Midterm 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 algorithm for checking
- Law of the unconcsious statistician (LOTUS)
- Monte Carlo error
  - How does the variance of our Monte Carlo
- Inversion Sampling
- Rejection 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
  - Identify the full conditional distributions
  - Pros and cons of a Gibbs sampling relative to MH sampling
- MCMC convergence Diagnostics
  - Run mulitple chains, different initalizations
  - ACF
  - Traceplot
  - Rejection rate
  - Effective sample size
  - How the size of the "jump" proposal effects the sampler
- Hierarchical / multilevel models (to be covered 12/3)
  - Complete pooling vs no pooling
  - Partial pooling
  - Relation to signal and noise variance

Practice Problems

1. We are interested in the parameter λ of a Poisson(λ) distribution. We have a prior distribution for λ with density

```
$$p(\lambda)=\left\{\begin{array}{ll}{0} & {(\lambda<0)} \\ {k_{0} \lambda^{3} e^{-\lambda}} & {(\lambda

#. Find the value of $\kappa_0$

#. Find the prior mean and prior standard deviation of $\lambda$.

    We now observe $y_1,..., y_n$ which are independent observations from the Poisson($\lambda$).

#.  Write the likelihood
#.  Write the posterior density of $\lambda$
#.  Write the posterior mean of $\lambda$
```

Solution:

(a) Notice that this looks like an unnormzlied Gamma density: $\lambda \sim$ Gamma(a, b) with $a = 4$ and $b = 1$. Then, referring to the gamma density we have

$$\kappa_0 = \frac{b^a}{\Gamma(a)} = \frac{1}{\Gamma(4)} = \frac{1}{3!} = 1/6$$

(b) The mean of a gamma is $a/b = 4$. The variance of a Gamma is $a/b^2 = 4$ so the standard deviation is 2.

(c)

$$L(\lambda|y) = \frac{\lambda^{\sum_{i=1}^n y_i} e^{-n\lambda}}{\prod_{i=1}^n y_i!}$$

(d)

$$p(\lambda|y) \propto \lambda^{\sum_{i=1}^n y_i} e^{-n\lambda} \times \lambda^3 e^{-\lambda} = \lambda^{\sum_{i=1}^n y_i + 4 - 1} e^{-(n+1)\lambda}$$

Then,

$$p(\lambda|y) = \frac{(n+1)^{\sum y_i+4}}{\Gamma(\sum y_i+4)}\lambda^{\sum_{i=1}^{n} y_i+4-1}e^{-(n+1)\lambda}$$

(e)

$$E(\lambda|y) = \frac{\sum_{i=1}^{n} y_i+4}{n+1}$$

2. In a fruit packaging factory apples are examined to see whether they are blemished. A sample of n apples is examined and, given the value of a parameter θ, representing the proportion of aples which are blemished, we regard y, the number of blemished apples in the sample, as an obervation from the binomial(n, θ) distribution. The chosen prior density for θ is

```
$$p(\theta)=\left\{\begin{array}{ll}{k_{0}\left(20 \theta(1-\theta)^{3}+1\right)} & {(0 \leq \theta \le
```

```
We observe $n=10$ apples and $y=4$.
```

```
#. Find the likelihood function.
#. Find the posterior density of $\theta$.
#. Find the posterior mean of $\theta$.
```

Solution

(a)

$$L(\theta|y) = \binom{10}{4}\theta^4(1-\theta)^6$$

(b)

$$p(\theta|y) \propto \theta^4(1-\theta)^6[20\theta(1-\theta)^3+1]$$

To find the constant part we need to write $p(\theta|y) = c\theta^4(1-\theta)^6[20\theta(1-\theta)^3+1]$

$$\int_0^1 p(\theta|y)d\theta = (20Beta(6,10)+Beta(5,7))c = 1$$

,

$$c = 1/(20Beta(6,10)+Beta(5,7)) = 910$$

(c)

$$E(\theta|y) = \int_0^1 \theta p(\theta|y)d\theta = 20c\int_0^1 \theta^6(1-\theta)^9 + c\int_0^1 \theta^5(1-\theta)^6 d\theta = 20cBeta(7,10)+cBeta(6,7) = 0.3914$$

where c is given in question b. Note $Beta(a,b)$ refers to the Beta function, which is one over the normalizing constant of Beta(a,b) distribution.

