NHL Aging Curves Using Functional Principal Component Analysis

by

Elijah Cavan

M.Sc., Wilfrid Laurier University, 2019
B.Sc., University of Waterloo, 2017

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2022

Declaration of Committee

Name:	**Elijah Cavan**
Degree:	**Master of Science**
Project title:	**NHL Aging Curves Using Functional Principal Component Analysis**
Committee:	**Chair: Wei (Becky) Lin** Lecturer, Statistics and Actuarial Science
	Tim B. Swartz Co-Supervisor Professor, Statistics and Actuarial Science
	Jiguo Cao Co-Supervisor Professor, Statistics and Actuarial Science
	Sonja Isberg Committee Member Lecturer, Statistics and Actuarial Science
	Haolun Shi Examiner Assistant Professor, Statistics and Actuarial Science

Abstract

All major league sports teams are interested in projecting the performance of their players into the future. The seemingly most important feature of a model to project future performance is age. On average, players tend to improve from their rookie (earliest) season in the league, until they retire from the league (due to poor performance or injuries, for example). In this project we apply Functional Principal Component Analysis (FPCA) to the careers of NHL players in order to fit individual aging curves for each player. We compare the results of three methods: ImFuncPCA, SOAP and PACE.

Keywords: Functional Data Analysis; Sports Analytics; Aging Curves; Principal Components Analysis; National Hockey League

Acknowledgements

To my supervisors Jiguo and Tim, my family back in Ontario, and my friends from Vancouver: Brendan, Kim, Ryker, Gurashish, Robyn, Tim, Kaitlin, Grant, Rebecca, Renny, Mandy, Alice and others.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Many decisions made by the front office staff of a major league sports club are predominantly concerned with projecting the future performance of players. In the case of drafting younger players, teams would like to predict a player's future performance given their abilities relative to a player's peers in the junior leagues. The decisions the front office staff make based on trades and free-agency, however, are quite different. In this case, 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 the roster of a team is fixed. The composition of players is constantly changing as older players retire and younger players take on more prominent roles. Therefore, it is of interest to teams to approximate the value a player is likely to provide as they age; we call the models, aging curves. Aging curves can be difficult to study in team sports due to the fact that the performance of players is highly dependent on their teammates and the number of minutes they play; thus we are faced with the problem of multicollinearity, where the independent variables in the model are correlated. Aging curves have been studied in a number of sports including: golf [1], baseball [8], soccer [6], hockey [4] and basketball [7].

The effect of aging on the body is a common issue for all athletes [15]. For most major sports leagues, players reach their peak performance between ages 25-29, and 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 [6] and [8]. In [8], the authors show that different skills decline at different rates; baserunning, for example, declines at a much faster rate as compared to power in baseball. Most research reaches the general consensus that these age effects are positional dependent; for example in [4] the authors conclude that the peak age for NHL forwards is between ages 27-28, while the peak age of performance for defencemen is between 28-29. Each player has a unique aging curve (due to different body composition or previous athletic history, for example)- but there should

be some agreement between players who have similar styles of play. Previously, the most common way to build an aging curve was to calculate the difference between a player's performance between years that they have played in the league, and then average over a number of players (this is called the "Delta Method" [20]). A feature of the Delta Method is that it is not impacted by the differences in quality between players. Some research attempted to expand on the Delta Method by projecting forward for players who have left the league (creating so called "phantom players") [21].

There is an obvious risk in using the Delta Method, in that the tails of the age distribution (players who begin their careers at a very young age and players who stay in the league until a very old age) are bound to suffer from small sample sizes. This is called survivorship bias, and is a special case of the more general problem of selection bias in statistics. Selection bias is a systematic statistical error caused by drawing a non-random sample from a population. In the case of aging curves, samples of player performances are non-random because only the most talented players enter the league significantly sooner than the modal age, and these are usually the same players who stay in the league the longest (except for the case of early career ending/altering injuries). A number of different techniques have been studied to account for selection bias. For example: [12] uses modern imputation techniques, [14] uses a multilevel Bayesian model, [4] uses a fixed effects regression model and [1] uses a random effects model.

