NHL Aging Curves using Functional Principal Component Analysis

Elijah Cavan, Jiguo Cao and Tim B. Swartz *

Abstract

When considering future performance in sport, age is an important feature for prediction models. On average, players tend to improve from their rookie (earliest) season, plateau, and then decline in performance until they retire from the league. In this paper we apply Functional Principal Component Analysis to the careers of players from the National Hockey League in order to construct individual aging curves. The approach is nonparametric in the sense that a parametric structure is not imposed on the aging curves. A main aspect of our work is the consideration of selection bias.

Keywords : Aging curves, Functional data analysis, National Hockey League, Principal component analysis, Sports analytics.

*E. Cavan is an MSc graduate and data analyst at Pinnacle Sports, and J. Cao and T. Swartz are Professors, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. Cao and Swartz have been partially supported by the Natural Sciences and Engineering Research Council of Canada. The work has been carried out with support from the CANSSI (Canadian Statistical Sciences Institute) Collaborative Research Team (CRT) in Sports Analytics.

1 INTRODUCTION

Many decisions made by front office staff of major sports clubs involve projecting the future performance of players. In the case of drafting younger players, teams attempt to predict a player's future performance given their abilities relative to their peers in junior leagues. However, decisions involving trades and free-agency are different from drafting decisions. In these cases, a team must forecast the performance of players who have already been entrenched in the league, and this is subject to the constraint that the number of players on team rosters is fixed. The composition of players is constantly changing as older players retire and younger players take on more prominent roles. In fantasy sports, the relationship between performance and age is also greatly discussed (Cockcroft 2023). Therefore, it is of vital interest to predict the future values of players as they age. The results of such prediction models are referred to as aging curves. Aging curves can be difficult to construct in team sports due to the fact that player performance is highly dependent on teammates and the number of minutes played.

Aging curves have been studied in many sports including: golf (Berry, Reese and Larkey 1999), football (Young and Weckman 2008), baseball (Bradbury 2009), soccer (Swartz, Arce and Parameswaran 2013), hockey (Brander, Egan and Yeung 2014), tennis (Mlakar and Tušar 2015), cricket (Saikia, Bhattacharjee and Mukherjee 2019), basketball (Wakim and Jin 2014) and snooker (McHale 2023).

The effect of aging on the body is a common issue for all athletes (Distefano and Goodpaster 2018). For many sports, players reach their peak performance before the age of 30 years, and then generally decline as they age due to decreasing athleticism and increased injury risk. The observance of this "peak" in performance has been studied, for example, by Dendir (2016) and Bradbury (2009). Bradbury (2009) demonstrates that different skills decline in baseball at different rates; baserunning, for example, declines at a much faster rate than power. Most research reaches the general consensus that age effects are dependent on position. For example, Brander, Egan and Yeung (2014) conclude that the peak age for forwards in the National Hockey League (NHL) is between ages 27-28, while the peak age for defensemen is between 28-29. Each player has a unique aging curve (due to different body composition or previous athletic history, for example). However, it seems reasonable that there should be some agreement in age curves between players.

In the literature, there are two prominent general approaches for the construction of aging curves; (1) the so-called “Delta method” and (2) regression methods. The Delta method calculates the difference between a player’s performance between years, and then averages these differences over players. A feature of the Delta method is that it is not impacted by the differences in quality between players. There are variations to the Delta method; see for example, Lichtman (2009) and EvolvingWild (2017). With regression methods, the response variable of interest (player performance) is regressed against the covariate age. Regression methods may differ in the form of regression and whether additional covariates are considered in the regression model (e.g., a player effect). Early regression methods tended to take a parametric approach. For example, Fair (2008) and Bradbury (2009) imposed quadratic shapes on aging curves in baseball. Villaroel, Mora and Gonzalez-Parra (2011) consider both quadratic and cubic shapes with respect to the performance of triathletes. Although convenient, parametric approaches do not permit freedom in the shape of aging curves. For example, it has been observed in chess (Roring and Charness 2007) that performance improvement leading to peak performance may occur at a steeper incline than the decline in performance following the peak. Many of the more recent regression methods tend to take a nonparametric or semiparametric approach. For example, Turtoro (2019) uses generalized additive models (GAM) for constructing aging curves in the NHL.

Another distinguishing feature in the construction of aging curves is whether or not an approach considers selection bias. Selection bias is a systematic statistical error caused by drawing a non-random sample from a population. In sports, samples of player performances are typically non-random because only the most talented players enter leagues at very early ages. Moreover, these players are usually the same players who stay in the league the longest (except in the case of early career altering injuries). Therefore, gifted players are typically overly represented in both the left and the right tails of age distributions. Schuckers, Lopez and Macdonald (2023) demonstrate the considerable impact of selection bias using data from the NHL. Various approaches have been proposed that consider selection bias in the construction of aging curves. These all involve imputation schemes that “add” data over missing periods. For example, after a player retires, data are imputed for the years following retirement. Lichtman (2009) takes a basic approach where regression methods impute data based on best estimates. Schuckers, Lopez and Macdonald (2023) use more sophisticated imputation schemes based on the simulation of missing data. They introduce thresholds for player performance at different ages and ensure that simulated data do not breach the threshold. Nguyen and Matthews (2023) take the simulation approach one step further by investigating the cause of missingness, and use different simulation approaches based on the different causes of missingness.

Although there is now a considerable literature on methods developed for the construction of aging curves, there are several innovations in this paper. First, we use an approach based on Functional Principal Component Analysis (FPCA). This is a nonparametric approach which permits flexibility in the shape of growth curves. Second, within the FPCA framework, we account for selection bias by imposing an intuitive constraint on the corresponding likelihood function. Third, and most importantly, we construct aging curves for individual players. This is obviously an important contribution since players age differently. The curves have a predictive component where curves can be extended beyond a player's current age. An important aspect of the FPCA model is that there is an underlying relationship between the individual aging curves, and this enables prediction. Previously, aging curves have only been constructed for the so-called "average" player.

