**Maximum Likelihood Estimation of Home Court Advantage**

From time to time, I like to write with the emphasis being on the mathematics more so than on the sports. Today I am going to focus on a powerful statistical technique and show how I used it to quantify the NBA home court advantage. Our analysis will reveal what the league-wide home court advantage actually is as well as assigning a home-court advantage to each individual team.

This article is written with quite a bit of emphasis on the mathematics involved. If you are like me and passionate about the techniques, then please read every line of this article. Otherwise, all my conclusions are contained after the section titled ‘Home Court Advantage in the NBA’ and you may safely skip ahead.

**What are Parameters?**

To understand the technique I am using, we need to discuss a bit about what exactly a parameter. A parameter is a number that governs the randomness in a situation. Here are two examples:

* You have a weighted coin that comes up heads with probability 60%
* We run a store where customers arrive at random but at an average rate of 20/hour

The two above settings each involve a parameter for a probability distribution. In the first case, the parameter is the 60%; the 60% is what determines the random behavior the coin exhibits. A coin with 60% heads isn’t going to generate the same heads/tails sequence as a coin that is 50% heads.

The second situation also involves a parameter, the average arrival rate. In any given hour, the number of customers who arrive is going to behave like a random number. It can’t be predicted. However, the fact that the average arrival rate tells us we are more likely to see a sequence of 18, 22, and 27 customers during a three hour stretch than the sequence 115,120,119. This parameter governs the randomness.

**Problem: Parameters are Often Unknown**

Often we don’t know the precise value of a parameter. Information isn’t always given to you in a neat, condensed textbook problem. But, nonetheless, we may want to know the parameter’s value. All that is left to us is to estimate the value from some observed data.

Take the example of flipping a weighted coin. We know the coin is weighted, but we don’t know how often it will come up heads. If we flip it 20 times and observe:

* 10 heads & 10 tails then our best guess should be that the coin is heads 50% of the time
* 15 heads & 5 tails then our best guess should be that the coin is heads 75% of the time

This seems like common sense. Use what the coin says to inform your guess, or estimate, of the unknown parameter. In mathematics, though, we formalize everything. We start with a simple idea and distill the important characteristics in order to be able to extend ideas into murkier situations.

**Maximum Likelihood Estimation**

The technique of maximum likelihood estimation is a generic method to estimate unknown parameters from observed data. The idea is simple: if we *knew* the value of the parameter, we could compute the probability of observing our sample:

* If a coin is 50% heads, then observing two heads out of four happens 37.5% of the time
* If a coin is 70% heads, then observing two heads out of four happens 26.5% of the time
* If a coin is 45% heads, then observing two heads out of four happens 36.8% of the time.

In fact, observing two heads out of four flips is most likely provided the coin is fair. I plotted below the probability of observing two heads out of four for the entire range of possible heads probabilities. Notice the peak is at exactly the coin being 50% heads.

Then, estimating the coin to come up heads 50% of the time is the same as picking the parameter that maximizes the probability of our sample. To put this another way, we pick the parameter value that best explains the data we have observed.

**The Problem with Estimating Home Court Advantage another Way**

I am nowhere near the first person to study the amount of home court advantage (HCA) in various sports. I have observed two other categories of ways people estimate home court advantage:

* Aggregate over every game ever and define home court advantage to be the average margin of victory of the home team.
* For each team, [look at their average scoring margin at home and on the road](https://kenpom.com/blog/how-to-measure-home-court-advantage/). This way you get each team’s specific HCA

(Note: If you average each team’s HCA in the second case, you will get the result of the first case. That is, these methods aren’t actually different mathematically).

I can argue that, at least in college basketball, both these ideas are flawed and suffer from the same crucial issue. Ask yourself: what does the beginning of a basketball season look like?

At the beginning of the 2019-2020 season, Kentucky played Georgetown College, Kentucky State, Eastern Kentucky University, and University of Evansville all at home. (Sorry big blue fans, I will never let you forget the University of Evansville loss). Good college basketball teams start the season playing many inferior opponents at home. Bad college basketball teams start the season playing superior opponents on the road.

What I am trying to say is that the quality of opponent at home and on the road is often very different. Therefore, comparing margin of victory at home and on the road isn’t a fair comparison. Because better teams play at home more often, the home-court advantage by this method is over exaggerated.

**Home Court Advantage with Maximum Likelihood Estimation.**

If we are going to perform maximum likelihood estimation, we need to model professional sports using a random setting. Our [ensemble ratings](https://thedatajocks.com/introducing-ensemble-ratings/) give us a way to predict margin of victory on a neutral court in the NBA. Alternatively, if you had any other method to predict margins of victory in every game in the NBA, you could use that here.

Then, for each game we observe the variable **home team margin of victory**. From our ensemble ratings we observe **predicted neutral court margin of victory for the home team.** The difference between these two acts as a piece of evidence for the true value of that team’s home court advantage.

Even if our favorite team loses by 30, if we predicted that they would lose by 35 points on a neutral court, then our guys out performed expectations to the tune of 5 points. This five points can be attributed to the effect of playing at home.

But wait!

That’s just one game, how can we be sure those 5 points are a home-court effect? And one more thing, just because on average we expect to lose by 35 there is going to be natural variance in the margin of victory. That 30 could have happened on a neutral court or on the road.

All true.

But, due to the nature of statistics, the more games we include, the surer we can be that this effect is attributed to the home-court than natural variance. If you perform 5 points better at home on average over the course of 200 games, it would be extremely unlikely for this to a normal deviation. The only explanation for the increased performance at home is a home-court advantage.

