Motivation

As players in the National Basketball Association continue to increase their shooting abilities and push the boundary of the definition of a "good" shot, the league is evolving to be ever more shooter-centric. Additionally, the advent of the basketball data revolution has ushered in a new era of emphasis on analytics as teams vie for the smallest edge on their competition. As a byproduct of both of these trends, there is a clear demand for advanced shooting-ability metrics. However, current evaluation of shooting centers around simple percentages and shot charts or ventures into the realm of subjective opinion. This paper proposes a model to evaluate players' shooting ability that includes game and play-specific situational variables that have an obvious effect on shooting but are not commonly included in analyses. Including these additional factors will allow for conclusions to be made on their effect on individual player ability and show how teams can best exploit the unique advantages afforded to them by their players' abilities.

Data & Model

This analysis was done using data on over 128,000 shots taken during the 2014-2015 NBA season. These shots were segmented by player, and for ease of analysis only nine players[1] of interest were considered. The response was the result of each shot, an ordinal variable with categories for 0, 2, or 3 points scored. Free throws were not considered. There were five covariates of interest, given below. Figure 1 and 2 display an exploratory glance at the data from one player.

- **Location**: Whether the game was played at home or away, from the perspective of the shooting player
- **Shot Clock**: Time remaining on the shot clock when the shot was taken
- **Dribbles**: Number of times the ball was dribbled by the player prior to the shot being taken
- **Touch Time**: Amount of time the player has possessed the ball prior to the shot being taken
- **Closest Defender Distance**: Distance from the nearest defender to the player taking the shot

Jimmer Fredette's 2014-15 Shot Data

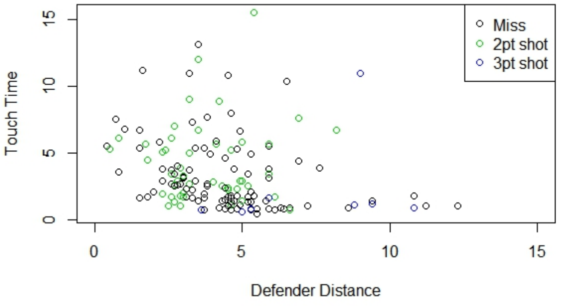

Figure 1: There are more made shots as distance from the nearest defender increases and touch time decreases, with almost all the three point makes coming from "catch and shoot" situations.

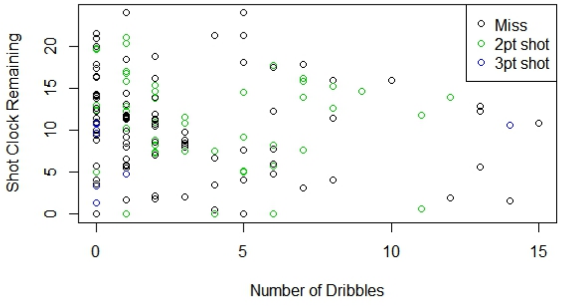

Figure 2: There are more attempted and a higher ratio of made shots as the number of dribbles decreases. More shots are attempted in the middle of the shot clock, but it is difficult to tell where more are made.

[1] Anthony Davis, Gordon Hayward, James Harden, Jimmer Fredette, Klay Thompson, Kobe Bryant, LeBron James, Russell Westbrook, Stephen Curry

Hierarchical Latent Variable Probit Model

The ordinal data is modeled with a latent variable Z, which uses a probit link to model the underlying distribution of the response on a standardized scale. The selection of this model was motivated by a desire to generate β effect sizes for the covariates and to fit a hierarchical structure that allows for comparison between players. This model also results in the derivation of posterior distributions from which prediction and meaningful statements of probability can be generated.

$$\mathbf{Z}_{ik} \sim \mathcal{N}(\mathbf{X}_k'\boldsymbol{\beta}_k, 1)$$
$$\mathbf{Y}_{ik} = j, \quad \text{if} \quad \gamma_{j-1} \leq \mathbf{Z}_{ik} \leq \gamma_j$$
$$j = \{0, 2, 3\}$$
$$k = 1, \ldots, n_{players}$$
$$i = 1, \ldots, n_{shots}$$
$$\boldsymbol{\beta}_k \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\mu} \sim MVN(\mathbf{m}, \mathbf{V})$$
$$\boldsymbol{\Sigma} \sim InverseWishart(w, \mathbf{I})$$