The proposed FPCA approach is predicated on Functional Data Analysis (FDA). In FDA, we use spline basis functions to determine functional relationships. This provides a relationship between performance and age that is essentially nonparametric. In addition, to fitting separated aging curves for each player, an FDA approach can identify clusters from principal component scores, allowing us to readily compare players. FPCA aging curves have been previously considered in the context of basketball by Wakim and Jin (2014) where implementation is based on conditional expectation using the PACE method. However, the PACE approach does not account for the selection bias issue.

In Section 2, we describe the data and the player evaluation metric. In Section 3, we carry out some exploratory data analysis to investigate aging patterns. In Section 4, we outline the methods for constructing an aging curve using FPCA. Here, we briefly outline some of the underlying mathematical background. In Section 5, we present the results of our modelling. In Section 6, we obtain predictions involving some of the best players. We conclude with a short discussion in Section 7.

2 DATA

The data in this project was scraped from Sports Reference LLC at https://www.hockey-reference.com. The website contains summary statistics (goals, points, games played, etc) for each player in the NHL during the period 1920-2022. Originally, there were approximately 50,000 rows in the dataset before adjusting for duplicate rows (due to players changing teams mid-season). This led to $n = 7,393$ unique players with 43,705 player-seasons worth of data. The maximum player age in the dataset is 51 years (Gordie Howe), and the minimum player age is 17 years (Wayne Gretzky). From an alternative study (Diamond 2000), the average career length in the NHL is 4.5 years.

2.1 Player Value

As a measure of player value, we use the point share PS statistic (Kubatko 2010). It is a measure derived from the win share metric (James and Henzler 2002) that was originally used to evaluate baseball players. Here, "points" refer to the points a team gains from winning games, and not the points a player gains from scoring a goal or assisting on a goal in hockey. Hence, the metric attempts to credit a player's contribution to their team's success. This metric was chosen because we require a composite measure of performance that adjusts for the quality of linemates. Goals and assists metrics are not composite measures since they are not reflective of contributions to team defense, and would consequently overvalue forwards.

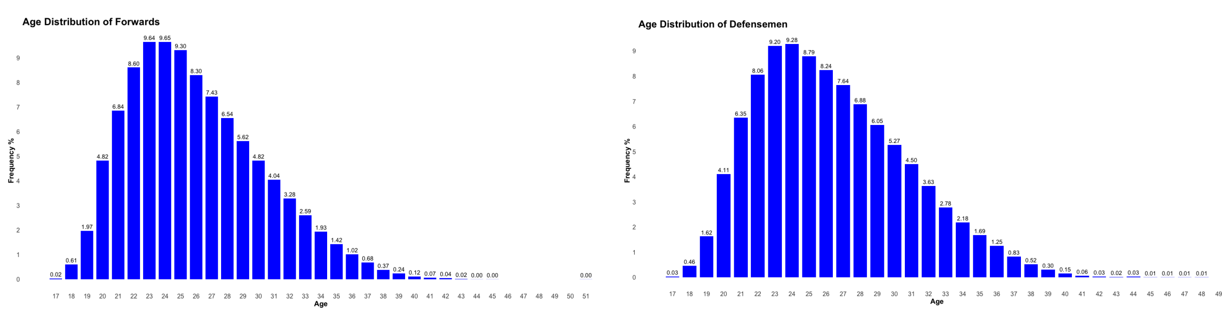
The point share metric has been developed so that a hockey team with 100 team points (e.g., 40 wins and 10 overtime losses) will have players whose individual point shares sum to 100. Players may have a negative point share. Negative point shares indicate that a player is losing team points relative to a replacement level player. Point shares are obtained from both offensive and defensive component point shares. The offensive point share for a player during a season is based on their goals created (a weighted sum of the player's goals and assists divided by team goals and assists), adjusted by the player's minutes and adjusted for the league environment (league goals divided by league points). There are also positional adjustments for forwards and defensemen.

Over time, adjustments have been made to the point share calculation based on various rule changes. For example, player minutes on ice have only been recorded since the 2000-2001 NHL season. Also, the number of games in the NHL season has increased over the years and the NHL previously permitted tied games, which yielded one point to both teams. Consequently the point share metric is not perfect; however it is still regarded as an advanced statistic in hockey analytics. The point share PS statistic can be obtained from https://www.hockey-reference.com. More details on it's calculation are found in Kubatko (2023).

Alternative metrics for assessing performance include those based on salary considerations (Swartz, Arce and Parameswaran 2013). Also, there are limitations on the availability of advanced match statistics for use in aging curves. For example, the NHL only started to track individual shots and plus-minus statistics in 1960-1961.

3 EXPLORATORY DATA ANALYSIS

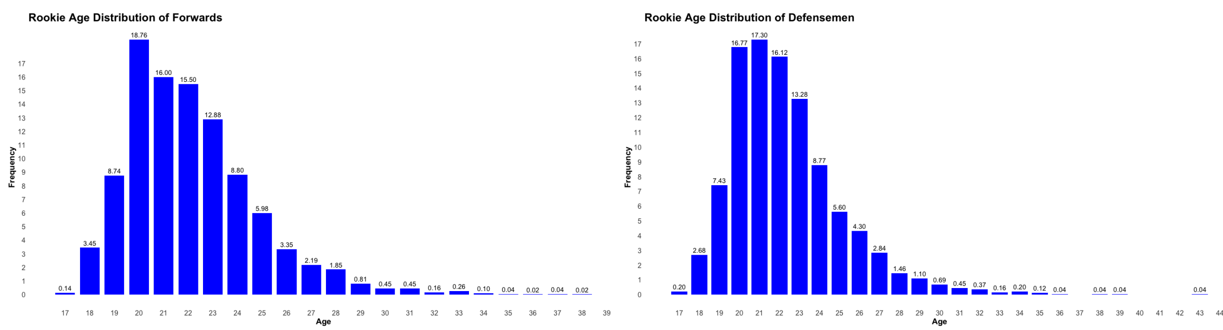
In Section 5, our FPCA analyses consider forwards and defensemen separately. In this section, we provide some basic exploratory plots that motivate this distinction. It is suggested that forwards and defensemen age differently in the NHL.



