Applied Bayesian Analysis : NCSU ST 540

Homework 1

Bruce Campbell

Bayesian analysis of Clutch Shots

In this problem we're asked to analyze success rate of NBA player clutch shots. We're given an overall success rate and data for clutch successes and attempts. Let i denote player i θ_i be our clutch success proportion, p_i the player's overall success proportion N_i the number of clutch attempts and Y_i the number of clutch successes. We will then model the likelihood with a $Binomial(N, \theta)$ distribution and for mathematical tractability we'll model the prior with a $Beta(\alpha, \beta)$ distribution.

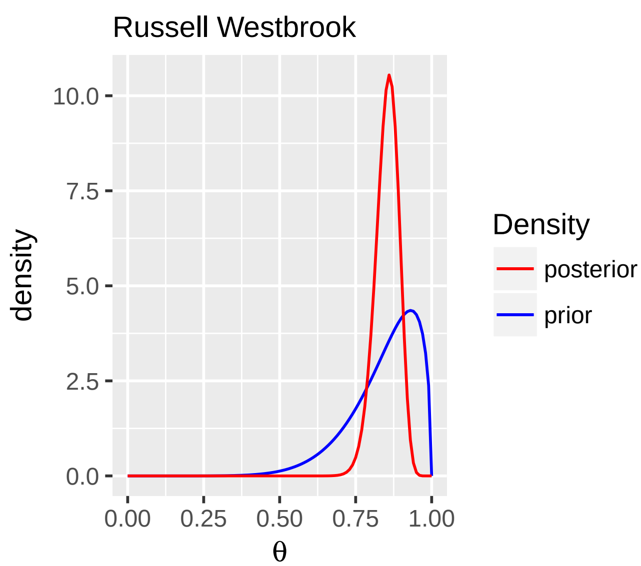
We'll use the overall proportion p_i to inform our prior. The text book does a great job explaining how the parameters of $Beta(\alpha, \beta)$ can be related to the mean $\mu = \frac{\alpha}{\alpha+\beta}$ and spread $\kappa = \alpha + \beta$ We'll choose (α, β) so the prior mean is p_i. Now the spread κ can be thought of as the minimum number of trials required for us to consider updating our belief in μ. 10 seems like the minimum number of samples we'd need to even begin to contemplate updating our beliefs and we'd like not to bias our analysis by choosing too large a number. Also the data available may play a role in helping us design our prior. The number of attempts $N_i \in [16, 95]$ so we'd not want to go above this range. It's worth contemplating if one should set $\kappa = N$. For now we will use $\kappa = 10$ and revisit that choice when we look at the sensitivity of our analysis to the prior.

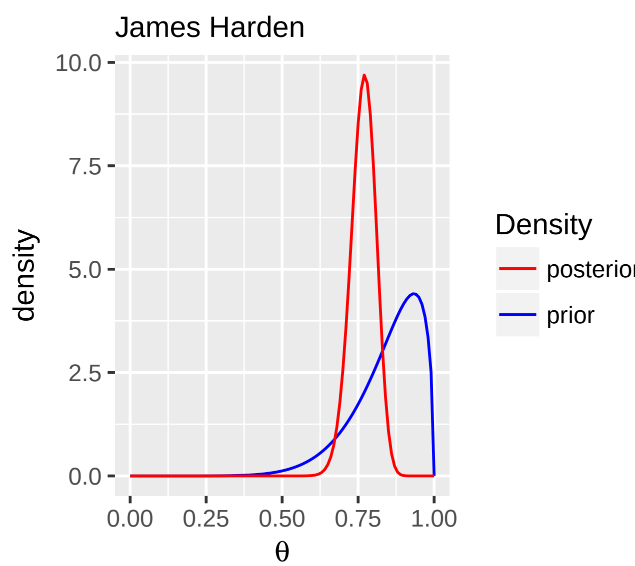
Plots of the posterior and prior used.

