100m Ranking Model

Tyrel Stokes

20/02/2022

This is a companion document to a presentation I gave at the IOC conference on injury and prevention. It was a workshop trying to think about how to model performance data in the Olympic context. I chose the 100m as an example because it has many nice common features for olympic sports and it is well known and understood. See the slides at the following link.

The model I use is a version of the so called plackett-luce which is a convenient model for rank data. It allows us to take into account the strength of competition which is extremely important in these kind of sports, particularly with the tournament structure which guarantees stronger athletes are likely to compete against each other more ofter as the events go on. The other thing that is nice is it models the joint structure of all the ranks and not just the probability of getting first or some other dichotimized event like a medal.

Here is the likelihood.

$$P(Y_1 > Y_2 > \cdots > Y_p) = \prod_{i=1}^{P} \frac{exp(\beta_i/\sigma)}{\sum_{j=i}^{P} exp(\beta_j/\sigma)}$$

Each player that is being ranked has a parameter β_i. The larger it is the more likely they are to place high. There is also a noise parameter σ. The smaller this is, the more deterministic the outcome is assumed to be. That is the better players are increasingly likely to win out. In many applications it is assumed to be 1 or fit with some hyper-parameter searching procedure. One thing that is useful to notice is that it is really difficult to separate this parameter from the general scale and how much variability we might expect in the β's. We need to be careful and make smart choices with our priors to not run into identifiability issues.

A little hitch I added was allowing the noise parameter to vary by type of race. In the 100m, there are heats, semis, and finals. The assumption I have is that in the heats there is more likely an upset. This is because racers are playing optimal strategy over the tournament, which means it is often in the top athletes best strategy to try and win while expending the least energy possible. We might expect then some results that are not because someone is better but because players are using different strategies (favorites easing up, underdogs going for broke) or miscalculations in this strategy.

So we let the noise parameter be larger in less important races. Implicitly this weights those results slightly less, but without us having to explicitly choose the weighting!

So I let

$$\sigma = [\sigma_{finals}, \sigma_{semis}, \sigma_{heats}]$$

where I impose

$$\sigma_{finals} = 1 \leq \sigma_{semis} \leq \sigma_{heats}$$

This is achieved using the positive ordered vectors in stan.

I set weakly informative priors over (β, σ).

The data in this git repo is a cleaned version of all races in the 2011 World Championship + 2012 Olympics. It seems to behave reasonably well given how little data is here. Feel free to find more data and add it or adapt this model to a different context. I haven't seen many plackett-luce models in stan so hopefully this is useful to someone.

The model includes code to take random samples from the posterior predictive of a plackett-luce model. I use this to re-simulate the 2012 Olympic finals.

Data and Model Fitting

```
######################
## Load the data list

race_list <- readRDS("race_data_list.Rds")

stan_dt <- race_list[[1]]

x <- stan_dt$x
x <- apply(x,2,as.integer)

stan_dt$x <- x

stan_dt$n_types <- 3

stan_dt$type <- plyr::mapvalues(stan_dt$type,from = c(1:3),to = c(3:1))

rank_data <- race_list[[2]]

athletes <- race_list[[3]]

fints <- stan_dt$finals
```

Now we fit the model

```
#########################

library(cmdstanr)
```

```
## This is cmdstanr version 0.4.0.9000
## - Online documentation and vignettes at mc-stan.org/cmdstanr
## - CmdStan path set to: C:/Users/tyrel/Documents/.cmdstan/cmdstan-2.29.0
## - Use set_cmdstan_path() to change the path
```