(a) Forwards (b) Defencemen

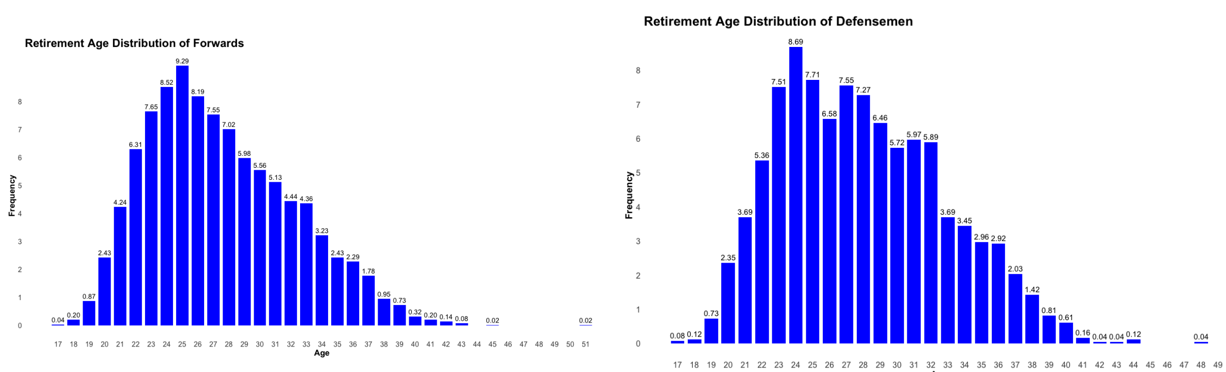
Figure 1: Histograms of the age of NHL players based on player-seasons from 1920-2022.

Figure 1 is a histogram based on what is called the "participation method" (Brander, Egan and Young 2014). Here, every player-season is taken as an observation where age is the recorded variable based on player age at the end of the season. The idea is that players participate in the NHL during the seasons of their peak performance. For some, this window may be short. Therefore, the modal regions of the histogram indicate peak periods of performance. In the histograms, there are 28,871 observations corresponding to forwards and 14,834 observations corresponding to defensemen. From Figure 1, we observe that the modal age for both positions lies between 23-24 years of age. We also observe that there is greater participation longevity for defenseman than forwards.



(a) Forwards (b) Defensemen

Figure 2: Histograms of rookie age by position.



(a) Forwards (b) Defencemen

Figure 3: Histograms of retirement season age by position.

We also provide histograms for the ages of the rookie and retirement seasons of NHL players in Figure 2 and Figure 3, respectively. From Figure 2, there is little difference between the time when forwards and defensemen begin their careers. It appears that most players begin their careers around 20-22 years of age. From Figure 3, there is some indication that defensemen end their careers slightly later than forwards. There is much greater variability in the retirement age than the rookie age; most players retire before 38 years of age. Of course, “early retirement” may simply be a case of players no longer achieving the standards of NHL play.

4 METHODS

4.1 Overview of FDA

FDA is a highly flexible statistical framework which is concerned with the modelling of longitudinal data. FDA is a modern approach to multivariate statistical modelling, with many applications as seen in Ramsay and Silverman (2005). A review of the current advances in the topic can be found in Wang, Chiou and Müller (2016). Recently, FDA has been used in the analysis of sports data, as seen in Guan et al. (2022) where it is used to specify conditional distributions for in-game win probabilities in rugby. In Chen and Fan (2018), FDA is used to model the score difference process in basketball. Statistical contributions to sport are highlighted in the handbook by Albert et al. (2017).

FDA is different from well-known related methods. For example, unlike time series analysis, FDA does not impose underlying assumptions regarding stationarity. Also, unlike standard multivariate statistics whose data consist of multiple measurements for each subject, FDA curves are viewed as infinite-dimensional random vectors in a functional space. There is an underlying assumption in FDA that observed samples are independent stochastic processes. FDA is particularly suited to handle sparse data.

FDA can be used to perform a number of common machine learning tasks such as classification, clustering, ANOVA, regression, principal component analysis, interpolation and extrapolation. The enhancement provided by FDA is that point estimators are replaced with functional estimators. Some benefits of FDA include the representation of observed data as smooth functions, dimensionality reduction and the ability to compute derivatives of smoothed estimators. The are numerous packages that can be used to perform FDA; the common ones being *scikit-fda* (Python) and the R package *fda* available on CRAN.

4.2 Spline Regression for FDA

When performing FDA, we consider the observations from each individual arising from a random, smooth function. For example, suppose that we observe data from $i = 1, \ldots, N$ players, and for the ith player, we observe m data points $(y_{i1}, \ldots, y_{im})$ at time points $t_1, \ldots, t_m$. Then the core assumption in FDA is that

$$y_{ij} = X_i(t_{ij}) + \epsilon_{ij} \tag{1}$$

where the ϵ_{ij} are independent and normally distributed with mean θ and variance σ^2. Equation (1) explicitly assumes that the observations from individual i can be modelled by a single stochastic function $X_i(t)$ after accounting for measurement error ϵ_{ij}. Although we suppress additional notation, we note that the number of observed points m can actually vary from player to player. In order to approximate the functions $X_i(t)$, we introduce Q spline basis functions $b_1(t), \ldots, b_Q(t)$. Spline functions are piecewise polynomials joined at specific points, called knots. The number of knots τ is determined by $\tau = Q - d + 1$ where d is the degree of the polynomial. We write

$$X_i(t) \approx \sum_{q=1}^{Q} \alpha_{iq} b_q(t)$$

for sufficiently large Q where the α_{iq} are the coefficients of the spline basis functions. We estimate $\alpha_{i1}, \ldots, \alpha_{iQ}$, by minimizing the least squares loss function

