Modeling NBA Performance Curves

Abhijit Brahme

2024-09-20

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)

Data Overview

1. $\approx$ 2k NBA players from years 1997 - 2021, from the ages of 18 - 39
2. Longitudinal mixed tensor valued data $\mathcal{Y}$ 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 $\mathcal{Y}_{ntk}$ is missing if player n is not observed for metric k at age t.
   - Non-missing entries are observations from exponential families (i.e Binomial, Gaussian, Exponential, Poisson, etc.)
3. Ω is binary tensor of same size as $\mathcal{Y}$ indicating missingness.

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.

Modeling Assumptions

1. Space of players live in low dimensional latent space $X \in \mathbb{R}^{N \times D}$
2. For a given time t, and metric k, $f_{tk} \sim \mathcal{GP}(0, K_X)$ is a vector of size N, with K_X capturing correlation between players
    - Approximation of f_{tk} is given by Random Fourier Features such that $f_{tk} \approx Z(X)^T \beta_{tk}$ (Gundersen, Zhang, and Engelhardt 2020)
    - Inducing correlation across time t and metric k comes from inducing correlation amongst linear weights β_{tk}.
    - We assume a separable covariance structure for time, metric, and player.

Random Fourier Features (TL;DR)

1. Approximation of Gaussian Process can be turned into a linear operation, $f_{tk}(X) \approx Z(X)^T \beta_{tk}$
2. Number of random features, R, determines how good the approximation is
3. Choice of $p(\omega)$ determines covariance function of the Gaussian Process

Model Parameters

1. $X \sim \mathcal{N}(\mu_0, \Sigma_0)$
2. $\sigma_k \sim IG(1,1) \forall k \in \mathcal{G}$
    - Variance term for normally distributed observations
3. $\omega_r \sim \mathcal{N}_D(0, I_d)$
    - random feature map approximation
    - $Z(X) \in \mathbb{R}^{N \times 2 \cdot D}$
4. $\gamma \sim IG(1,1)$ represents the lengthscale of the subsequent GP
5. $\beta_{rtk} \sim \mathcal{GP}(0, I_{2D} \otimes I_K \otimes K_T(\gamma))$
    - $K_T(\gamma)$ is the covariance function capturing auto-correlation among time observations
    - Separable covariance structure for time, metric
6. $\mu = Z(X)^{pr} \beta^r_{tk} \in \mathbb{R}^{N \times T \times K}$ is represented as a tensor contraction between $Z(X)$ and β over the second index.

Modeling Assumptions

We include the following metrics and distribution families

1. Poisson
   - $\mathcal{R} =$ {FG2A, FG3A, FTA, BLK, OREB, DREB, TOV, AST, STL}
2. Gaussian
   - $\mathcal{G} =$ {DBPM, OBPM}
3. Binomial
   - $\mathcal{B} =$ {FG2M, FG3M, FTM}
   - $\mathcal{N} =$ {FG2A, FG3A, FTA}
4. Exponential
   - $\mathcal{M} =$ {Minutes}
5. Bernoulli
   - $\mathcal{K} =$ {Retirement}

Model Assumptions (contd.)

$$Y_{ptk} \sim \begin{cases} Pois(Y_{ptm}e^{\mu_{ptk}}) \text{ if } k \in \mathcal{R} \text{ , } \forall m \in \mathcal{M} \\ Bin(Y_{ptj}, logit^{-1}(\mu_{ptk})) \text{ if } k \in \mathcal{B} \text{ , } j \in \mathcal{N} \\ \mathcal{N}(\mu_{ptk}, \frac{\sigma_k^2}{Y_{ptm}}) \text{ if } k \in \mathcal{G} \text{ , } \forall m \in \mathcal{M} \\ Bern(logit^{-1}(\mu_{ptk})) \text{ if } k \in \mathcal{K} \\ Exp(e^{\mu_{ptk}}) \text{ if } k \in \mathcal{M} \end{cases} \quad (1)$$

Challenges

1. MCMC Convergence (multi-modal posterior)
2. Identifiability (rotational / scale invariance of model)
3. Modeling temporal and within-metric correlation

Methods (Approach 1)

In order to address identifiability issues and MCMC convergence, we propose the following scheme to estimate the latent space X and functional coefficients β_{rtk}.