In this MSc. project, we attempt to build an aging model using Functional Principal Component Analysis (FPCA). This approach has been taken by [7], with the authors making use of the PACE package. One benefit of this method is that we do not need to specify the functional relationship of the aging model; whereas many papers, such as [4] assume a quadratic or cubic [13] functional form for the model. In FDA, we can use an arbitrary number of basis spline functions to determine functional relationships. Another benefit of FDA models is that we are able to fit separate aging curves for each player since each player has an individually estimated principal component score which is added to the mean function. Furthermore, we can form clusters from these principal component scores, allowing us to compare players in the sample.

1.2 Data

The data studied in this project was scraped from Sports Reference LLC at `https://www.hockey-reference.com` [23]. It contains summary statistics (goals, points, games played, etc) for each player in the National Hockey League (NHL) from the years 1920-2022. There are $n = 7393$ unique players in our dataset, with a total of $p = 43689$ seasons worth of data;

and there is a total of approximately 52 thousand rows in our dataset before adjusting for duplicated rows (due to a player changing teams mid season). The observed maximum player age is 51, and the minimum observed player age is 17. From an alternative study,[24] the average career length is 4.5 years.

There are limitations on the statistics available for modelling because the NHL only started to track individual shots and plus-minus statistics since 1960-1961; and time on ice (TOI) was not tracked for individual players until the 2000-2001 NHL season [23].

1.3 Measures of Player Value

As a measure of the value of a player we used the metric "point shares" (PS) [23]. It is a measure derived from a metric created by Bill James in 2002 [22] that was originally used to evaluate baseball players. Here "points" refers 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, and thus can be thought of as an earlier version of WAR (wins above replacement) for hockey. This metric was chosen because we wanted a metric that takes into account a composite measure of performance, and adjust for the linemates that an individual player plays alongside. The goals and assists metrics are not reflective of contributions to the team such as defense; and would overvalue forwards (who's principal contribution to a team is to score goals) compared to defensemen.

The point shares metric is adjusted so that a hockey team with 100 team points (representing 50 wins, or 40 wins and 10 overtime losses, etc.) in the standings, will have players whose individual point shares sum to 100. Players may have negative point shares. Negative point shares would indicate that you are losing your team points relative to a replacement level player. Point shares are obtained from both offensive and defensive components point shares. To calculate offensive point shares for each player we calculate goals created (a weighted sum of a players goals and assists divided by team goals and assists), adjust for the minutes played by the player and adjust for the league environment (league goals divided by league points). There are different positional adjustments for forwards compared to defensemen. The point shares metric is obtained from `https://www.hockey-reference.com` and more details of it's calculation are found in [23].

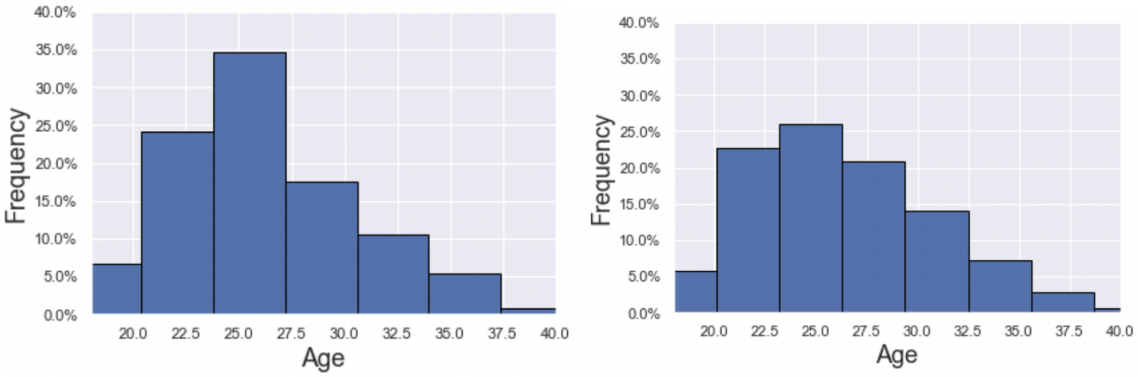
The rest of the project is organized as follows: in Chapter 2, we perform exploratory data analysis to investigate the distribution of key variables. In Chapter 3, we outline the methods we will use to model the problem. We briefly outline some of the underlying

