A functional data approach to model score difference process in professional basketball games

Ting-Hung Yu

University of Iowa

ting-hung-yu@uiowa.edu

Nov. 8, 2023

1. Introduction
2. Estimation
3. Application
4. Conclusion

Motivation

1. In the p.12 of the first lecture note, Dr. Z mentioned that "...*A functional principal component analysis to uncover the dominant pattern of curves for a given team in a given season might be interesting*..."
2. I like playing basketball and watching some NBA games.
3. Applying functional data analysis (FDA), aka a fancy stat approach, to understand relationship between momentum and home court advantage.
4. Found the paper:
   Tao Chen & Qingliang Fan (2016): A functional data approach to model score difference process in professional basketball games, Journal of Applied Statistics, DOI:10.1080/02664763.2016.1268106

Data description

1. 8211 NBA regular season games from the period of 1 November 2002 to 28 November 2014.
2. Collected from game logs at espn.go.com using data scraping techniques.
3. Construct the data frame of score changes using 30 sec as the time unit. (i.e. a 48-min game has 96 units.)
4. Why 30 sec?
   "*...the average total score (home team plus away team) in 48 minutes of regulation is 194.4 (194.2 if we include the regulation period of the games with overtimes). Therefore, the per 30-second score is 2, which is approximately the expected value of each scoring event. ...*"

$$X_j(t) = \texttt{HomeTeamScore}_j(t) - \texttt{AwayTeamScore}_j(t),$$

with $t = 0, 1/96, 2/96, ..., 1,\ j = 1, 2, ..., 8211$

Score difference and home court advantage

1. Using score difference process between game opponents (home and away teams) to uncover the home court advantage.
2. NBA games are more frequent in having scoring events (than soccer, ice hockey, or American football...).
3. Abstract away from
   - game-specific factors: assume the change of score difference itself has reflected all those factors' effects in aggregate. (distribution for score increments in different time units)
   - the specific possessions: i.e. it is independent of the per-possession structure of basketball games. Eg. jump ball.

Why FDA?

- Path of score difference could be viewed as continuous due to the relative high frequency.
- Some previous methods, eq. Brownian motion (Stern, 1994 JASA) and Random walk (Gabel and Redner, 2010 JQAS), ignore the dependence among score increments.
- FDA allows to capture the momentum effect in games.
- FDA method restores continuous sampling which ensures further inference based on FDA estimates does not suffer from the inconsistency issue: the model is motivated in a continuous time framework while the data are inevitably sampled discretely, we might obtain inconsistent estimators if we ignore this difference. (Ait-Sahalia, 2002 Econometrica)

FDA modelling

$$X_j(t) = Y_j(t) + \epsilon_j(t)$$

where

- $\epsilon_j(t)$'s are random disturbances with $E(\epsilon_j(t)) = 0$ and some dependence structures across j and t.
- $Y_j(t)$'s are assumed to be smooth and pass through the origin. One way to approximate it is:

$$Y_j(t) \approx \sum_{k=1}^{K} c_{j,k}\phi_k(t) = C_{j,K}^T \Phi_K(t)$$

where K is a positive integer, and $\phi_k(t)$'s are B-spline basis functions.

FDA estimation – objective function

Objective function $m(\cdot)$:

$$m(C_{1,K},\dots,C_{n,K},\lambda,K) \triangleq \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{T}\sum_{i=1}^{T}[Y_j(t_i)-X_j(t_i)]^2+\lambda\int_0^1\left(\frac{d^2}{dt^2}Y_j(t)\right)^2 dt\right)$$

$$\approx \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{T}\sum_{i=1}^{T}[C_{j,K}^T\Phi_K(t)-X_j(t_i)]^2+\lambda C_{j,K}^T R_K C_{j,K}\right),$$

where

$$R_K=\int_0^1\left(\frac{d^2}{dt^2}\Phi_K(t)\right)\left(\frac{d^2}{dt^2}\Phi_K(t)\right)^T dt$$

FDA estimation – Optimization

1. Goal: find the minimizer $(C^*_{1,K}, ..., C^*_{n,K}, \lambda^*, K^*)$ of $m(\cdot)$.
2. When fixed (λ, K), we have

$$\hat{C}_{j,K} = \Big[\frac{1}{T}\sum_{i=1}^{T}\Phi_K(t_i)\Phi_K^T(t_i) + \lambda R_K\Big]^{-1}\Big(\frac{1}{T}\sum_{i=1}^{T}\Phi_K(t_i)X_j(t_i)\Big)$$

   for each j.
3. Grid search over the pair (λ, K):
   - range of K: integers from 5 to 200.
   - range of λ: a sequence of length 200 from 10^{-20} to 0.1 equally-spaced in the sense of unit of $\log_{10}(\lambda)$.

