Modeling NBA Performance Curves

Abhijit Brahme

2024-08-10

Motivation

A key goal of both sports teams and sports journalists is understanding how player ability evolves over time. Forecasting player ability is essential for teams considering player acquisition, either via the draft, free agency or trades as well as strategic planning. Often, production curves are used to describe similarities in production across related players. *As a result, analysis of production curves could provide useful insight into player archetypes, and how various archetypes change with age* .

Production Curves in Sports

Most commonly, in a production curve analysis, a continuous measurement of aggregate skill (i.e. RAPM or VORP), denoted Y_{pt} is considered for a particular player at time t, $Y_{pt} \approx f_p(t) + \epsilon_p(t)$, where $f_p(t)$ is the average production for player at any time t and $\epsilon_p(t)$ represents the residual uncertainty about player production, typically assumed to be uncorrelated over time. Athletes not only exhibit different career trajectories, but their careers occur begin and end at different ages, can be interrupted by injuries, and include different amounts of playing time. As such, the statistical challenge in production curve analysis is to infer smooth trajectories $f_p(t)$ from sparse, irregular observations of Y_{pt} across players.

Relevant Work

1. Bayesian Hierarchical Framework
   - Hierarchical aging model to compare player abilities across different eras in three sports: hockey, golf, and baseball (Berry, Reese, and Larkey 1999)
   - Gaussian Process regressions to infer how production evolves across different basketball positions (Page, Barney, and McGuire 2013)
   - Parametric curves to describe trajectories before and after peak-performance (Vaci et al. 2019)
2. Functional Data Analysis
   - Functional principal components metrics can be used in an unsupervised fashion to identify clusters of players with similar trajectories (Wakim and Jin 2014)
   - Nearest Neighbor algorithm to characterize similarity between players (Silver 2015)
   - Each player's production curve is represented as a convex combination of curves from the same set of archetype (Vinué, Epifanio, and Alemany 2015)

Current Contribution

In this work, we propose a model for jointly inferring how multiple of athleticism and skill co-evolve over a player's career. Our model explicitly accounts for multiple sources of variability in the metrics by accounting for dependence across similar player types, dependence between metrics which measure similar aspects of latent player ability and, of course, autocorrelation in time. Further, unlike previous approaches, we give more careful consideration to the sampling distribution of observed metrics.

Data Overview

1. $\approx$ 2k NBA players from years 1997 - 2021, from the ages of 18 - 39
2. Longitudinal mixed tensor valued data $\mathcal{X}$ of size N by T by K where N is the number of players, T is the number of years in a player's career, and K are the number of production metric curves with $\mathcal{X}_{ntk}$ is missing if player n is not observed for metric k at age t.
3. Ω is binary tensor of same size as $\mathcal{X}$ indicating missingness.

Modeling Assumptions

Let $\mathcal{B}$ be the set of metric indices i associated with metrics reflecting the number of successes in n attempts, i.e. Binomial metrics, $\mathcal{R}$ correspond to count metrics which occur at a given rate for each minute played, i.e. Poisson metrics, and $\mathcal{G}$ correspond to metrics which can be characterized as approximately Gaussian.

1. $\mathcal{R} = \{$FGA, FG3A, FTA, Blocks, OREB, DREB, TOV, AST, STL, Fouls$\}$
2. $\mathcal{B} = \{$FGM, FG3M, FTM$\}$
3. $\mathcal{N} = \{$FGA, FG3A, FTA$\}$
4. $\mathcal{G} = \{$DBPM, OBPM, log(Minutes)$\}$.

$$Y_{pi}(t) \sim \begin{cases} Pois(M_{pit} e^{f_{pi}(t)}) \text{ if } i \in \mathcal{R} \\ Bin(N_{pit}, logit^{-1}(f_{pi}(t))) \text{ if } i \in \mathcal{B} \\ \mathcal{N}(f_{pi}(t), \frac{\sigma_i^2}{M_{pit}}) \text{ if } i \in \mathcal{G} \end{cases} \tag{1}$$

Modeling Assumptions (contd.)

1. Space of players live in low dimensional latent space $X \in \mathbb{R}^{NxD}$
2. For a given time t, and metric k, $f_{tk} \sim \mathcal{GP}(0, K_x)$
   - Approximation of f_{tk} is given by Random Fourier Features such that $f_{tk} \approx \phi(X)^T \beta_{tk}$ (Gundersen, Zhang, and Engelhardt 2020)
   - Inducing correlation across time t and metric k comes from inducing correlation amongst linear weights β_{tk}.
   - Separable covariance structure for time, metric, and player.

Challenges

1. MCMC Convergence (multi-modal posterior)
   - show example posterior chains
2. Identifiability (rotational invariance of model)
3. Modeling temporal and within-metric correlation
4. Missing Not At Random (MNAR) due to selection bias

Methods

In order to address identifiability issues and MCMC convergence, we propose an alternating scheme to estimate the latent space X and functional bases β_{tk}.

1. Initialize X
   - Exponential PPCA, Probabilistic Tensor Decomposition, Standard PCA, etc.
   - Make note of how different methods yielded "better" clustering / initialization in the latent space (i.e probabilistic methods accounted for sampling variance / outlier shrinkage, and make note of how each of these methods affected results)
2. Using X from above, conduct inference on β_{tk} using Hamiltonian Monte Carlo techniques
3. Fixing β_{tk} from above, conduct inference on X using HMC techniques.

Current Results

1. (plots of example player production curves)
2. (plots of the different types of curves)
3. Correlation matrix of latent curves of metrics
4. Dendrogram of latent space
5. Separation in latent space

Future Direction

1. Use model to identify causal effect of types of injuries across varying types of players.
2. Address selection bias issue.

References

Berry, Scott M, C Shane Reese, and Patrick D Larkey. 1999. "Bridging Different Eras in Sports." *Journal of the American Statistical Association* 94 (447): 661–76.

Gundersen, Gregory W., Michael Minyi Zhang, and Barbara E. Engelhardt. 2020. "Latent Variable Modeling with Random Features." https://arxiv.org/abs/2006.11145.

Page, Garritt L, Bradley J Barney, and Aaron T McGuire. 2013. "Effect of Position, Usage Rate, and Per Game Minutes Played on NBA Player Production Curves." *Journal of Quantitative Analysis in Sports* 9 (4): 337–45.

Silver, Nate. 2015. "We're Predicting the Career of Every NBA Player. Here's How." *FiveThirtyEight*. https://fivethirtyeight.com/features/how-were-predicting-NBA-player-career/; FiveThirtyEight.

Vaci, Nemanja, Dijana Cocić, Bartosz Gula, and Merim Bilalić. 2019. "Large Data and Bayesian Modeling—Aging Curves of