1. Initialize X
   - Exponential PPCA, Probabilistic Tensor Decomposition, Standard PCA, etc.
2. Using the fixed X from above, conduct inference on $\beta_{rtk}, \sigma_k, \omega_r$

Methods (Approach 2)

In order to address identifiability issues and MCMC convergence while also recovering sampling variability in the latent space, we propose an alternating scheme to estimate the latent space X and functional coefficients β_{rtk}.

1. Let $X \sim \mathcal{N}(\mu_0, \Sigma_0)$ where μ_0 and Σ_0 come from an initialized latent space X_0.
    - Exponential PPCA, Probabilistic Tensor Decomposition, Standard PCA, etc. can be used to create X_0
2. Using a hybrid Gibbs-HMC routine, perform the following updates:
    - Sample X, γ while holding all other parameters fixed using HMC proposal step
    - Conditional on the sampled X and γ, sample the remaining parameters using HMC proposal step

Methods (Approach 3)

In order to address identifiability issues and MCMC convergence while also recovering sampling variability in the latent space, we propose an alternating scheme to estimate the latent space X and functional coefficients β_{rtk}.

1. Let $X \sim \mathcal{N}(\mu_0, \Sigma_0)$ where μ_0 and Σ_0 come from an initialzed latent space X_0.
   - Exponential PPCA, Probabilistic Tensor Decomposition, Standard PCA, etc. can be used to create X_0
2. Conditional on the fixed X_0, sample the remaining parameters using HMC until convergence.
3. Taking the posterior mean of all parameters resulting from (2), sample X using HMC until convergence.

Current Progress

1. Shiny App

Future Work

1. Address trend in baseline rate of 3PA, etc over time
2. Impose correlation across metrics
3. Look at hold-out coverage interval
4. Loosen separable covariance assumption

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

Appendix

Random Fourier Features

Attempt to approximate the inner product $k(x, y) = \langle \phi(x), \phi(y) \rangle$ with a randomized map $z : \mathbb{R}^D \rightarrow \mathbb{R}^R$. Computational savings arise if $R << N$.

In our case, we let $k(x, y) = k(x - y) = exp(\frac{-||x-y||^2}{2})$ be the standard radial basis kernel.

From Bochner's theorem, we have that $k(x - y) = \int p(\omega) exp(i\omega(x - y))d\omega$, and it can be shown that to produce the radial basis kernel, $\omega \sim \mathcal{N}_D(0, I_d)$.

Thus the map is composed of $z_{\omega_r} = [cos(\omega_r^T x), sin(\omega_r^T x)]^T$.

$Z(X) = \frac{1}{\sqrt{R}}[z_{\omega_1}, z_{\omega_2}, \ldots, z_{\omega_R}]^T$

Probabilistic Tensor Decomposition

This model seeks to factorize the $N \times T \times K$ linear scale tensor A using CP Decomposition. Since we have various outputs that are not normally distributed, this becomes a form of exponential family CP Decomposition.

We seek to approximate the following:

$A \approx \mu + \sum_{i=1}^{R} \lambda_i \cdot x_i \otimes v_i \otimes w_i$

where, μ, x_i, v_i, $w_i \sim \mathcal{N}(0, I)$, $\lambda \sim Dirichlet(1/R)$

$X \in \mathbb{R}^{N \times R}$

$V \in \mathbb{R}^{T \times R}$

$W \in \mathbb{R}^{K \times R}$

$\mu \in \mathbb{R}^{T \times K}$.

Here μ is used to de-mean the data and act as an intercept term.

Probabilistic Tensor Decomposition (contd.)

Let $\tilde{A}_{pit} = g_{pit}^{-1}(A_{pit})$, where g_{pit} is the appropriate link function transforming the linear scale parameter into the appropriate exponential family parameterization. Consequently, X, V, W, μ are estimated by maximizing the following loss function using gradient descent.

$\max_{X,V,W,\mu} \sum_{p,i,t} log(F_{pit}(Y_{pit}|\tilde{A}_{pit})) \cdot \Omega_{pit}$

where F_{pit} is the appropriate distribution density function associated with entry Y_{pit}.

This offers the following benefits:

1. Latent space X is created while accounting for sampling variability
2. Latent space is created while also accounting for correlations across each mode of the tensor, which is representative of the final model.