mathematical background required to understand Functional Data Analysis. In Chapter 4, we present the results of our modelling. In Chapter 5, we discuss some of insights provided by the modelling.

Chapter 2

Exploratory Data Analysis

In this chapter, we present some plots and tables to investigate the distribution of key variables. As this project is concerned with developing models to simulate aging curves, we are principally interested in the distribution of related variables across the sampled seasons (1920-2022). Table 2.1 presents some summary statistics for the age and career PS variables in our dataset based on the two position groups - forwards and defensemen. Note that the NHL requires 12 forwards and 6 defensemen on the roster for each game; this fact, along with the different responsibilities of the two positions, leads to differences in the distributions of these key variables. Figure 2.1 shows the frequencies of ages for the two positional groups we are studying. The number of observations is 12766 for defensemen and 24970 for forwards. The modal age for both positions is between 25-27 years. It is apparent that defensemen play to more advanced ages than forwards. We show the the distributions of the starting (rookie) and ending (retirement season) ages of NHL players in Figures 2.2 and 2.3, respectively. Most players tend to start around 20-21 years of age; to be eligible for the NHL draft, players must be at least 18 years old, and younger than 20 years - hence we can conclude that most drafted NHL players start their careers a couple years after



(a) Forwards

(b) Defencemen

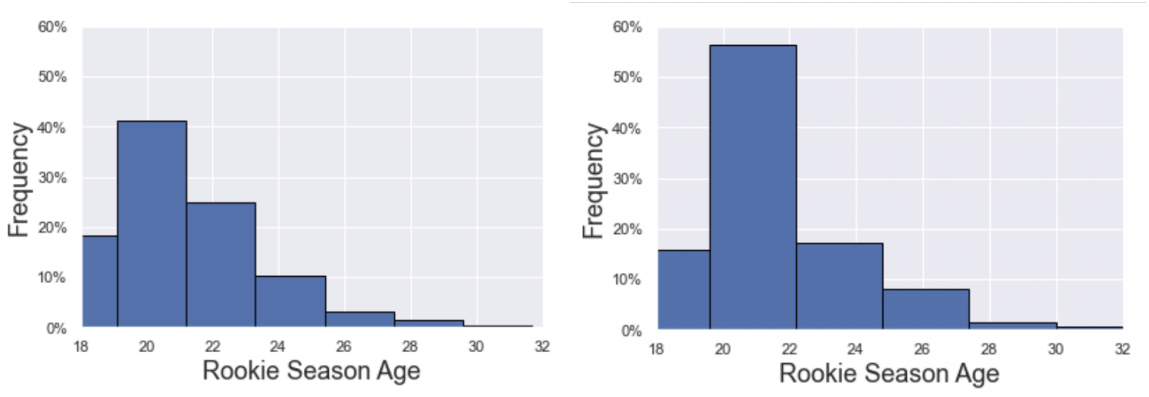
Figure 2.1: Histograms of age by position.