$$\sum_{j=1}^{m} \left(y_{ij} - \sum_{q=1}^{Q} \alpha_{iq} b_q(t) \right)^2 .$$

Once the α_{ik} have been estimated via $\hat{\alpha}_{ik}$, this leads to the estimated aging curve

$$\hat{X}_i(t) = \sum_{q=1}^{Q} \hat{\alpha}_{iq} b_q(t) \tag{2}$$

for the ith player. We can also compute statistics such as the mean aging curve across all players

$$\hat{\mu}(t) = \frac{\sum_{i=1}^{N} \hat{X}_i(t)}{N} \ .$$

4.3 Functional Principal Component Analysis

In our problem, assume for the time being that the number of observations m is constant across all players, $i = 1, \ldots, N$ and that $t_{ik} = t_{jk}$ for all players i, j and all ages $k = 1, \ldots, m$. FPCA begins with the estimated covariance function

$$\hat{v}(s,t) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{X}_i(t) - \hat{\mu}(t) \right) \left(\hat{X}_i(s) - \hat{\mu}(s) \right)$$

which describes the correlation of aging curves at ages s and t, where $\hat{X}_i(t)$ is the estimated aging curve in (2) corresponding to player i.

Our objective is to obtain the orthonormal eigenfunctions $\hat{\xi}_1(t), \ldots, \hat{\xi}_p(t)$ and the eigenvalues $\rho_1, \ldots, \rho_p$ by solving the equation

$$\int \hat{v}(s,t)\xi(s) \ ds = \rho\xi(t)$$

where we introduce the dimensionality reduction $p \leq m$. The eigenfunctions $\hat{\xi}_1(t), \ldots, \hat{\xi}_p(t)$ are known as functional principal components. This leads to FPC scores

$$s_{i\ell} = \int \hat{\xi}_\ell(t)(\hat{X}_i(t) - \hat{\mu}(t)) \ dt \tag{3}$$

for players $i = 1, \ldots, N$ and the ℓth functional principal component, $\ell = 1, \ldots, p$.

It is the FPC scores $s_{i\ell}$ in (3) that readily permit the comparison of players in lower-dimensional settings. Once we have computed all of the FPCs terms, individual aging curves are expressed by

$$\hat{X}_i(t) = \hat{\mu}(t) + \sum_{\ell=1}^{p} s_{i\ell}\hat{\xi}_\ell(t) \ . \tag{4}$$

The formula (4) follows from the Karhunen–Loève expansion. The truncation is based on the first p principal components that explain most of the variation between curves.

4.4 FPCA using imFunPCA

The conventional FPCA method assumes that missing data are missing at random (MAR). For the aging curves, we know that this is not the case due to selection bias. In Shi et al. (2021), the authors adjust for this bias using a constrained likelihood approach for missing data. For example, assuming that the data y_{ij} are normally distributed with the mean $X_i(t_{ij}) = \mu(t_{ij}) + \sum_{\ell=1}^{p}\{s_{i\ell}\xi_\ell(t_{ij})\}$ and the variance σ^2, the authors show that the functional principal components can be calculated by maximizing the likelihood function

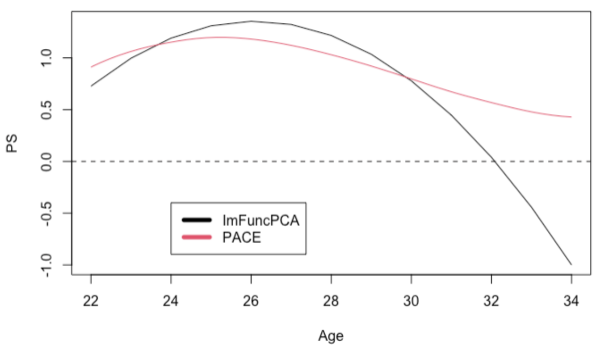
$$\prod_{i=1}^{N}\prod_{j=1}^{M} f(y_{ij})^{1-\delta_{ij}} P(y_{ij} \leq c_i)^{\delta_{ij}} , \tag{5}$$

where M is the index corresponding to maximum age, $f(y_{ij})$ is the probability density function of y_{ij}, δ_{ij} is an indicator function corresponding to whether the jth observation for player i is missing where player i has m_i observations. Here, we allow for different numbers of observations per player. The missing data in the right tail have the natural constraint that once a player is out of the league, future values of performance are no larger than c_i; c_i is set to be the final observation for the ith player. More details concerning the *imFunPCA* method are provided in Shi et al. (2021).

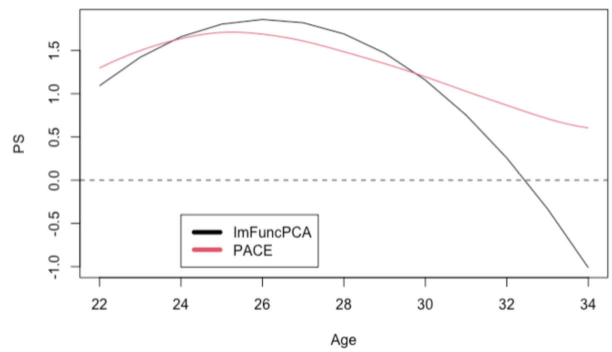
5 RESULTS

We restrict the data to include only player-years 22-34. We also limit the data to players who had an NHL career lasting at least seven seasons. This provides us with 1785 forwards and 883 defencemen. We randomly partitioned the data into training sets (1420 forwards, 706 defensemen) and testing sets (356 forwards, 177 defensemen) such that the testing sets form approximately 20% of the data. First, we obtained the estimated mean function $\hat{\mu}(t)$ using the spline regression method outlined in Section 4.2 on the training data. We used $Q = 6$ spline basis functions of degree three. Figure 4 displays the estimated mean point share (PS) functions for forwards and defencemen using the imFunPCA method and the conventional PACE method, where we observe concave shapes with peaks around 26-28 years. This period of peak performance differs from the participation peak observed in

Figure 1. It is observed that the mean point share functions, as estimated by the imFunPCA method, exhibit faster declines compared to those estimated by the conventional FPCA method. This discrepancy arises from the fact that the imFunPCA method takes into account selection bias, specifically, the retention of only skilled players in the league after reaching their peak performance.