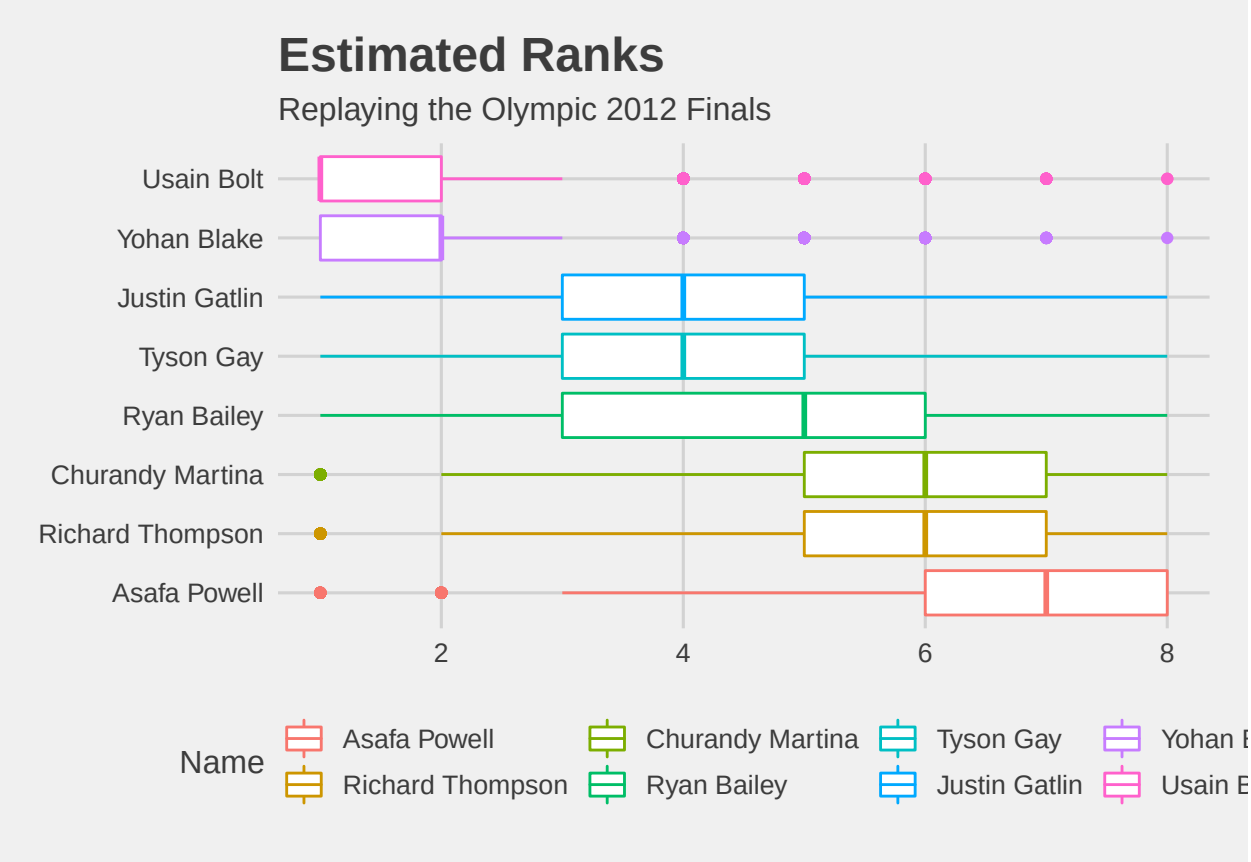
```
mod <- cmdstan_model("plackett_luce.stan")


fit <- mod$sample(data = stan_dt,iter_warmup = 2000, iter_sampling = 2000,parallel_chains = 4)


############################
################################
```

Now a recreation of some of the plots seen in the slides.

```
##############################
################################

### Collect information and make some plots


cv <- fit$summary()


chc2 <- fit$summary("posterior_latent_ranks")

post_data <- data.frame(Name = athletes,rank = chc2$mean, Rl = chc2$q5,
                        Ru = chc2$q95)

fdata <- post_data[fints,] ## Rank information for just those in the finals


########################################
########################################

### This gets you the finals data

finals_draws <- fit$draws("replay_ranking", format = "df")

# Reformat this to work well with ggplot2
library(foreach)

fdata <- foreach(i = 1:8, .combine = rbind)%do%{

 out <-  data.frame(Name = athletes[fints[i]], Rank = as.vector(finals_draws[,i]))

  names(out)[2] <- "Rank"
  out
}
```

Make plots

```
library(ggplot2)
library(ggthemes)

fdata$Name <-  factor(fdata$Name,levels = rev(athletes[fints]))

ggplot(fdata,aes(x= Name,y= Rank))+ geom_boxplot(aes(color = Name)) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Estimated Ranks",
       subtitle="Replaying the Olympic 2012 Finals",
       x="Finalist Name",
       y="Rank") + coord_flip() +theme_fivethirtyeight()
```