Position	Variable	Mean	Max	Min
FWD	Age	26.17	51	17
	PS	27.44	250.9	-10.6
DEF	Age	26.54	48	17
	PS	36.78	242.5	-6.4

Table 2.1: Summary statistics grouped by position.

they have been drafted. They spend these early post-draft years developing in the minor leagues. The median rookie age for both positions is 21 years. The age distribution for the retirement age shows significant variance; while the distribution is clearly centered between ages 30-35 years. The median retirement ages are 31 and 32 years for forwards and defensemen respectively. We see there are NHL players who end their careers quite early (due to poor performance or injuries) and in rarer cases, NHL players who end their careers past age 36 years. Note that the early "retirement" ages often correspond to players who are deemed not sufficiently "good" to play in the NHL. We would assume the older age players are NHL stars who have garnered much respect in the league, and who's careers have been relatively free from injuries or poor performance. This point is emphasized in Figure 2.4; the distribution for career point shares is clearly left skewed - with the majority of players ending their careers with between 0-20 point shares. Career PS is the sum of a player's PS for each season he has played in his career, which would reward players with longer careers. If a player is able to consistently perform at a high level, while avoiding career threatening injuries, then they are likely to continue playing in the NHL.

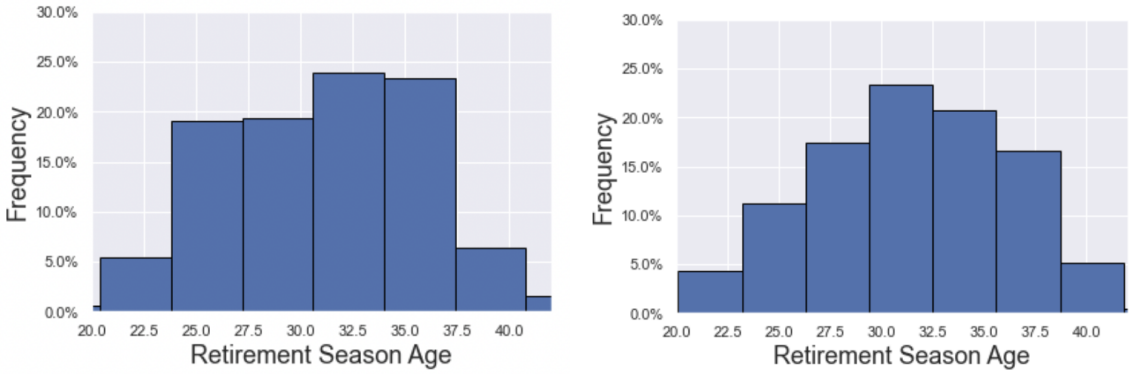
We are also concerned with the number of samples we get from each player as illustrated in Figure 2.5. Some NHL players may have 20-year careers, while others may have relatively short careers. We are also interested in visualizing the average minutes played and average



(a) Forwards

(b) Defensemen

Figure 2.2: Histograms of rookie age by position.



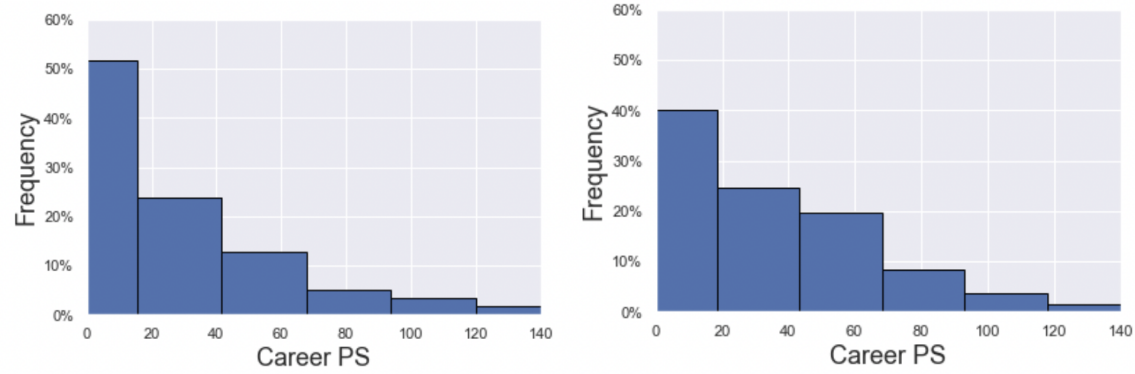
(a) Forwards

(b) Defencemen

Figure 2.3: Histograms of retirement season age by position.