There are two γ cutpoints, which are fixed at 2 and 3 as those are natural points to cut. Values for the prior on $\boldsymbol{\mu}$ are set as $\mathbf{m} = \mathbf{0}$ and $\mathbf{V} = \mathbf{I}$, a 5x5 identity matrix. Values are the prior on $\boldsymbol{\Sigma}$ are set as w = 1 + $n_{players}$ and $\mathbf{I}$ = a 5x5 identity matrix. The model is implemented by deriving complete conditionals (see Appendix) and using a Gibbs Sampling technique to obtain draws from the posterior distributions. The results here are from a run of 10,000 draws with a burn of 500.

Results

Examining the β posterior distributions answers the research question of the effect of different factors on individual player ability. As this is a probit model, the effect sizes are interpreted as the change in Z-score across the scale of the ordinal response (from missed shot to two points to three points). A few interesting observations from these results are highlighted below.

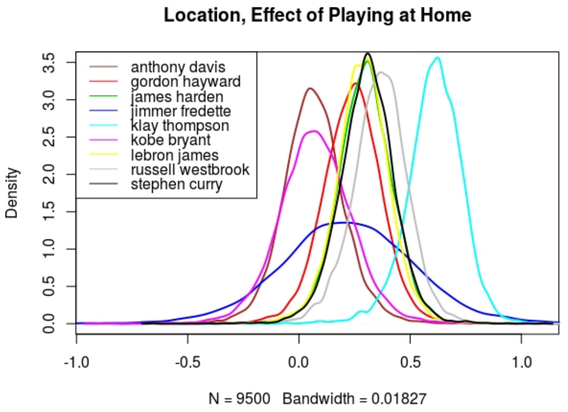

Figure 3: Thompson is significantly more efficient when at home, in contrast to Bryant and Davis, who are hardly effected.

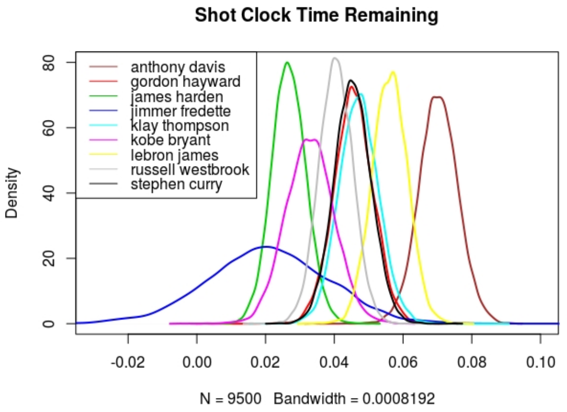

Figure 4: Davis is more efficient with more time on the shot clock, in contrast to Fredette and Harden, who are much less effected.

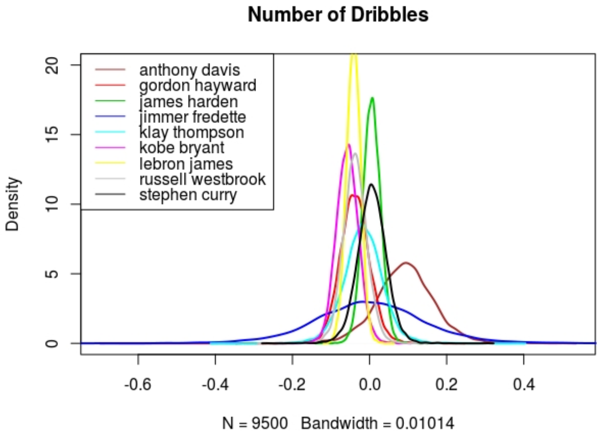

Figure 5: Davis is more efficient as he takes more dribbles, in contrast to the other players, who are very slightly less efficient bordering on no effect.