3. In a medical experiment, patients with a chronic condition are asked to say which of two treatments, A, B, they prefer. (You may assume for the purpose of this question that every patient will express a preference one way or the other). Let the population proportion who prefer A be θ. We observe a sample of n patients. Given θ, the n responses are independent and the probability that a particular patient prefers A is θ. Our prior distribution for θ is a beta(a, a) distribution with a standard deviation of 0.25.

```
#. Find the value of a.
#. True or false: the 95\% _prior_ quantile interval is the same as the 95\% prior HPD interval?
#. We observe n = 30 patients of whom 21 prefer treatment A. Find the posterior distribution of $\theta$
#. True or false: the 95\% _posterior_ quantile interval is the same as the 95\% prior HPD interval?
```

Solution

(a)

$$Var(Beta(a,a)) = \frac{a^2}{4a^2(2a+1)} = 0.25^2, \text{ then } a = 1.5$$

(b) In this example it is true, because the distribution is symmetric.

(c) The model we are using here is a binomial

$$L(\theta|y) \propto \theta^{21}(1-\theta)^9$$

$$p(\theta|y) \propto \theta^{21.5}(1-\theta)^{9.5}$$

The posterior distribution is a Beta(22.5, 10.5)

(d) False. The posterior is no longer symmetric, so the HPD will have different probabilities in left and right tails.

4. We observe a sample of 10 observations from a normal distribution with mean μ and precision $\frac{1}{\sigma^2}$ The data, $y_1, ..., y_{10}$, are such that

   (a) Suppose the value of $\frac{1}{\sigma^2}$ is known to be 0.004 and that our prior distribution is $p(\mu) \sim N(\mu_0 = 20, \sigma^2 = 100)$. 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.

   (d) What is the bias of the posterior mean, $E[\mu \mid Y_1, ...Y_{10}]$?

   (e) What is the variance of the posterior mean, $E[\mu \mid Y_1, ...Y_{10}]$?

   (f) What is the MSE of $E[\mu \mid Y_1, ...Y_{10}]$?

   (g) 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, $\bar{Y}$?

Solution
(a)

From the lecture notes, we know that

$$p(\mu \mid y, \sigma^2) \propto N(\mu_n, \sigma_n^2),$$

where $\mu_n = w\bar{y} + (1-w)\mu_0, \quad w = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2}$ and $\sigma_n^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}$

As given by the question, $n = 10$, $1/\sigma^2 = 0.004$, $\tau^2 = 100$, $\mu_0 = 20$, thus $w = 0.8$

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

$$\sigma_n^2 = \frac{1}{n/\sigma^2 + 1/\tau^2} = 20$$

95 % HPD interval:

$$\mu_n \pm 1.96 * \sigma_n = (0.8\bar{y} + 4) \pm 1.96 * \sqrt{20} = (0.8\bar{y} - 4.765386,\ 0.8\bar{y} + 12.76539)$$

(b)

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

(c)

Posterior mean estimator: $\mu_n = 0.8\bar{Y} + 4$

(d)

Bias: $E(0.8\bar{Y} + 4 - \mu) = (1 - w) * (\mu_0 - \mu) = 4 - 0.2 * \mu$

(e)

Variance: $Var(0.8\bar{Y} + 4) = w^2 * Var(\bar{Y}) = 16$

(f)

MSE: MSE = bias^2 + Var = $(4 - 0.2 * \mu)^2 + 16$

(g)

Let

$$(4 - 0.2 * \mu)^2 + 16 = Var(\bar{Y}) = 25,$$

solve it for μ:

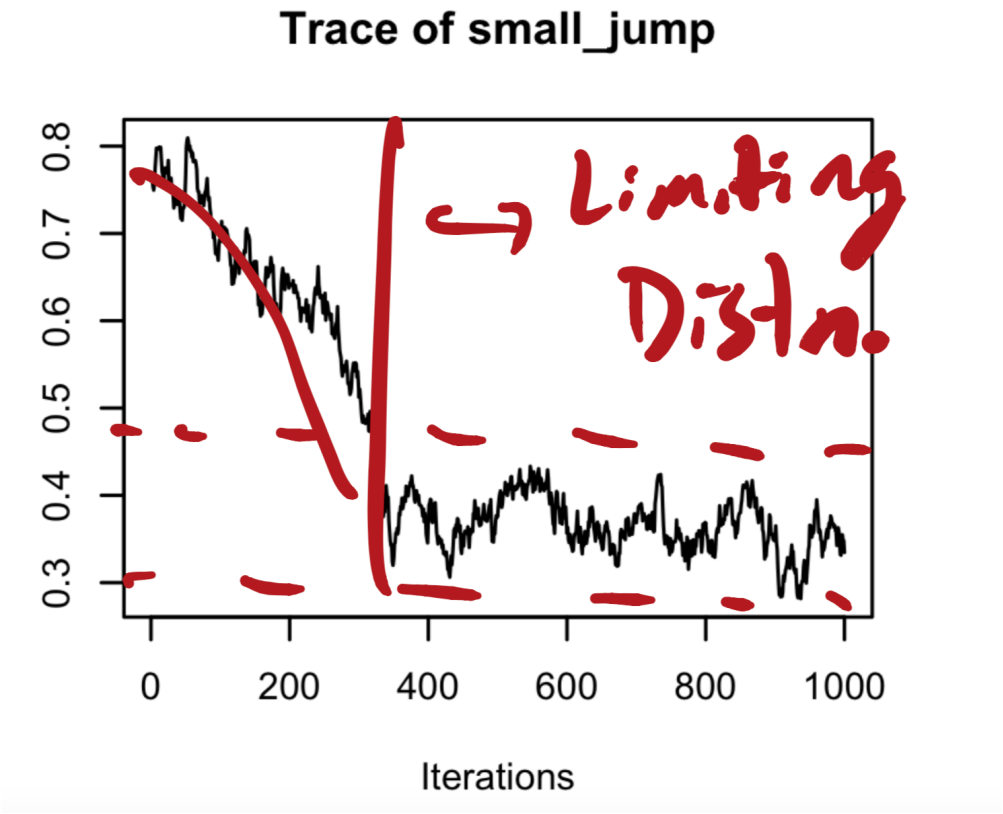
```
## [1] 5
```

Thus, when $\mu = 5$, the posterior mean estimator has equal MSE to the the maximum likelihood estimator.

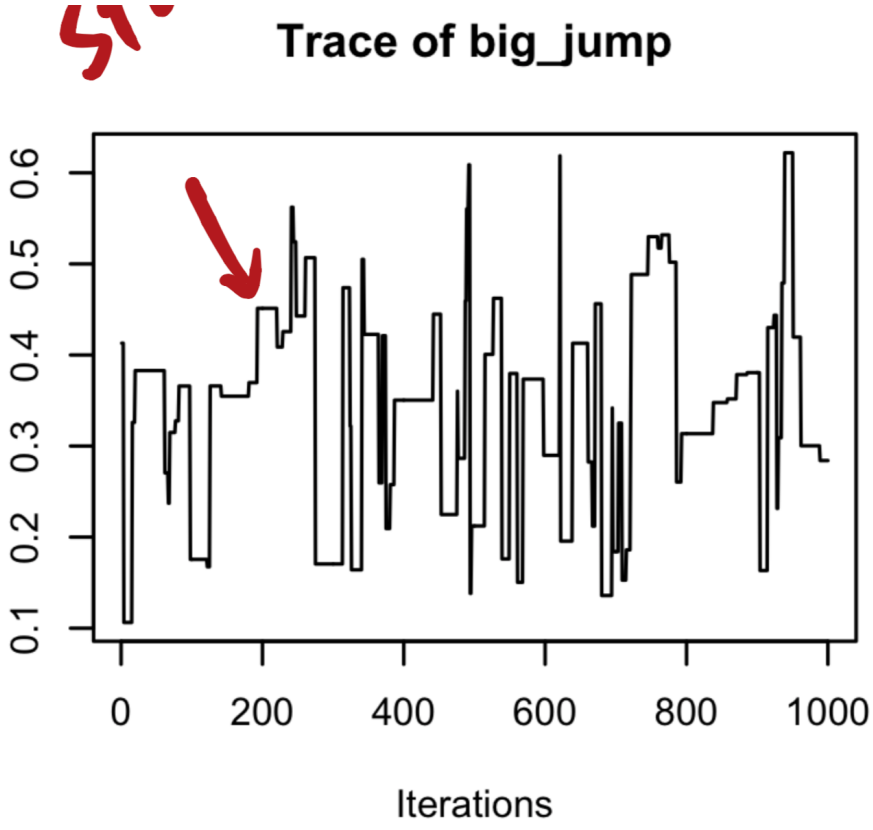
5. Draw a picture of a traceplot of a Markov Chain with high / low rejection rate.

solution:

Low rejection rate (small jumps, high autocorrelation):



High rejection rate (big jumps, "sticky" chain):