**How does MLE come into this?**

For each team, their declared home court advantage is the **parameter value** that best explains the observed difference between our predicted neutral court values and the team’s actual winning margins at home. There is a statistical model, a quadratic optimization, and derivatives floating around behind all this, but the details aren’t the focus. The point is, each of the computed values is the number that best explains a team’s performance on the road and at home even while taking into the quality of opponent in each game.

**A Note for the Detail Oriented or Mathematically Inclined**

If you’re like me and the specific mathematical details are the interesting part, then this section is written for you. Otherwise, my conclusions and findings are entirely contained in the following sections.

Above, I spoke about ensemble ratings predicting the margin of victory on a neutral court. The way we predict margins of victory are also using maximum likelihood to assign each team a rating and the difference between two team’s ratings is the predicted margin of victory on a neutral court.

However, I don’t have any data about games played on a neutral court [yet, the NBA restart in Orlando is the only time I can remember where this is going to happen]. So, I have to be able to estimate neutral court margins given games where a team has a home-court advantage. It seems as if I need to know HCA before computing a team’s rating in the first place.

It turns out that in the model I use, I have to simultaneously estimate home court advantage and team ratings. Essentially, I am solving a system of 60 equations in 60 variables (2 variables per team). The bad part though: the resulting matrix is singular, it isn’t invertible. You can’t uniquely solve a system of equations that aren’t invertible. This is because we only ever observe differences between team’s ratings and shifting every team’s rating by the same amount results in the same predicted margin.

There are two ways to overcome this. The first is to add an extra row to your matrix forcing the average rating to be whatever value you like. The second is to just use the Moore-Penrose (pseudo)inverse to compute *a* solution to the system.

**Home Court Advantage in the NBA**

Traditionally, home court advantage has been quoted as being anywhere from 2 to 5 points using various methods. While different methods will give you different numbers, for us the optimal number to assign to the home court advantage league wide is….

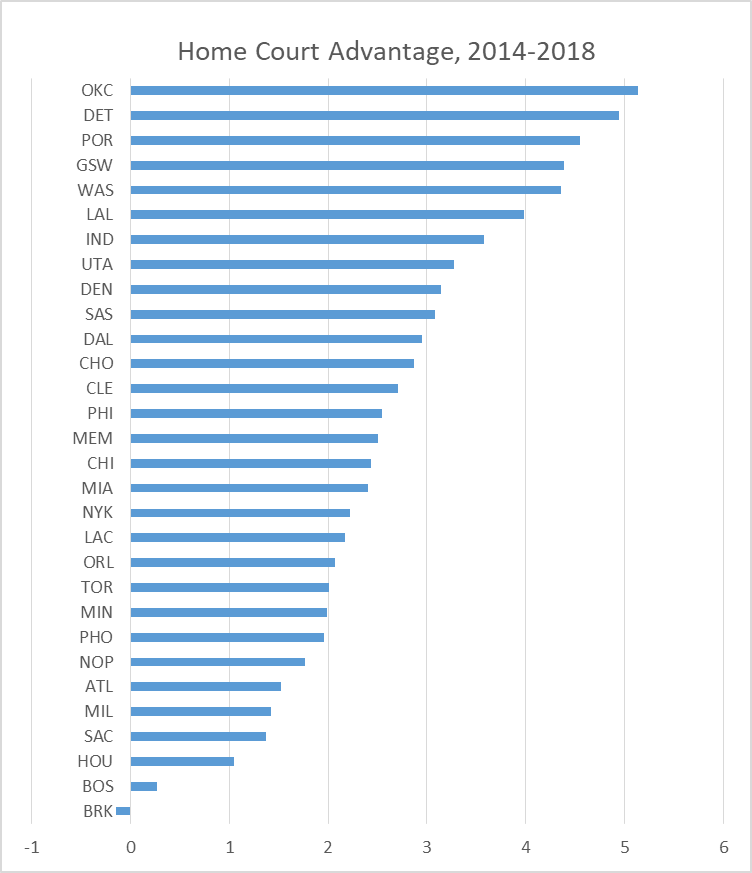
2.6 points. This is probably quite a bit lower than you might expect. It might even seem insignificant. But, it is close to the difference between the Nuggets and the Trailblazers. It is roughly comparable to the point differential we assigned to explain how much harder the [West was than the East](https://thedatajocks.com/how-imbalanced-are-the-nba-conferences/).

One last way to view this 2.6 number. If two evenly matched teams play on a neutral court, they each win about 50% of the time. A home court advantage of 2.6 points gives the home team a 60% chance of winning against an evenly matched opponent.

**Individual Team Home Court Advantage**

Perhaps more interesting is the different home court advantages for different teams across the NBA. The 2,6 number above was the league average. Some teams have much different numbers. If you compute the home court advantage for Philadelphia in the 2019-2020 season, it comes out to about 13 points of difference. That is insane and certainly the Philadelphia fans are not worth 13 points per game.

So, to try to assign home court advantages to different teams we looked at four years of data. The numbers we show below are the home court advantages that best explain the home/away differences in performance for each team between 2014 and 2018. Here is the chart.



Breaking it down like this gives me some reservations. Thought Oklahoma City and Golden State were well-known to have an excellent home atmosphere during this timeframe, Detroit was not. I actually attribute the surprise-factor of this chart to two things.

First, perhaps we are dead wrong about which team’s home environments are the hardest for opposing teams. Perhaps Detroit fans really do add almost five points of value to their team. Perhaps Brooklyn would rather play with nobody in the building.

Second, perhaps it takes more than four years of data to actually reveal who has the best home-court. After all, four seasons is only 160 home games for each team. Typically I like to run at least 1000 simulations to minimize the variance. However, 160 is still a fairly large sample size.

I think it is more likely that the common consensus for the dominant home courts is dead wrong.