(a) Forwards



(b) Defensemen

Figure 4: The estimated mean point share (PS) functions for forwards and defencemen using the imFunPCA method (black lines) and the conventional PACE method (red lines).

5.1 Functional Principal Components

Figure 5 shows the two functional principal components (eigenfunctions) of the point share (PS) functions for forwards estimated using the imFunPCA method. The first functional principal component consistently maintains a positive value throughout the entire career, with a greater magnitude in the early career compared to the later stages. We interpret the scores on this first functional principal component as a weighted average of point shares, where the component itself acts as the weighting factor. It is noteworthy that the first functional principal component assigns higher importance to the early career in comparison to the later stages. Forwards will obtain a substantial score on the first functional principal component if they exhibit strong performance throughout their entire career. Conversely, the second functional principal component is positive in the early career before age 26 and diminishes thereafter. Scores on the second functional principal component signify declining performance levels after age 26. A higher absolute value of score on the second functional principal component indicates a pronounced decrease in performance after reaching the age of 26.

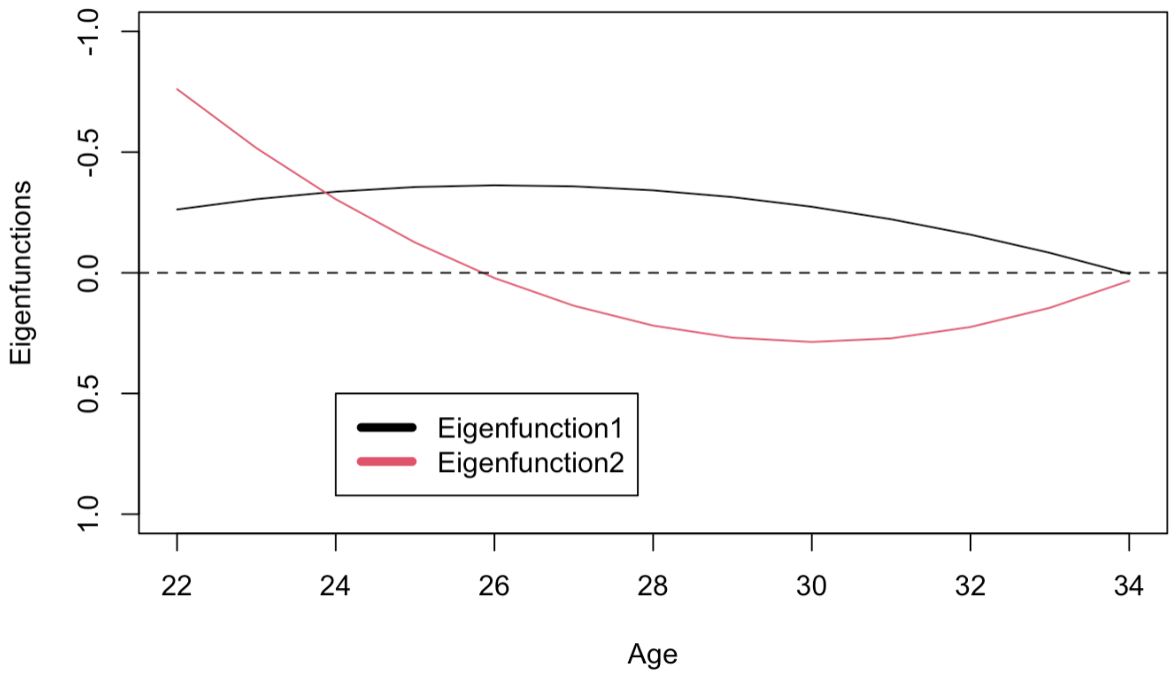


Figure 5: The two functional principal components (eigenfunctions) of the point share (PS) functions for forwards using the estimated imFunPCA method.

Figure 6 illustrates the two functional principal components (eigenfunctions) of the point share (PS) functions for defensemen, as determined by the imFunPCA method. The first functional principal component consistently maintains a positive value throughout the entire career, displaying a more pronounced magnitude in the early career compared to later stages. Scores on this first functional principal component are interpreted as a weighted average of point shares, where the component itself serves as the weighting factor. It is important to highlight that the first functional principal component places greater emphasis on the early career. Defensemen will achieve a substantial score on the first functional principal component if they demonstrate strong performance throughout their entire career. In contrast, the second functional principal component is positive in the early career before age 27 and decreases thereafter. Scores on the second functional principal component indicate diminishing performance levels after age 27. A more substantial absolute value of score on the second functional principal component signals a notable decline in performance after reaching the age of 27.