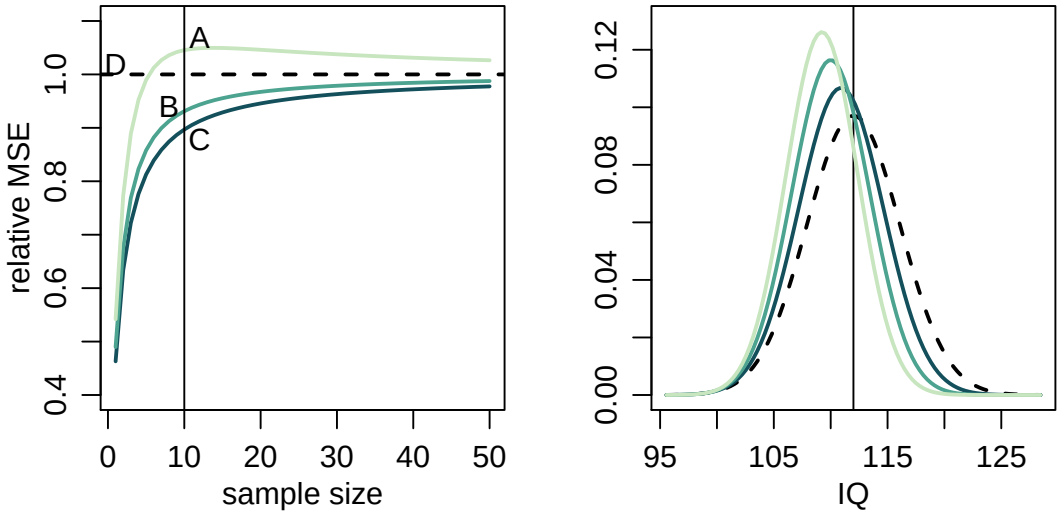


6. Consider Bayesian inference for p(μ, $\frac{1}{\sigma^2} \mid y$), where an $y \sim N(\mu, \sigma^2)$. Assume the non-infromative prior disribution $p(\mu, \sigma^2) \propto 1/\sigma^2$. This will lead to the diamond shaped posterior discussed in class and on the lecture notes. Argue in words (no 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?

solution:
When the uninformative prior has been employed, we can see the relationship between μ and $1/\sigma^2$ based on the sampling distribution.

- If $1/\sigma^2$ is very large, y is 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.

7. 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 $p(\mu) \sim N(100, \frac{13^2}{\kappa_0})$.

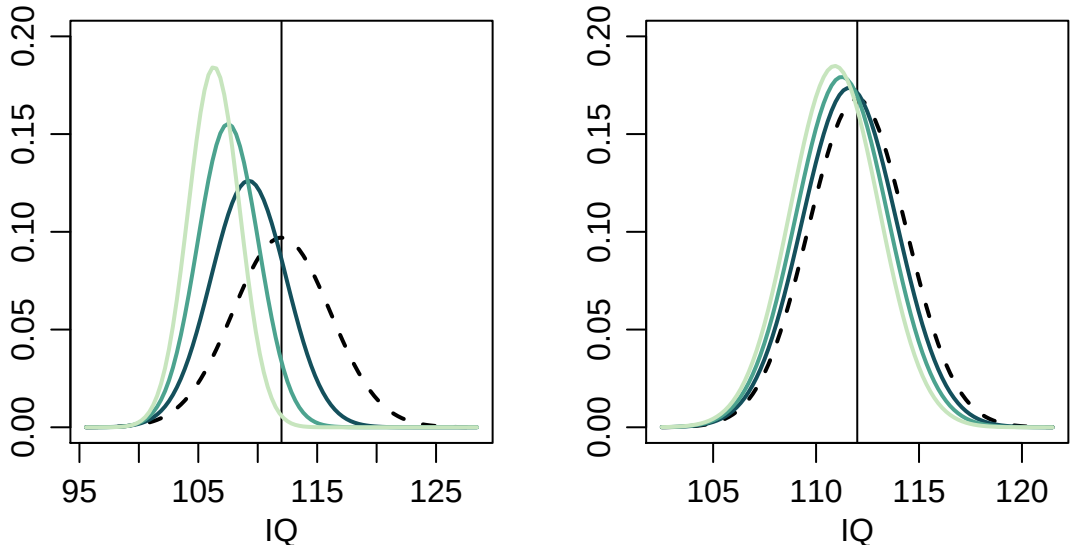


a. The left figure shows the mean squared error (MSE) of the posterior mean estimator relative to the maximum likelihood estimator. Fill in the blanks with the number 0, 1, 2, or 3. For line A $\kappa_0 =$ _____, for line B $\kappa_0 =$ _____, for line C, $\kappa_0 =$ _____, and for line D $\kappa_0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

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.

c. Circle all options that could describe the differences between the two figures below.



i. In the left figure, the values of κ_0 for each corresponding line are larger than they are for the right figure. Both have the same sample size, n.

ii. In the left figure, the values of κ_0 for each corresponding line are smaller than they are for the right figure. Both have the same sample size, n.

iii. In the left figure, n is larger than it is for the right figure. Both have the same values of κ_0.

iv. In the left figure, n is smaller than it is for the right figure. Both have the same values of κ_0.

solution:

(a)

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

(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 has bias when $\kappa > 0$ but reduced variance.

(c) Circle i, iv

Multiple Choice

Solution
1. a 2. b 3. a 4. ade 5. bd 6. FALSE 7. ab 8. complete pooling, partial pooling, no pooling. 9. Effective sample size, rejection rate and autocorrelation.

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

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

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, $p(\theta \mid y)$

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.
   (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. 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

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.

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

8. Consider esimates 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.

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