Homework 3

Your name here

Due Monday, February 20.

Theory problem:

Mixtures of independent distributions.

Suppose the distribution of $\theta = (\theta_1, \theta_2, ..., \theta_J)$ can be written as a mixture of independent and identically distributed components:

$$p(\theta) = \int \prod_j p(\theta_j|\phi)p(\phi)d\phi$$

Prove that $cov(\theta_i, \theta_j)$ are all non-negative. *Hint:* consider the law of total covariance.

Computing problem / Applied Problems:

Estimating Skill In Baseball

In baseball, the batting average is defined as the fraction of hits (successes) divided by "at bats" (attempts). We can conceptualize a player's "true" batting skill as $\theta = \lim_{n_i \to \infty} \frac{y_i}{n_i}$. In other words, if each at bat was independent (a simplifying assumption), θ_i describes the total fraction of success for player i as the number of attempts gets very large. Our goal is to estimate the true skill of all player as best as possible using only a limited amount of data. As usual, for independent counts of success/fail data it is reasonable to assume that $Y_i \sim \text{Bin}(n_i, \theta_i)$. The file "lad.csv" includes the number of hits, `y` and the number of attempts `n` for $J = 10$ players on the Los Angeles Dodgers after the first month of the 2022 baseball season. The variable `val` includes the end-of-season batting average, which we will take as our standin for the true θ_i and will be used to validate the quality of various estimates. If you are interested, at the end of the assignment I have included the code that was used to scrape the data.

```r
baseball_data <- read_csv("lad.csv", col_types=cols())
baseball_data
```

```
# A tibble: 12 x 4
   name                  y     n   val
   <chr>             <dbl> <dbl> <dbl>
 1 Austin Barnes         4    19 0.216
 2 Chris Taylor         19    66 0.221
 3 Cody Bellinger       16    77 0.207
 4 Edwin Rios            5    21 0.244
 5 Freddie Freeman      25    81 0.329
 6 Gavin Lux            14    54 0.28
 7 Hanser Alberto        5    18 0.234
 8 Justin Turner        14    75 0.28
 9 Max Muncy             9    66 0.2
10 Mookie Betts         18    78 0.271
11 Trea Turner          21    80 0.299
12 Will Smith           13    51 0.266
```

```r
## observed hits in the first month
y <- baseball_data$y
## observed at bats in the first month
n <- baseball_data$n
## observed batting average in the first month (same as MLE)
theta_mle <- y/n
## number of players
J <- nrow(baseball_data)
## end of the year batting average, used to evaluate estimates
val <- baseball_data$val
```

a) Compute the standard deviation of the empirical batting average, $\hat{\theta}_{MLE} = y/n$, and then compute the sd of the "true skill", (the `val` variable representing the end of season batting average which we take as a proxy for θ). Which is smaller? Comment on why this meets (or does not meet) your expectations.

b) Consider two estimates for the true skill of player i, θ_i: 1) $\hat{\theta}_i^{(\text{MLE})} = \frac{y_i}{n_i}$ and 2) $\hat{\theta}_i^{(\text{comp})} = \frac{\sum_j y_j}{\sum n_j}$. Estimator 1) is the MLE for each player and ignores any commonalities between the observations. This is sometimes termed the "no pooling" estimator since each parameter is estimating separately without "pooling" information between them. Estimator 2) assumes all players have identical skill and is sometimes called the "complete pooling" estimator, because the data from each problem is completely "pooled" into one common set. In this problem, we'll treat the end-of-season batting average as a proxy for true skill, θ_i. Compute the root mean squared error (RMSE), $\sqrt{\frac{1}{J}\sum_i(\hat{\theta}_i - \theta_i)^2}$ for the "no pooling" and "complete pooling" estimators using the variable `val` as a stand-in for the true θ_i. Does "no pooling" or "complete pooling" give you a better estimate of the end-of-year batting averages in this specific case?

$$
\begin{aligned}
y_i &\sim Bin(n_i, \theta_i) \\
\theta_i &= Beta(a, b) \\
p(a, b)
\end{aligned}
$$

a) Implement the hierarchical binomial model following the example on the model for rat tumors in Section 5.3 (or the basketball example from class). Choose your own proper subjective prior on the prior mean $a/(a+b)$ and on $log(a+b)$ based on your knowledge about baseball and convert this to a prior on $p(a,b)$ (don't forget the jacobian!).

   - Use `optim` to find the mode of the marginal posterior. Use this mode as a guide to set limits on a grid and then use grid sampling to draw samples from the marginal posterior $p(\alpha, \beta \mid y)$.
   - Use Monte Carlo samples from the marginal posterior to create a histogram of the marginal posterior distribution of $a/(a+b)$. Report the marginal posterior mean and a 95% credible interval for $a/(a+b)$.

b) Given the samples from α and β now generate Monte Carlo samples of θ_i for each player. Compute the posterior mean for each player. Use the following function to make a shrinkage plot. Pass in `y/n` and the posterior means of θ_i and the posterior mean of $a/(a+b)$ as arguments. Report the RMSE for the posterior mean in the hierarchical model. How does this compare to the RMSEs under the no pooling and complete pooling models?