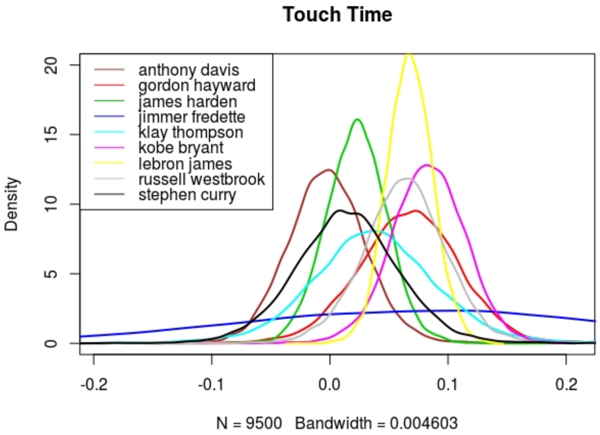

Figure 6: Bryant and James are more efficient with longer touch times times, in contrast to Curry and Davis, who are hardly effected.

Figure 3 needs no further explanation. Shot clock time remaining in Figure 4 is an interesting covariate because of unique behavior for different types of players at the beginning and end of the shot clock. However, the results are able to bear out an overall trend. Number of dribbles and touch time in Figures 5 and 6 are obviously correlated, but both were included in the regression model in hopes that different players would have different correlation relationships, which the results partially supported. Closest defender distance in Figure 7 shows interesting results, but this variable does not account for a player's ability to create space from a defender on his own, which discounts a player like Harden, who utilizes a dynamic step-back move. In addition, as a way of explanation, Fredette's distribution has much more variability due to the comparatively low number of shot attempts he recorded.

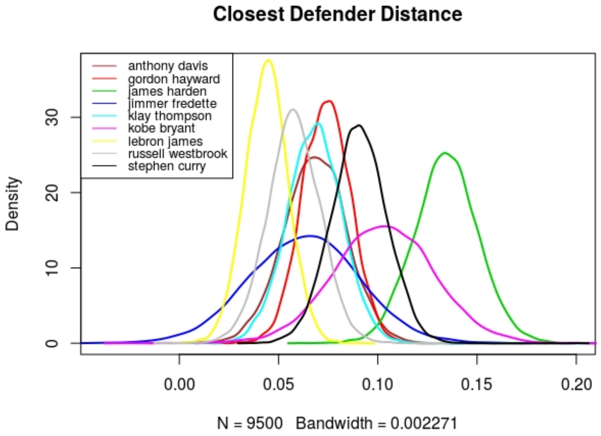

Figure 7: Harden is significantly more efficient when a defender is further away, in contrast to James and Westbrook, who are hardly effected.

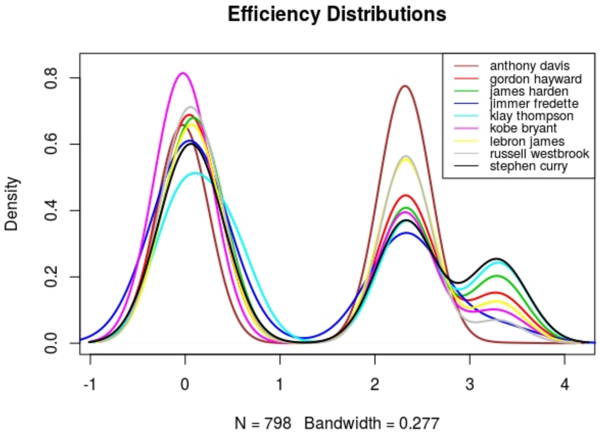

Figure 8: Bryant is much more likely to miss a shot, Davis excels at two point shots, and the "Splash Brothers", Curry and Thompson top the three point category.

Figure 8 shows a cumulative distribution of player efficiency, obtained from the density of the means of the draws of each player's Z-distributions. This distribution shows a player's ability to either miss or make a two point shot or three point shot. To answer the research question of direct player ability evaluation, the cursory glance at player ability shown by this plot can be formalized by extension to a posterior predictive distribution. This distribution reflects the probability of a player's next shot being either a miss, two point make, or three point make and is seen in Table 1.