Tunning process for (K^*, λ^*)

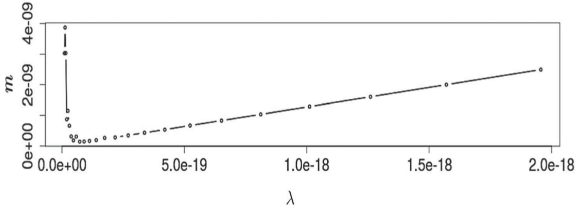


Figure 1. Optimization of the tuning parameter λ.

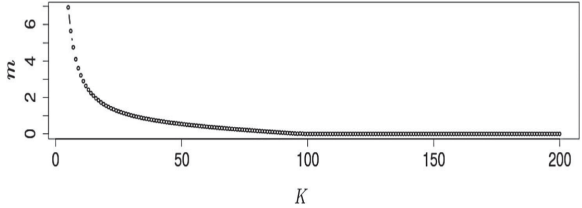


Figure 2. Optimization of K.

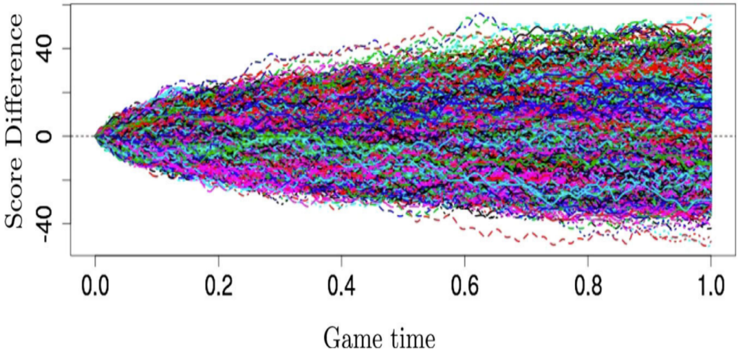
Optimized (K^*, λ^*)

Table 3. (K^*, λ^*) in various cases.

Without overtimes	With 1 overtime	With 2 overtimes	With 2+ overtimes
$(99, 7.23 \times 10^{-20})$	$(99, 7.23 \times 10^{-20})$	$(99, 7.23 \times 10^{-20})$	$(99, 7.23 \times 10^{-20})$

Fitted curve

The fitted curves for all games at (K^*, λ^*) (without overtime):



1. Introduction
2. Estimation
3. Application
4. Conclusion

Momentum

Momentum of size γ from time t to t+s with both s and $\gamma > 0$

1. Home team:

$$M(s,t,\gamma) = Y(t+s) - Y(t) > \gamma$$

2. Away team:

$$M(s,t,\gamma) = Y(t+s) - Y(t) < -\gamma$$

3. Example: if we choose t $= 0, 1/96, 2/96, 3/96, \ldots, 1-s$ with s $= 6/96$, and $\gamma = 11.5$, then a home team's momentum means there exists a 3-minute period such that the home team has the tendency to outscore the away team by at least 12 points.

Momentum and home court advantage

Games have one and only one such event where a team outscored the other by at least 12 points within a 3-minute period.

Table 4. Exactly one momentum.

Overtime(s)	0	1	2	>2
Number of games with exactly one momentum	884	952	961	962
% of games with the momentum from the home team	57	57	57	57
% of home winnings having the momentum	79	74	73	73
% of away winnings having the momentum	71	65	64	64

Momentum and home court advantage

Close games that have momentum in the second half.

Table 6. Close games, without overtime, $\gamma = 9.5$.

	Statistics
Number of games with no momentum	834
% of home winnings	52
Number of games with exactly one momentum	203
% of games with the momentum from the home team	52
% of home winnings having the momentum	60
% of away winnings having the momentum	44
% of either team with the momentum wins	52

1. With momentum, the winning percentage is 60% for 105 home teams and is 44% for 98 away teams.
2. Momentum has a big impact for the outcome of close games and home court advantage effect is strong.

Conclusion

1. FDA is a fancy tools for studying the relationship between momentum and home court advantage.
2. Momentum and home court advantage play important role in determining the outcome.
3. Could we incorporate some other interesting/important factors to study the home court advantage under FDA framework?