   ::: {.cell}

```
shrinkage_plot <- function(observed_frac,
                           posterior_mean,
                           shrink_point=1/2) {

  tibble(y=observed_frac, pm=posterior_mean) %>%
    ggplot() +
    geom_segment(aes(x=y, xend=pm, y=1, yend=0), linetype="dashed") +
    geom_point(aes(x=y, y=1)) +
    geom_point(aes(x=pm, y=0)) +
    theme_bw(base_size=16) +
```

```
    geom_vline(xintercept=shrink_point, color="blue", size=1.2) +
    ylab("") + xlab("Estimate") +
    scale_y_continuous(breaks=c(0, 1),
                       labels=c("Posterior Mean", "MLE"),
                       limits=c(0,1)) +
    scale_color_continuous(type="viridis")
}
```

:::

c) True or false: as the number of at bats, n_i goes to infinity for all i, then the global batting average, $a/(a+b)$, is perfectly identified, e.g. the marginal posterior $p(a/(a+b) \mid y)$ collapses to a point at the true value. Argue why or why not.

d) Assume you have access to hits and at bats for the first month of the season, for all baseball players across all teams, not just the LA dodgers. You can also assume that you have access to any additional metadata about the players themselves (age, type of player etc). Propose a data generating process for a hierarchical model of this data. Use any prior knowledge you have to argue that your proposed model is reasonable.

Analyzing Bike traffic

A survey was done of bicycle and other vehicular traffic in the neighborhood of the campus of the University of California, Berkeley, in the spring of 1993. Sixty city blocks were selected at random; each block was observed for one hour, and the numbers of bicycles and other vehicles traveling along that block were recorded. The sampling was stratified into six types of city blocks: busy, fairly busy, and residential streets, with and without bike routes, with ten blocks measured in each stratum. For this problem, we will restrict your attention to residential streets labeled as 'bike routes,' which we will use to illustrate this computational exercise. In this exercise we'll focus on modeling only on the total amount of traffic at each location. The counts represent the observed traffic at the intersection in one hour.

```
bikes <- c(16, 9, 10, 13, 19, 20, 18, 17, 35, 55)
other <- c(58, 90, 48, 57, 103, 57, 86, 112, 273, 64)
total <- bikes + other
```

1. Set up a model in which the total number of vehicles observed at each location j follows a Poisson distribution with parameter θ_j, the 'true' rate of traffic per hour at that location. Assign a gamma(a, b) population distribution for the parameters θ_j and a noninformative hyperprior distribution for a and b. Write down the joint posterior distribution for all parameters.

2. Compute the marginal posterior density of the hyperparameters and plot its contours. Simulate random draws from the posterior distribution of the hyperparameters and make a scatterplot of the simulation draws.
3. Is the posterior density integrable? Answer analytically by examining the joint posterior density at the limits or empirically by examining the plots of the marginal posterior density above. If the posterior density is not integrable, alter it and repeat the previous two steps.
4. Draw samples from the joint posterior distribution of the parameters and hyperparameters.

5. Make a plot analogous to figure 5.4 in BDA, where the x-axis is and the y-axis includes the posterior medians and 95% credible intervlas for θ_j. Include the posterior mean estimate for the overall average rate of traffic in Berkeley city streets as a horizontal line.
6. Was it reasonable for you to assume the observations are exchangeable in this problem? Discuss in high level terms when it might make sense to assume a non-exchangeable prior distribution for θ_j's and what kinds of models might be needed.

Appendix: Code used for scraping Dodgers baseball data

http://billpetti.github.io/baseballr/

```
## Install the baseballr package
devtools::install_github("BillPetti/baseballr")

library(baseballr)
library(tidyverse)

## Download data from the chosen year
year <- 2022

one_month <- daily_batter_bref(t1 = sprintf("%i-04-01", year),
                               t2 = sprintf("%i-05-01", year))
one_year <- daily_batter_bref(t1 = sprintf("%i-04-01", year),
                              t2 = sprintf("%i-10-01", year))

## filter to only include players who hat at least 10 at bats in the first month
one_month <- one_month %>% filter(AB > 10)
one_year <- one_year %>% filter(Name %in% one_month$Name)

one_month <- one_month %>% arrange(Name)
one_year <- one_year %>% arrange(Name)
```

```r
## Look at only the Dodgers
LAD <- one_year %>% filter(Team == "Los Angeles" & Level == "Maj-NL") %>% .$Name

lad_month <- one_month %>% filter(Name %in% LAD)
lad_year <- one_year %>% filter(Name %in% LAD)

write_csv(tibble(name=lad_month$Name,
                 y=lad_month$H,
                 n=lad_month$AB,
                 val=lad_year$BA),
          file="lad.csv")
```