point shares for players grouped by age, this is presented in Figure. 2.6. The selection bias in the sample is evident due to the fact that the mean function is not a smooth curve - which is what we would expect if we had an equal number of samples for each age. As seen in Figure. 2.1, very few players play past 36 years; and the ones who do are highly skilled players - we would expect the plot to increase to a maximum at age 27, and then show a decline. However, Figure 2.6 shows that the average PS increases past age 40, and then is subjected to an aggressive decline. Figure 2.6 highlights the issue of the naive approach of calculating the aging curve by simply averaging player performances by age.

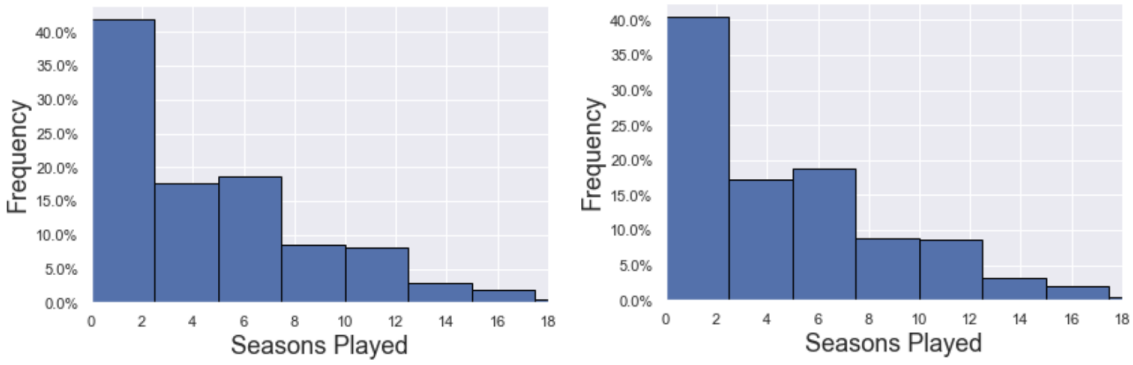
The final plot for this section is descriptive of the pattern we wish to predict from our model. Figure 2.7 presents the point shares trajectories for two forwards and two defensemen. One can see that the player with higher PS stayed in the league later (and experienced a much shallower decline in performance) relative to the less skilled players who exited the league at a much earlier age.



(a) Forwards

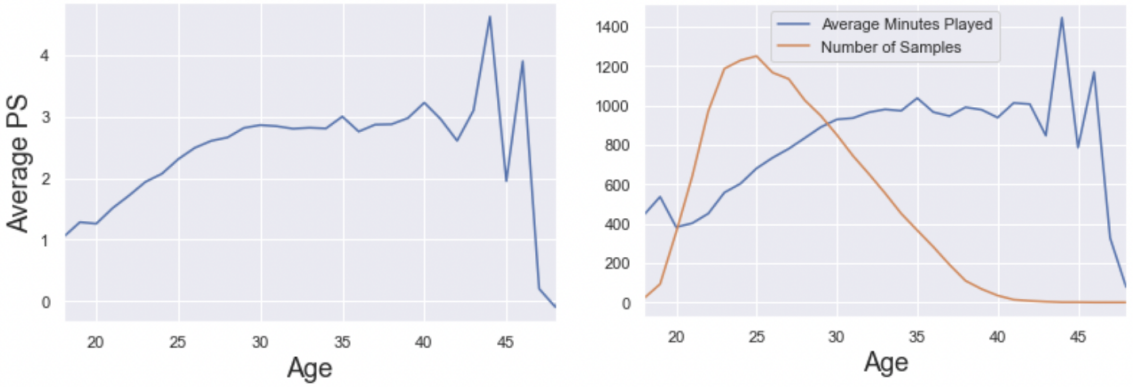
(b) Defensemen

Figure 2.4: Histogram of Career PS by position.