Here we see that Bolt and Blake are heavily favored. Looking at that same information in perhaps a more informative way gets us this heat table.

```
############################################################
############################################################
## Finals heat map rank
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
dd <- tabyl(fdata,Name, Rank)

rtab <- dd
rtab[,2:9] <- rtab[,2:9]/(nrow(fdata)/8)

rtab2 <- data.frame(Real_Result = c(8,6,3,7,5,4,1,2),rtab[,1:9])
names(rtab2)[3:ncol(rtab2)] <- c(1:8)

rtab2$Medal <- apply(rtab2[,3:5],1,sum)

names(rtab2)[1] <- "Actual Result"

library(gt)
```

```
library(scales)
library(readr)
```

```
##
## Attaching package: 'readr'
```

```
## The following object is masked from 'package:scales':
##
##     col_factor
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
rtab2 %>% arrange(`Actual Result`) %>% mutate_if(is.numeric,~round(.,2))%>% gt() %>%
  tab_spanner(label = "Predicted Ranks",
              columns = c(3:10) )%>%
  data_color(columns = 3:10,
             colors = col_numeric(palette = c("white","firebrick"),
                                  domain = c(0,1))) %>%
  data_color(columns = 11,
             colors = col_numeric(palette = c("white","gold"),
                                  domain = c(0,1)))
```

		Predicted Ranks								
Actual Result	Name	1	2	3	4	5	6	7	8	Medal
1	Yohan Blake	0.35	0.42	0.15	0.06	0.02	0.01	0.00	0.00	0.92
2	Usain Bolt	0.54	0.30	0.11	0.04	0.01	0.01	0.00	0.00	0.94
3	Churandy Martina	0.01	0.01	0.05	0.08	0.14	0.22	0.26	0.24	0.07
4	Justin Gatlin	0.03	0.08	0.20	0.24	0.21	0.13	0.08	0.03	0.31
5	Tyson Gay	0.05	0.10	0.24	0.23	0.18	0.11	0.06	0.03	0.39
6	Richard Thompson	0.00	0.01	0.05	0.08	0.14	0.23	0.27	0.21	0.07
7	Ryan Bailey	0.03	0.06	0.17	0.21	0.22	0.16	0.10	0.05	0.26
8	Asafa Powell	0.00	0.01	0.03	0.05	0.09	0.14	0.22	0.45	0.04

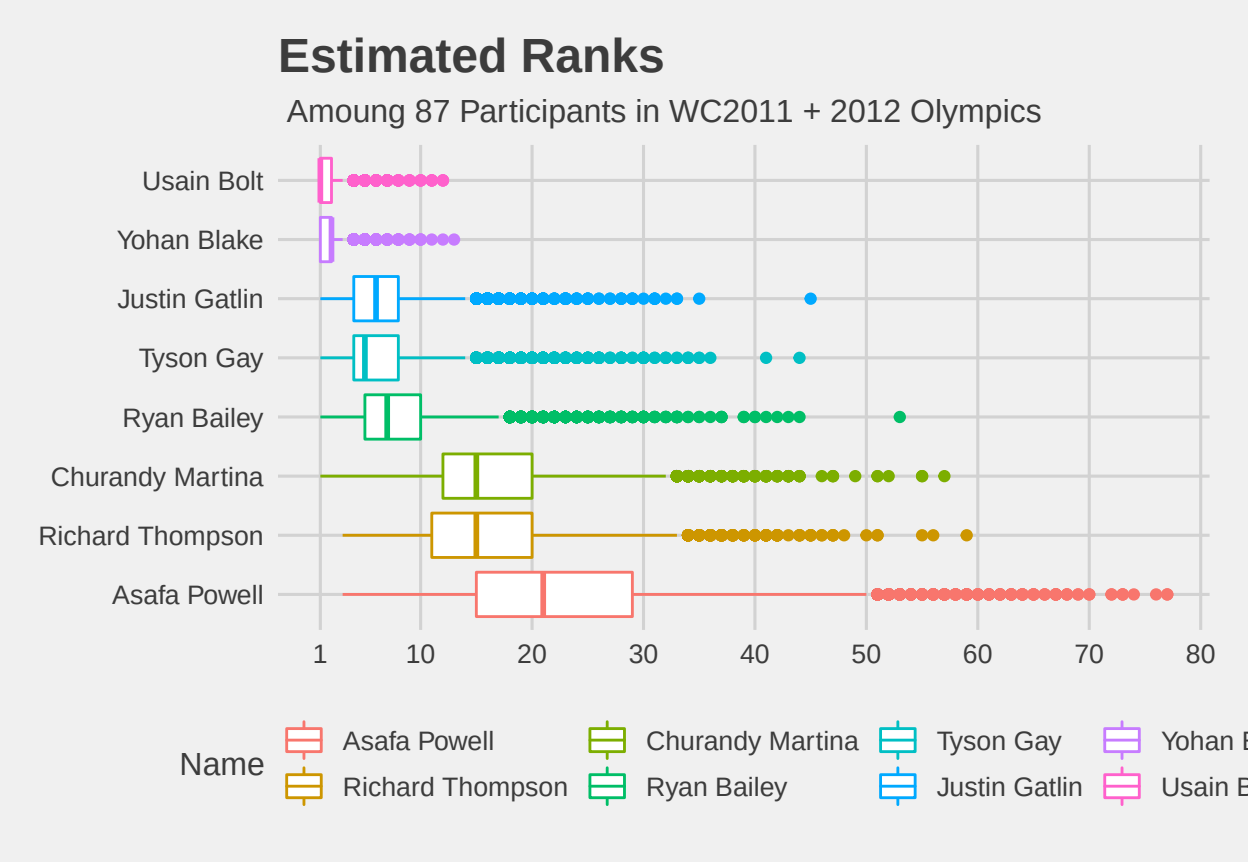
We see both Bolt and Blake are over 90% likely to get a medal according to the model. I looked up the betting odds for gold and the market seemed to suggest something in the 50-60% range chance of Bolt getting gold at the time. Not bad for a model with only 22 total races from 2 events.