	Miss	2 Points	3 Points	Composite Score
Stephen Curry	0.7355	0.2022	0.0623	0.5914
Klay Thompson	0.7483	0.1938	0.0579	0.5613
LeBron James	0.7657	0.1754	0.0589	0.5276
Anthony Davis	0.7769	0.1755	0.0476	0.4937
James Harden	0.7832	0.1701	0.0467	0.4804
Gordon Hayward	0.7838	0.1702	0.0460	0.4784
Russell Westbrook	0.7947	0.1613	0.0440	0.4545
Jimmer Fredette	0.8160	0.1303	0.0537	0.4217
Kobe Bryant	0.8295	0.1336	0.0369	0.3780

Table 1: The ”Splash Brothers” reign supreme on the strength of their three point shooting. Despite a comparatively low number of shot attempts, Fredette received a higher rating than post-Achilles Kobe Bryant. Interestingly, players like James and Davis have high three point ratios although they don’t attempt many. This suggests they should shoot more threes, which is in line with the direction the league is heading.

Table 1 represents the posterior predictive distribution in columns one through three. This distribution reflects the discrete multinomial data with π_j probabilities of each player’s next shot of being in category j. The composite score ranks player ability by multiplying each π_j by its corresponding 0, 2, 3 point value and summing across each player.

Model Selection & Goodness of Fit

A comparable model is a simple Gaussian hierarchical model with $Y \sim N(\theta_k, \sigma^2)$ with defined priors and hyperpriors. I attempted to compare these models using DIC, but I was unable to obtain a realistic value for my model. I believe this may be due to the fact that the fixed γ_j values are in effect forcing the model to fit the data since the Z-values are constrained to be between the cut-points and evaluation of these Z’s is the point in the DIC algorithm where the results are incorrect. I also attempted to use a Bayesian χ^2 Goodness of Fit test to determine if the model fit the data at all, but again ran into nonsensical results in the same area of evaluation. This is a potential weakness of the model that bears looking into.

Conclusion

This analysis set out to explore factors that influence shooting ability and rank players according to that ability. However, once analysis was underway it became clear that the ability being explored was one of efficiency, not pure shooting. This is due to the decision to classify and model the data as ordinal when in fact there are relationships between misses, two point makes and three point makes that are more complex than a simple ordinal model can account for. However, this model was able to represent a metric of efficiency, which was taken to mean a player’s ability to produce three point shots over two point shots, and both over missed shots. If shooting ability was going to be modeled, a nominal model that differentiated misses into two point attempts and three point attempts would be better able to bear out worthwhile results.

Despite this unanticipated deviation, interesting, interpretable results that make intuitive sense to those well-versed in the NBA were still obtained. It makes sense that, for example, cold-blooded Kobe is less affected by road games and LeBron is closely guarded on his many crashes down the lane. There is plenty of room for future studies in player efficiency, especially for those that are able to include additional factors, like spatial data, fouls drawn and free throws (the exclusion of which certainly hurt James Harden here), and other commonplace statistical categories like rebounds and assists.