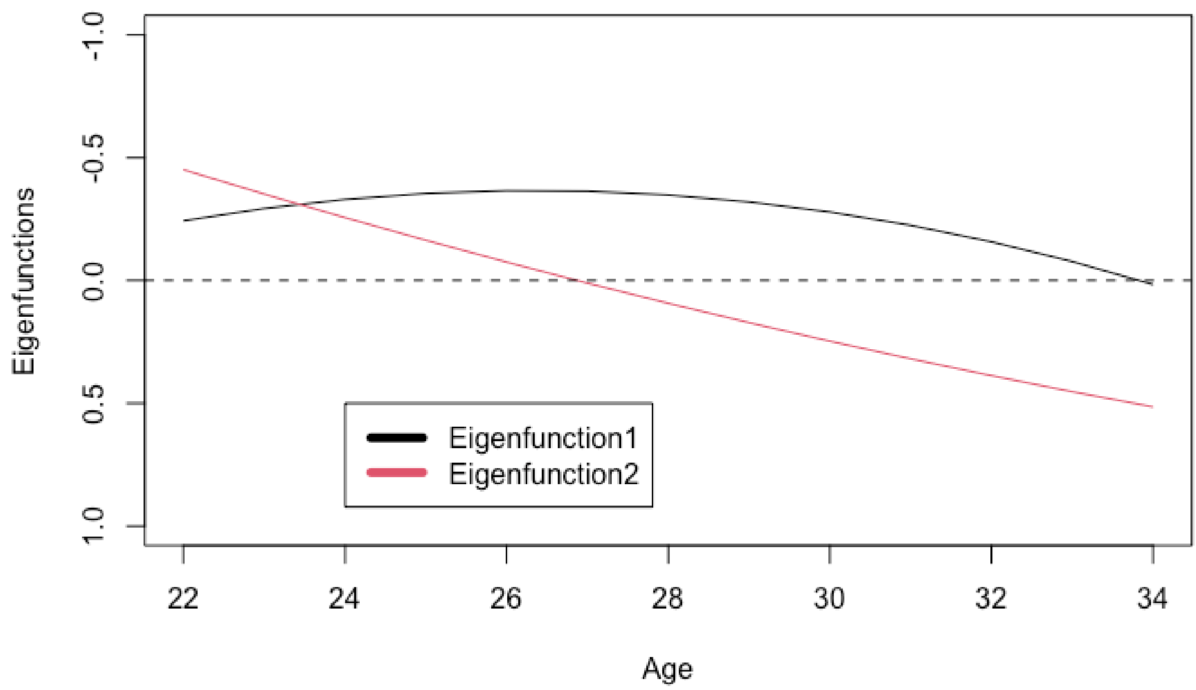


Figure 6: The two functional principal components (eigenfunctions) of the point share (PS) functions for defensemen using the estimated imFunPCA method.

5.2 Using FPCA for Prediction

To address the prediction problem, we randomly chose 80% of the players to form the training dataset and the remaining 20% comprise the test dataset. We start by deriving the first two principal components from the training dataset. Using these components, we calculate the FPC scores for the players in the test set, applying Equation (3). With these scores, we then use Equation (4) to predict the aging curves for the players in the test set.

The prediction errors are summarized by the median of mean absolute errors (MAE) and median absolute deviations (MAD) defined as follows:

$$\begin{aligned} \text{MAE} &= \underset{i \in \texttt{Test Set}}{\texttt{median}} \left\{ \frac{1}{D_i} \sum_{j \in D_i} |y_{ij} - X_i(t_{ij})| \right\} \\ \text{MAD} &= \underset{i \in \texttt{Test Set}}{\texttt{median}} \left\{ \underset{1 \leqslant j \leqslant D_i}{\texttt{median}}(|y_{ij} - X_i(t_{ij})|) \right\} \end{aligned} \tag{6}$$

where $j \in D_i$ are the observed ages for the ith player, y_{ij} is the true point share performance and $X_i(t_{ij})$ is the prediction based on FPCA.

Table 1 presents a comparison of prediction errors for the point share performance of forwards and defensemen using the imFunPCA, PACE, and Delta methods (Lichtman 2009). The imFunPCA method demonstrates the most accurate prediction, reducing the median of mean absolute errors (MAE) for forwards by 45.3% compared to the Delta method and 6.5% compared to the PACE method. For defensemen, the imFunPCA method also yields the most accurate prediction, reducing the median of mean absolute errors (MAE) by 33.0% compared to the Delta method and 10.9% compared to the PACE method. Similar conclusions can be drawn when comparing the three methods using median absolute deviations (MAD).

	Forwards		Defencemen	
Method	MAE	MAD	MAE	MAD
imFunPCA	0.799	0.768	1.079	1.058
PACE	0.855	0.839	1.211	1.073
Delta	1.489	1.405	1.611	1.611

Table 1: Medians of mean absolute errors (MAE) and median absolute deviations (MAD) defined in (6) for predicting the point share performance of forwards and defencemen using the imFunPCA, conventional FPCA and Delta methods.

We can also estimate the FPC scores from a player's early seasons, and then use the estimated mean curve and estimated functional principal components (FPCs) to forecast the player's future performance using Equation (4). This may be of great value to front office staff at major sports clubs. Figures 7 presents a predicted career path at future ages 32, 33, and 34 for Corey Perry and Marc-Edouard Vlasic based on their point share performance in their career before 31 years old using the imFunPCA, conventional FPCA and Delta methods. It shows that the imFunPCA method always has a more accurate prediction than the conventional FPCA and Delta methods.

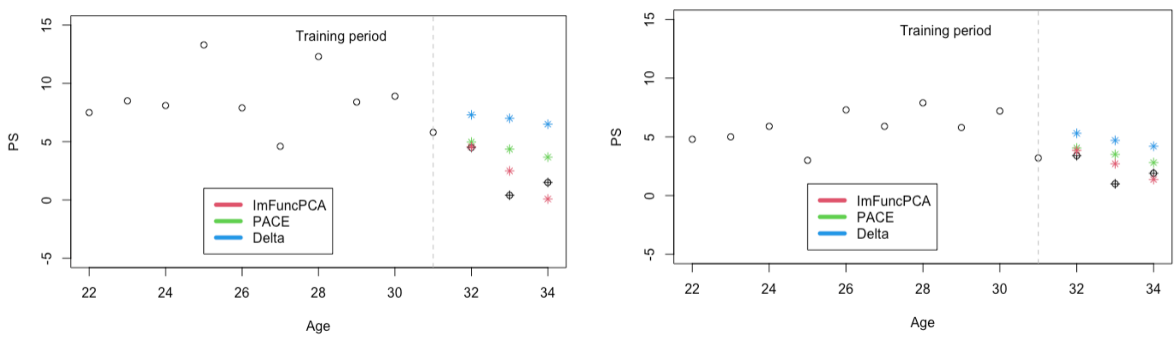
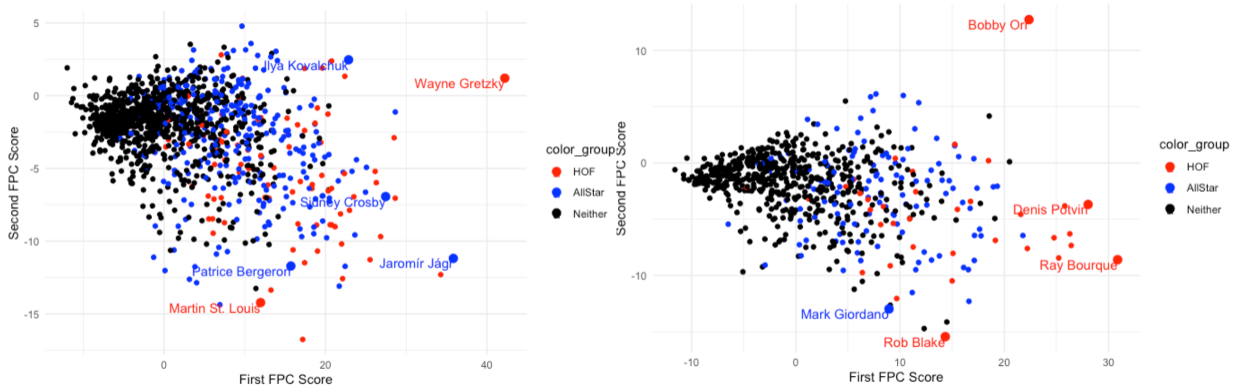