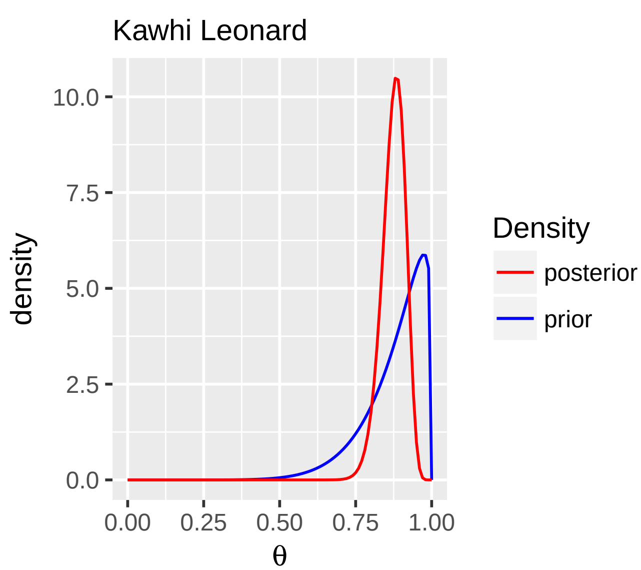
```r
# We load the data from a file here.
library(readr)
library(gridExtra)
NBA_Data <- read_csv("NBA_Data.csv", col_types = cols(ClutchAttempts = col_number(),
    ClutchMakes = col_number()))
NBA_Data$clutchProportion <- NBA_Data$ClutchMakes/NBA_Data$ClutchAttempts
# pander(NBA_Data)

# We'll plot the posteriors on a single
# page after our calculations. This list
# stores the plots for us to use later.
plots <- c()

# We'll store the posterior mean and mode
# for summarizing the poseterior in a
# table.
NBA_Data[, "posteriorMean"] <- NA
NBA_Data[, "posteriorMode"] <- NA
```

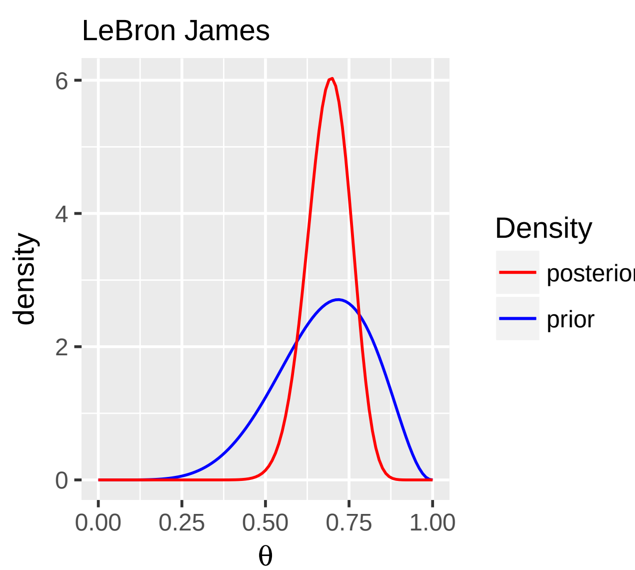
```r
# We also store the posterior parameters
# for calculating some probabilities
NBA_Data[, "posteriorA"] <- NA
NBA_Data[, "posteriorB"] <- NA


# Calculate posterior for each player
# give the prior described above.
for (i in 1:nrow(NBA_Data)) {
    # Calculate the prior parameters
    mean <- NBA_Data[i, ]$proportion
    kappa <- 10
    a = mean * kappa
    b = (1 - mean) * kappa

    if (a < 1 || b < 1)
        warning("Prior parameter warning")
    # Extract the N and Y from the data to
    # form the likelihood
    N <- NBA_Data[i, ]$ClutchAttempts
    Y <- NBA_Data[i, ]$ClutchMakes

    titleString <- NBA_Data[i, ]$PlayerName

    # Parameters for the posterior
    # distribution
    posteriorA <- Y + a
    posteriorB <- N - Y + b

    NBA_Data[i, ]$posteriorA <- posteriorA
    NBA_Data[i, ]$posteriorB <- posteriorB

    # Calculate posterior mean and variance
    posteriorMean <- posteriorA/(posteriorA +
        posteriorB)
    posteriorMode <- (posteriorA - 1)/(posteriorA +
        posteriorB - 2)
    posteriorVariance <- posteriorA * posteriorB/((posteriorA +
        posteriorB)^2 * (posteriorA + posteriorB +
        1))
    posteriorSD <- sqrt(posteriorVariance)

    # We use ggplot2's stat_function instead
    # of curve
    p1 <- ggplot(data.frame(x = c(0, 1)),
        aes(x)) + stat_function(fun = function(x) dbeta(x,
        shape1 = a, shape2 = b), aes(colour = "prior")) +
```