Appendix

Convergence

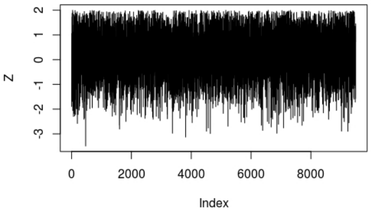

Figure 9

atte

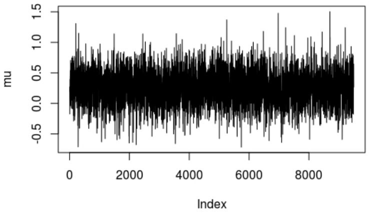

Figure 10

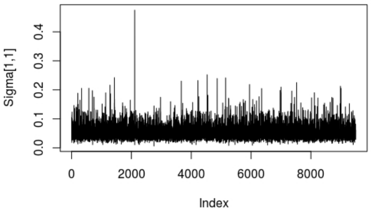

Figure 11

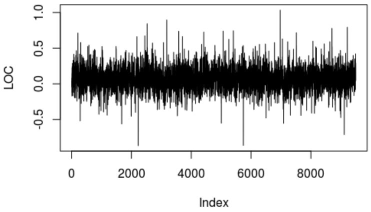

Figure 12

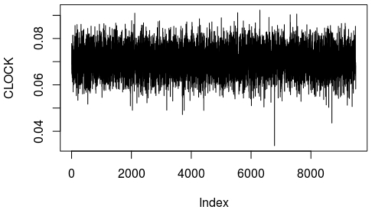

Figure 13

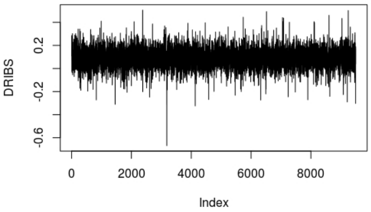

Figure 14

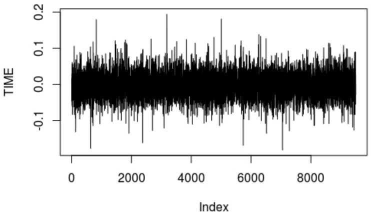

Figure 15

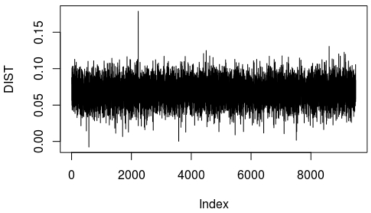

Figure 16

Complete Conditionals

$$\boldsymbol{\beta}_k|* \sim MVN(\Sigma^*(\Sigma^{-1}\mu + X_k'Z_k), (\Sigma^{-1} + X_k'X_k)^{-1})$$

$$\boldsymbol{\mu}|* \sim MVN(V^*(V^{-1}\mu + \Sigma^{-1}\bar{\boldsymbol{\beta}}_k), (V^{-1} + \Sigma^{-1})^{-1})$$

$$\boldsymbol{\Sigma}|* \sim InverseWishart(w + n_{players}, (I + \sum_{k=1}^{n_{players}} (\beta_k - \mu)(\beta_k - \mu)')^{-1})$$

$$\mathbf{Z}_{ik}|*, Y_{ik} = j \sim \mathcal{TN}(\mathbf{X}_k'\boldsymbol{\beta}_k, 1, \gamma_{j-1}, \gamma_j)$$