You might remember that Bolt was disqualified from the 2011 final, so bolt only has 5 races in this data set and yet the model really strongly prefers him.

The stan file could be easily modified to make any other athletes in the data race and simulate the results. With more work, one could simulate entire tournaments. I wish I had to time to do this, hopefully one day.

Here is another plot where we look at the latent ranks the model gives for the 8 finalists compared to all 87 athletes who ran at least one race in either event.

```r
## This gets you the ranking plot amoung all participants
overall_rank_draws <- fit$draws("posterior_latent_ranks", format = "df")


odata <- foreach(i = 1:8, .combine = rbind)%do%{

  out <-  data.frame(Name = athletes[fints[i]], Rank = as.vector(overall_rank_draws[,fints[i]]))

  names(out)[2] <- "Rank"
  out$mrank <- mean(out$Rank)
  out
}

odata$Name <- factor(odata$Name,levels = rev(athletes[fints]))

odata %>%  ggplot(aes(x= Name,y= Rank))+ geom_boxplot(aes(color = Name)) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Estimated Ranks",
       subtitle=" Amoung 87 Participants in WC2011 + 2012 Olympics",
       x="Finalist Name",
       y="Rank") + coord_flip() +theme_fivethirtyeight()+scale_y_continuous(breaks = c(1,10,20,30,40,50
```