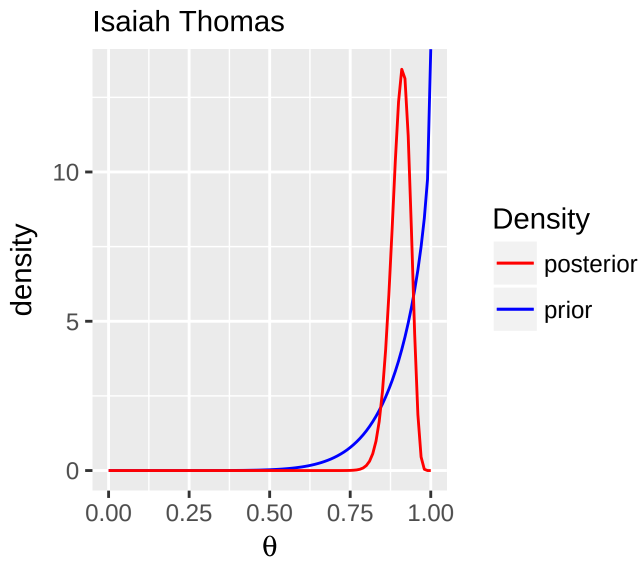
```
        ylab("density") + xlab(TeX("$\\theta$"))
    p2 <- p1 + stat_function(fun = function(x) dbeta(x,
        shape1 = posteriorA, shape2 = posteriorB),
        aes(colour = "posterior")) + ggtitle(titleString) +
        scale_colour_manual("Density", values = c("red",
            "blue")) + theme(plot.title = element_text(size = 11))
    print(p2)
    plots[[i]] <- p2
    p2 <- NULL

    NBA_Data[i, ]$posteriorMean <- posteriorMean
    NBA_Data[i, ]$posteriorMode <- posteriorMode

}
```