Code

```
y<-shots$PTS

X.list <- list()
for(i in 1:n.players) {
  X.list[[i]] <- shots[shots$player_name==unique(shots$player_name)[i],
  c(”LOCATION”,”SHOT_CLOCK”,”DRIBBLES”,”TOUCH_TIME”,”CLOSE_DEF_DIST”)]
}

#X.list has n.players # of matricies, each with a different n # of rows and 5 columns
#n x 5 x n.players

beta<-matrix(0,ncol(X.list[[1]]),nrow=n.players)
beta.save<-array(dim = c(n.players,ncol(X.list[[1]]),(burn+length)))
beta.save[,,1]<-beta

mu<-matrix(0,ncol=ncol(X.list[[1]]),nrow=(burn+length))
mu[1,]<-rep(0,ncol(X.list[[1]]))

m<-rep(0,ncol(X.list[[1]]))
V<-1*diag(ncol(X.list[[1]]))

w<-1+n.players
I<-diag(ncol(X.list[[1]]))
Sigma<-array(dim = c(ncol(X.list[[1]]),ncol(X.list[[1]]),(burn+length)))
Sigma[,,1]<-I

Z.list <- list()
for(i in 1:n.players) {
  Z.list[[i]] <- matrix(0,nrow=(length+burn),ncol=n.player.shots[i])
}
#length+burn x n.player.shots x n.players

gamma<-c(2,3)

devs <- matrix(0,ncol=n.players,nrow=length+burn)

for(d in 2:(length+burn)){
  # update beta
  for (k in 1:n.players){
    sigstar <- solve( solve(Sigma[,,(d-1)]) + t(X.list[[k]])%*%as.matrix(X.list[[k]]) )
    mustar <- sigstar %*% (solve(Sigma[,,(d-1)])%*%mu[(d-1),] + t(X.list[[k]])%*%Z.list[[k]][
    beta.save[k,,d] <- mvrnorm(1,mustar,sigstar)
  }

  # update mu
  Vstar <- solve( solve(V) +  solve(Sigma[,,(d-1)]) )
  mstar <- Vstar %*% (solve(V)%*%m + solve(Sigma[,,(d-1)])%*%colMeans(beta.save[,,d]))
  mu[d,] <- mvrnorm(1,mstar,Vstar)

  # update Sigma
  wstar<-w+n.players
  S.mu<-matrix(0,nrow = ncol(X.list[[1]]), ncol = ncol(X.list[[1]]))
  for(k in 1:n.players){
    S.mu<-S.mu+(beta.save[k,,d]-mu[d,])%*%t((beta.save[k,,d]-mu[d,]))
  }
  Istar<-solve(I+S.mu)
  Sigma[,,d] <- riwish(wstar,Istar)

  # update Z
  index <- 1
  for(k in 1:n.players){
    for (i in 1:n.player.shots[k]){
      Z.list[[k]][d,i] <-rtnorm(1,t(X.list[[k]][i,])%*%beta.save[k,,d],1,gamma[2],Inf)
      if(y[index]==2)
      {
        Z.list[[k]][d,i] <-rtnorm(1,t(X.list[[k]][i,])%*%beta.save[k,,d],1,gamma[1],gamma[2])
      }
      if(y[index]==0)
      {
        Z.list[[k]][d,i] <-rtnorm(1,t(X.list[[k]][i,])%*%beta.save[k,,d],1,-Inf,gamma[1])
      }
      index <- index+1
    }
  }
  index <- 1
  for(k in 1:n.players) {
    p <- hist(Z.list[[k]][d,],plot = F,breaks=c(-Inf,2,3,Inf))$count/n.player.shots[[k]]
    devs[d,k] <- -2*loglike(y[index:(index+n.player.shots[k]-1)],p)
    index <- index + n.player.shots[k]
  }
}

loglike <- function(y,p) {
  counts <- c(sum(y[which(y==0)]),sum(y[which(y==2)]),sum(y[which(y==3)]))
  sum(dmultinom(counts,prob=p,log=TRUE))
}

#attempt at dic
dbar <- mean(apply(devs[-c(1:burn),],1,sum))
dthetabar <- numeric()
index <- 1
for(k in 1:n.players) {
  p <- hist(apply(Z.list[[k]],2,mean),plot = F,breaks=c(-Inf,2,3,Inf))$count/n.player.shots[[
  dthetabar[k] <- -2*loglike(y[index:(index+n.player.shots[k]-1)],p)
  index <- index + n.player.shots[k]
}
psubd <- dbar-sum(dthetabar)

#Chi Square goodness of fit
bincnts<-0
nbin<-floor(n.player.shots^0.4)
quants<-matrix(0,nrow = n.players,ncol = 4)
index <-1
for(k in 1:n.players){
  quants[k,]<-c(0,as.numeric(cumsum(table(y[index:(index+n.player.shots[k]-1)])/n.player.shot
  index <- index + n.player.shots[k]
}
BX2<-matrix(0,nrow = (length+burn),ncol = n.players)

for(d in 1:(length+burn)){
  index <- 1
  for(k in 1:n.players){
    cdf<-numeric(n.player.shots[k])
    for(i in 1:n.player.shots[k]){
      cdf[i]<-pnorm(0,t(X.list[[k]][i,])%*%beta.save[k,,(d+burn)],1)
      if(y[index]==2){
        cdf[i]<-pnorm(1,t(X.list[[k]][i,])%*%beta.save[k,,(d+burn)],1)
      }
      if(y[index]==3){
        cdf[i]<-1
      }
      index <- index + 1
    }
    bincnts<-hist(cdf,plot=F,breaks=quants[k,],na.rm=T)$count
    n<-sum(bincnts)
    BX2[d,k]<-sum((bincnts-n/nbin[k])^2/(n/nbin[k]))
  }
}

#post pred
z.pred<-matrix(0,nrow = length,ncol = n.players)

X.samp<-array(dim = c(n.players,5,length))
for(j in 1:5){
  for(k in 1:n.players){
    X.samp[k,j,]<-sample(X.list[[k]][,j],length,replace = TRUE)
  }
}

for(k in 1:n.players){
  index<-sample((burn+1):M,length,replace = T)
  for(d in 1:length){
    z.pred[d,k]<-rnorm(1,t(X.samp[k,,d])%*%beta.save[k,,index[d]],1)
  }
}

prob.pred<-matrix(0,nrow = n.players,ncol = 3)

for(k in 1:n.players){
  prob.pred[k,]<-hist(z.pred[,k],plot = F,breaks=c(-Inf,2,3,Inf))$count/length
}

pred.avg.score<-numeric(9)
values<-matrix(c(0,2,3),ncol = 1)
pred.avg.score<-(prob.pred%*%values)
```