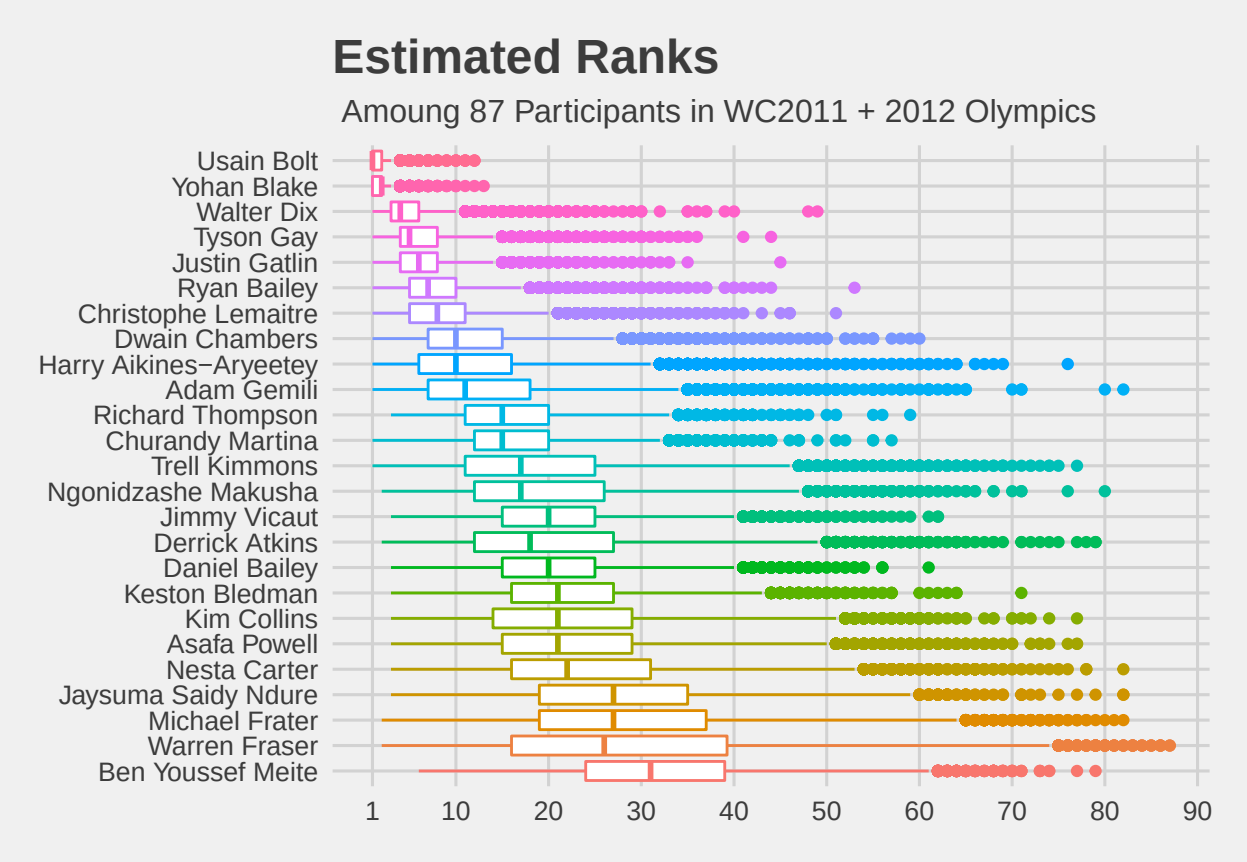
Usain and Blake are strongly considered 1 and 2 respectively by the model. Notice also that the top 5 in the event (before Tyson Gay's suspension) are all for sure top 8 runners. The other 3 in the finals, not so much.

Here we can make a massive graph with all of the players.

```
## This gets you the ranking plot amoung all participants



odata2 <- foreach(i = 1:length(athletes), .combine = rbind)%do%{

  out <-  data.frame(Name = athletes[i], Rank = as.vector(overall_rank_draws[,i]))

  names(out)[2] <- "Rank"
  out
}


mrank <- odata2 %>% group_by(Name) %>% summarise(mrank = mean(Rank))

mrank <- mrank %>% arrange(mrank)



odata2$Name <- factor(odata2$Name,levels = rev(mrank$Name))
```

Take the top n number of players and plot

```
top_n <- 25

odata2 %>% filter(Name %in% mrank$Name[1:top_n]) %>% ggplot(aes(x= Name,y= Rank))+ geom_boxplot(aes(col
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Estimated Ranks",
       subtitle=" Amoung 87 Participants in WC2011 + 2012 Olympics",
       x="Finalist Name",
       y="Rank") + coord_flip() +theme_fivethirtyeight()+scale_y_continuous(breaks = c(1,10,20,30,40,50
  theme(legend.position="none")
```


Estimated Ranks
Amoung 87 Participants in WC2011 + 2012 Olympics
Usain Bolt
Yohan Blake
Walter Dix
Tyson Gay
Justin Gatlin
Ryan Bailey
Christophe Lemaitre
Dwain Chambers
Harry Aikines−Aryeetey
Adam Gemili
Richard Thompson
Churandy Martina
Trell Kimmons
Ngonidzashe Makusha
Jimmy Vicaut
Derrick Atkins
Daniel Bailey
Keston Bledman
Kim Collins
Asafa Powell
Nesta Carter
Jaysuma Saidy Ndure
Michael Frater
Warren Fraser
Ben Youssef Meite
1
10
20
30
40
50
60
70
80
90