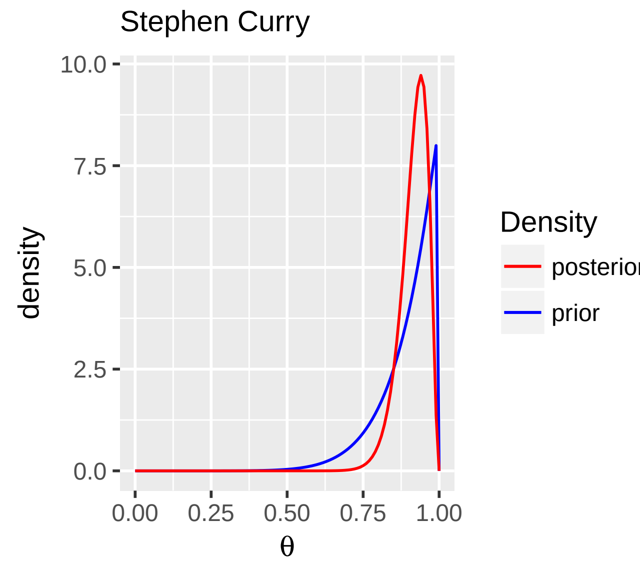


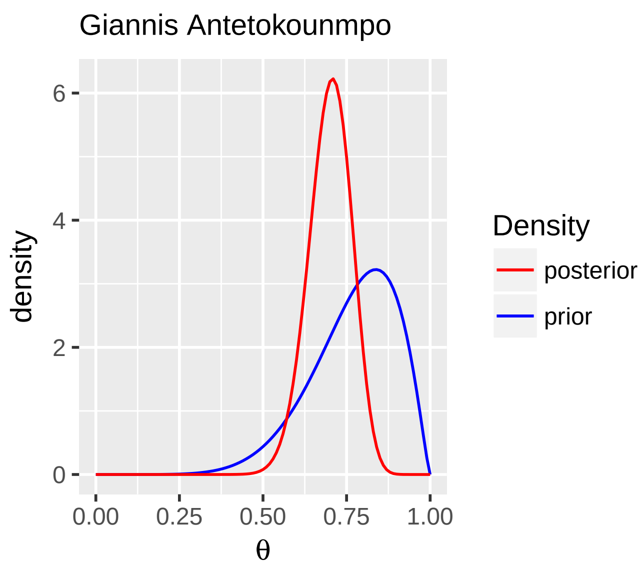


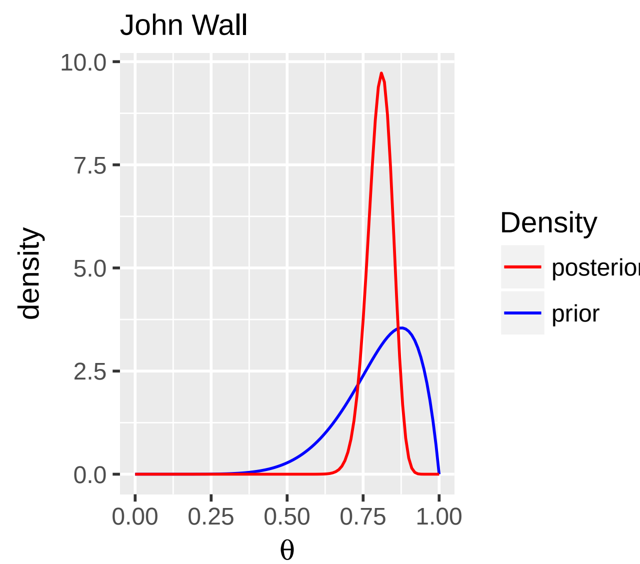


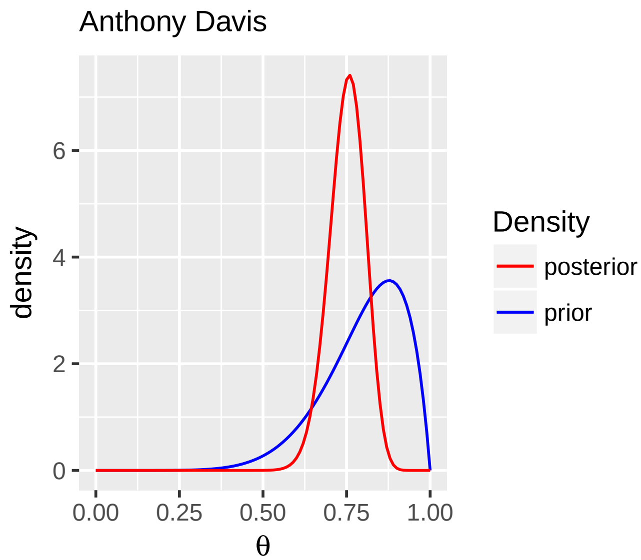


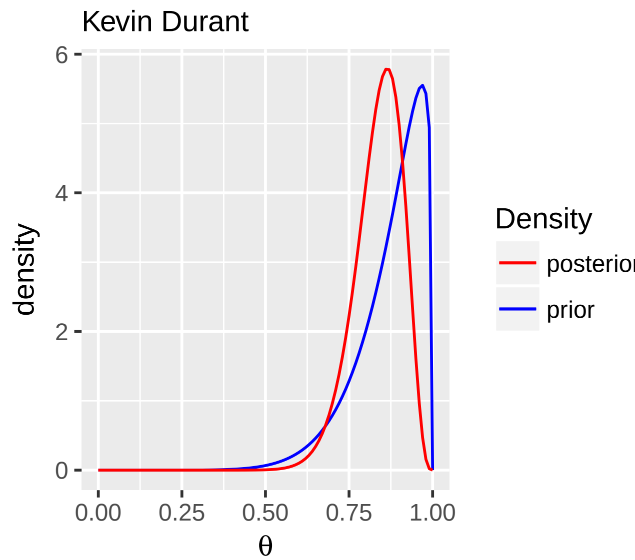

Isaiah Thomas
density
0
5
10
0.00
0.25
0.50
0.75
1.00
θ
Density
posterior
prior


Stephen Curry
density
0.0
2.5
5.0
7.5
10.0
0.00
0.25
0.50
0.75
1.00
θ
Density
posterior
prior


Giannis Antetokounmpo
density
0
2
4
6
0.00
0.25
0.50
0.75
1.00
θ
Density
posterior
prior


John Wall
density
0.0
2.5
5.0
7.5
10.0
0.00
0.25
0.50
0.75
1.00
θ
Density
posterior
prior






```
# MADDENING! I tried to plot them all on
# one ggplot page The titles differ but
# the plots are all the first plot.
# grid.arrange(grobs = plots, ncol = 2)
```

Summary of Posterior distributions

```
pander(data.frame(names = NBA_Data$PlayerName,
    posterior.mean = NBA_Data$posteriorMean,
    posterior.mode = NBA_Data$posteriorMode))
```

names	posterior.mean	posterior.mode
Russell Westbrook	0.8524	0.8608
James Harden	0.7664	0.7716
Kawhi Leonard	0.874	0.8845
LeBron James	0.6886	0.6966
Isaiah Thomas	0.9042	0.9131
Stephen Curry	0.9161	0.9406
Giannis Antetokounmpo	0.7	0.7082
John Wall	0.8045	0.8112
Anthony Davis	0.7503	0.7584
Kevin Durant	0.8365	0.8646

Testing the hypothesis that the clutch proportion is less than the overall proportion