Figure 7: Predicting the career path at future ages 32, 33, and 34 for Corey Perry (left) and Marc-Edouard Vlasic (right) based on their point share performances in their career before 31 years of age using the imFunPCA, conventional FPCA and Delta methods. The black circles and diamonds are the observed values.

5.3 Using FPCA for Clustering

Another benefit of FPCA is dimension reduction. We can project aging curves into corresponding FPC scores. For instance, after obtaining the FPC using the imFunPCA method, we can obtain the FPC scores for each aging curve. These FPC scores can be understood as features of these aging curves. The interpretation for FPC scores are provided in Section 5.1.

Figure 8 illustrates the top two FPC scores for each aging curve among forwards and defensemen. The players are also colour coded according to whether they were Hall of Famers, All-Stars or neither. The FPC scores clearly delineate a clustering amongst these three types of players. In addition, FPC scores denote intriguing career trajectories for individual players. Among the forwards, Wayne Gretzky exhibits notably high first FPC scores, indicative of his consistent high point shares throughout his career. In the case of defensemen, Bobby Orr boasts a substantial first FPC score due to his sustained high point shares early in his career, but his significant second FPC score reflects a sharp decline later



(a) Forwards (b) Defensemen

Figure 8: Clustering the FPC scores by position using the imFunPCA method.

in his career primarily due to injuries suffered while playing.

6 PREDICTION OF CAREER CONTRIBUTIONS

An often heated discussion in sport concerns the identification of the best player. Such a player is often referred to as the GOAT (greatest of all time). There can be considerable disagreement about the GOAT due to various personal preferences including the determination of a suitable performance metric and how one ought to compare players from different eras.

We entertain this problem in the context of the NHL. As a metric, we use career point shares. This metric is appealing since it considers both excellence and longevity.

In Table 2, we present the top 10 forwards and top 10 defencemen according to career point shares. There are two active players in the list at the completion of the 2021/2022 season. These players are both forwards (Ovechkin and Crosby). We note that all of the players in Table 2 had exceptionally long careers, and this points to the emphasis on longevity in the career point share metric. It is remarkable that Gordie Howe's 26 playing years do not include the six years that he played in the WHA (World Hockey Association). It is curious that Bobby Orr who is considered by many to be the GOAT does not make the top 10 list for defencemen. This is explained by a knee injury that limited his career to 12 NHL playing seasons.

Forwards		Defencemen	
Player	PS (YP)	Player	PS (YP)
Wayne Gretzky	250.9 (20)	Ray Bourque	242.5 (22)
Gordie Howe	217.1 (26)	Nicklas Lidström	211.8 (20)
Jaromir Jágr	217.0 (22)	Al MacInnis	195.0 (23)
Alex Ovechkin*	188.9 (17)	Paul Coffey	185.7 (21)
Teemu Selänne	172.3 (21)	Larry Murphy	176.9 (21)
Sidney Crosby*	170.4 (17)	Phil Housley	170.6 (21)
Joe Sakic	167.9 (21)	Scott Stevens	170.0 (22)
Mario Lemieux	167.9 (17)	Larry Robinson	169.4 (20)
Steve Yzerman	166.7 (22)	Chris Chelios	168.5 (26)
Phil Esposito	163.7 (18)	Denis Potvin	160.7 (15)

Table 2: Top 10 NHL forwards and defencemen according to career point shares with years played (YP). Asterisks denote active players at the end of the 2021/2022 season.

From Section 5, our preferred method of prediction is imFUNPCA. We use imFUNPCA to predict future years for three active NHL forwards and defencemen. The players were all 31 years of age during the 2021/2022 NHL season (the endpoint of our analysis) with the highest career point shares. We then predict point share accumulation for three more years of play until they reach 34 years of age (the endpoint of our predictions). Of course, it is entirely possible that these players could play beyond 34 years.

Again, a great benefit of the imFUNPCA approach is that aging curves can be predicted for individual players in contrast to typical aging curves that refer to the "average" player. Another benefit of imFUNPCA prediction is that the procedure yields standard errors. Therefore, this permits the calculation of predictive confidence intervals.

In Table 3, we present the prediction of these six players and the corresponding lower and upper 90% prediction intervals. We observe that the six players lie significantly short of the top 10 lists even with extended careers going beyond 34 years of age. From these predictions, it seems that Wayne Gretzky's status as the GOAT using the career point shares metric will not be challenged in the near future.

Predictions for Forwards				Predictions for Defencemen			
Player	P	Lower	Upper	Player	P	Lower	Upper
Steven Stamkos	102.5	96.5	108.5	Victor Hedman	121.3	115.1	127.5
John Tavares	96.5	91.8	101.2	Roman Josi	118.3	111.8	124.7
Matt Duchene	67.8	62.2	73.4	Erik Karlsson	89.7	82.5	96.9

Table 3: Three active 31-year old NHL forwards and defencemen with career point share predictions extended to 34 years of age. The prediction P, lower and upper 90% limits are provided.