Here we test using the principle of MAP - maximum a poseritori. We are performing the test $H_0 : \frac{Y}{N} < p$ versus the slternative $H_A : \frac{Y}{N} \geq p$. We'll use the posterior distribution to perform the test and will accept H_0 when $P(H_0|Y) > P(H_A|Y)$

```r
NBA_Data[, "posteriorProbC"] <- NA
NBA_Data[, "posteriorProbP"] <- NA

NBA_Data[, "testResult"] <- NA

for (i in 1:nrow(NBA_Data)) {
    a <- NBA_Data[i, ]$posteriorA
    b <- NBA_Data[i, ]$posteriorB
    p <- NBA_Data[i, ]$proportion

    N <- NBA_Data[i, ]$ClutchAttempts
    Y <- NBA_Data[i, ]$ClutchMakes

    NBA_Data[i, ]$posteriorProbP <- pbeta(p,
        a, b)

    NBA_Data[i, ]$posteriorProbC <- pbeta(Y/N,
        a, b)

    NBA_Data[i, ]$testResult <- (NBA_Data[i,
        ]$posteriorProbC < NBA_Data[i,
        ]$posteriorProbP)

}

pander(data.frame(name = NBA_Data$PlayerName,
    overall = NBA_Data$posteriorProbP,
    clutch = NBA_Data$posteriorProbC, test.result = NBA_Data$testResult))
```

name	overall	clutch	test.result
Russell Westbrook	0.3975	0.4814	FALSE
James Harden	0.9827	0.4035	TRUE
Kawhi Leonard	0.5272	0.4551	TRUE
LeBron James	0.3989	0.5069	FALSE
Isaiah Thomas	0.5255	0.4546	TRUE
Stephen Curry	0.2972	0.4938	FALSE
Giannis Antetokounmpo	0.8653	0.381	TRUE
John Wall	0.4457	0.4827	FALSE
Anthony Davis	0.8301	0.4118	TRUE
Kevin Durant	0.671	0.3317	TRUE

Sensitivity Analysis

Here we calculate the sensitivity of the results to the choice of prior. We redo the hypothesis test using the completely uninformative prior $Beta(1, 1)$ first and extend to priors with smaller κ if necessary.

```
NBA_Data[, "uninformativeTestResult"] <- NA
NBA_Data[, "uninformativePosteriorProbC"] <- NA
NBA_Data[, "uninformativePosteriorProbP"] <- NA

for (i in 1:nrow(NBA_Data)) {
    a = 1
    b = 1

    N <- NBA_Data[i, ]$ClutchAttempts
    Y <- NBA_Data[i, ]$ClutchMakes

    p <- NBA_Data[i, ]$proportion

    # Parameters for the posterior
    # distribution
    posteriorA <- Y + a
    posteriorB <- N - Y + b

    NBA_Data[i, ]$uninformativePosteriorProbP <- pbeta(p,
        posteriorA, posteriorB)

    NBA_Data[i, ]$uninformativePosteriorProbC <- pbeta(Y/N,
        posteriorA, posteriorB)

    NBA_Data[i, ]$uninformativeTestResult <- (NBA_Data[i,
        ]$uninformativePosteriorProbC <
        NBA_Data[i, ]$uninformativePosteriorProbP)

}

pander(data.frame(name = NBA_Data$PlayerName,
    overall = NBA_Data$uninformativePosteriorProbP,
    clutch = NBA_Data$uninformativePosteriorProbC,
    test.result = NBA_Data$uninformativeTestResult))
```

name	overall	clutch	test.result
Russell Westbrook	0.4793	0.5606	FALSE
James Harden	0.991	0.5326	TRUE
Kawhi Leonard	0.6402	0.5738	TRUE
LeBron James	0.4354	0.535	FALSE
Isaiah Thomas	0.6446	0.5785	TRUE
Stephen Curry	0.4707	0.6552	FALSE
Giannis Antetokounmpo	0.9167	0.5322	TRUE
John Wall	0.5092	0.5448	FALSE
Anthony Davis	0.8867	0.5393	TRUE
Kevin Durant	0.8457	0.6015	TRUE

name	overall	clutch	test.result

We see that the test results are not senitive to the choice of prior. This is really striking given that some of the data elemets have a low count for the attempts. The lowest being 16.

Conjugate Prior for Independent Binomial Random Variables.

Suppose $Y \sim Binomial(N, \theta)$, $Z \sim Binomial(M, \theta)$ and that Y Z are independent. Identify a conjugate prior and find the posterior distribution.

A conjugate prior is one that comes from a family of distributions such that when inserted into the expression for Bayes theorem results in a posterior that comes from the same family. The Beta distribution is the conjugate prior for the Binomial, Bernoulli, Geometric, and Negative Binomial likelihood distributions. There are several ways to think about this problem. One is that both distributions come from a series of independent Bernoulli trials and we may simply calculate the posterior from the likelihood of $Y + Z$ from a $Binomial(N + M, \theta)$. The other way is to consider the joint likelihood distribution (Y, Z) and utilize the independence to factor the likelihood. We can show both methods yield the same $Beta$ posterior.

$$P(\theta|Y+Z) \propto \binom{N+M}{Y+Z}\theta^{Y+Z}(1-\theta)^{N+M-(Y+Z)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Rearranging and dropping constants we have

$$P(\theta|Y+Z) \propto \theta^{\alpha+Y+Z-1}(1-\theta)^{\beta+N+M-(Y+Z)-1}$$

Which we recognize to be the kernel of a $Beta(\alpha + Y + Z, \beta + N + M - (Y + Z))$ distribution.

Now let's consider the second approach

$$P(\theta|Y,Z) \propto \binom{N}{Y}\theta^{Y}(1-\theta)^{N-Y}\binom{M}{Z}\theta^{Z}(1-\theta)^{M-Z}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Again rearranging and dropping constants we have

$$P(\theta|Y,Z) \propto \theta^{\alpha+Y+Z-1}(1-\theta)^{\beta+N+M-(Y+Z)-1}$$

Which we again recognize to be the kernel of a $Beta(\alpha + Y + Z, \beta + N + M - (Y + Z))$ distribution.