7 CONCLUSIONS AND DISCUSSION

Using techniques from Functional Principal Component Analysis, we have developed new methods for the construction of aging curves. The methods have several advantages including the avoidance of a fixed parametric shape on the aging curve and the consideration of selection bias where only the best players are represented in the older age cohorts. However, the most important contribution in our approach is the construction of individual aging curves where performance is predicted beyond the current age. This may be of great value to teams in terms of player retension and roster construction.

We have imposed an intuitive constraint on the likelihood function (5) using the im-FUNPCA procedure to address selection bias in the right tail of the age distribution. In future work, a similar constraint could be imposed on the left tail of the distribution to account for early bloomers. However, this is a lesser issue as players do not enter the NHL (no matter how good they are) at extremely young ages.

REFERENCES

Albert, J.A., Glickman, M.E., Swartz, T.B. and Koning, R.H., Editors (2017). *Handbook of Statistical Methods and Analyses in Sports*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Boca Raton.

Berry, S.M., Reese, C.S. and Larkey, P.D. (1999). Bridging different eras in sports. *Journal of the American Statistical Association*, 94(447), 661–676.

Bradbury, J.C. (2009). Peak athletic performance and ageing: evidence from baseball. *Journal of Sports Sciences*, 27(6), 599-610.

Brander, J.A., Egan, E.J. and Yeung, L. (2014). Estimating the effects of age on NHL player performance. *Journal of Quantitative Analysis in Sports*, 10(2), 241-259.

Chen, T. and Fan, Q. (2018). A functional data approach to model score difference process in professional basketball games. *Journal of Applied Statistics*, 45(1), 112-127.

Cockcroft, T.H. (2023). Fantasy football: What age do players peak/decline? *ESPN Fantasy*, Accessed September 29, 2023 at https://www.espn.co.uk/fantasy/football/story/_/id/37933720/2023-fantasy-football-players-peak-decline-quarterback-running-back-wide-receiver

Dendir, S. (2016). When do soccer players peak? A note. *Journal of Sports Analytics*, 2(2), 89-105.

Diamond, D. (2000). *Total Hockey (2nd ed.)*, Total Sports Publishing.

Distefano, G. and Goodpaster, B.H. (2018). Effects of exercise and aging on skeletal muscle. *Cold Spring Harbor Perspectives in Medicine*, 8(3), 1-15. a029785.

EvolvingWild (2017). A new look at aging curves for NHL skaters. *Hockey Graphs*, Accessed September 5, 2023 at https://hockey-graphs.com/2017/04/10/a-new-look-at-aging-curves-for-nhl-skaters-part-2/

Fair, R.C. (2008). Estimated age effects in baseball. *Journal of Quantitative Analysis in Sports*, 4(1), Article 1.

Guan, T., Nguyen, R., Cao, J. and Swartz, T.B. (2022). In-game win probabilities for the National Rugby League. *The Annals of Applied Statistics*, 16(1), 349-367.

James, B. and Henzler J. (2002). *Win Shares (1st ed.)*, STATS Pub.

Kubatko, J. (2010). Calculating point shares. *Hockey-Reference.com*, Accessed September 5, 2023 at https://www.hockey-reference.com/about/point_shares.html

Lichtman, M. (2009). How do baseball players age (Part 2)? *The Hardball Times*, Accessed September 5, 2023 at https://tht.fangraphs.com/how-do-baseball-players-age-part-2/

McHale, I. (2023). A flexible mixed model for the relationship between age and performance: Application to football, golf and snooker. *10th MathSport International Conference*, Budapest, Hungary. Abstract accessed September 25, 2023 at https://drive.google.com/file/d/19lGMQ0j0j3365mHEHiz6YpyxmzJS9nE2/view

Mlakar, M. and Tušar, T. (2015). Analyzing and predicting peak performance age of professional tennis players. In *Proceedings of the 13th International Symposium on Operational Research*, 99-104.

Nguyen, Q. and Matthews, G.J. (2023). Filling in the gaps: A multiple imputation approach to estimating aging curves in baseball. Accessed September 10, 2023 at https://arxiv.org/pdf/2210.02383.pdf

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis, 2nd Edition*, Springer Series in Statistics, Springer: New York.

Roring, R.W. and Charness, N. (2007). A multilevel model analysis of expertise in chess across the life span. *Psychology and Aging*, 22(2), 291-299.

Saikia, H., Bhattacharjee, D. and Mukherjee, D. (2019). Impact of age on performance of the cricketers. In *Cricket Performance Management*, Indian Statistical Institute Series. Springer, Singapore, 129-156.

Schuckers, M., Lopez, M. and Macdonald, B. (2023). Estimation of player aging curves using regression and imputation. *Annals of Operations Research*, 325, 681-699.

Shi, H., Dong, J., Wang, L. and Cao, J. (2021). Functional principal component analysis for longitudinal data with informative dropout. *Statistics in Medicine*, 40(3), 712-724.

Swartz, T.B., Arce, A. and Parameswaran, M. (2013). Assessing the value of draft positions in Major League Soccer's Superdraft. *The Sport Journal*, 16, Article 9.

Turtoro, C.J. (2019). Flexible aging in the NHL using GAM. Accessed September 9, 2023 at https://rstudio-pubs-static.s3.amazonaws.com/546176_89d68d30b12e47a1aa650529dd0addd2.html

Villaroel, C., Mora, R. and Gonzalez-Parra, G.C. (2011). Elite triathlete performance related to age. *Journal of Human Sport and Exercise*, 6(2), 363-373.

Wakim, A. and Jin, J. (2014). Functional data analysis of aging curves in sports. arXiv preprint arXiv:1403.7548.

Young, W.A. and Weckman, G.R. (2008). Evaluating the effects of aging for professional football players in combine events using performance-aging curves. *International Journal of Sports Science and Engineering*, 2(3), 131-143.

Wang, J.L., Chiou, J.M. and Müller, H.G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3, 257-295.