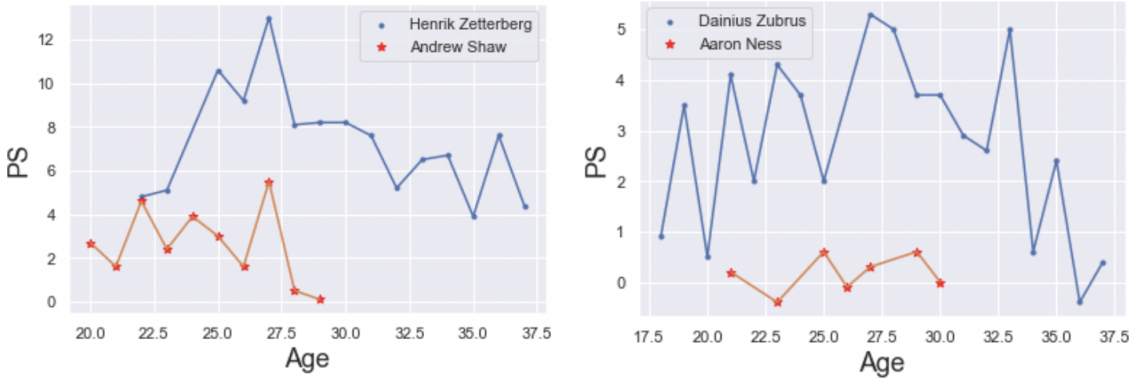
(a) Forwards (b) Defencemen

Figure 2.5: Histogram of seasons played by position.



(a) Plot of average point shares versus age. (b) Plot of average minutes played versus age.

Figure 2.6: Average PS (left) and minutes played (right) by age.



(a) Forwards (b) Defencemen

Figure 2.7: Career trajectories for two players by position.

Chapter 3

Methodology

3.1 Functional Data Analysis

3.1.1 Overview

Functional Data Analysis (FDA) is a highly flexible modelling technique which is concerned with the modelling of longitudinal or repeated measurements data. To be more specific, the models are appropriate for tracking the same sample at different points in time (i.e. each sample is a function of one or several variables). It is a modern approach to multivariate statistical modelling, with many applications as seen in [19]. A review of the current advances in the topic can be found in [11]. Recently, FDA has been used as an application for sports data, as seen in [10], where it is used to specify conditional distributions for the in-game win probability in rugby; and in [9], where FDA is used to model the score difference process in basketball.

FDA is different from more well known techniques such as time series analysis due to the fact that there are no underlying assumptions about the stationary of the underlying trajectories being studied. It is different from multivariate statistics, where each observation is a vector of observed values, because we can more readily perform analysis where nearby values are correlated. There is an underlying assumption (for the case of multiple subjects) in FDA that the observed samples are independent stochastic processes (i.e. the observation from one subject does not influence observed values from the other subjects). FDA is also uniquely built to be able to handle sparse data- i.e. the case of missing data at one or more time steps of a given subject. Another feature of FDA is that it allows for clustering of repeated measurements.

FDA can be used to perform a number of common machine learning tasks such as classification, clustering, ANOVA, regression, principal component analysis, interpolation/extrapolation and registration- the difference being we replace the more common point estimator with a functional estimator. The benefits of FDA include being able to represent the observed

data as smooth functions, dimensional reduction of the observed data and the ability to compute derivatives of the smooth estimator. The are numerous packages in R and Python that can be used to perform FDA; the common ones being scikit-fda (python) and the R package fda available on CRAN.

3.1.2 Mathematical Background

When performing FDA, we consider the set of observations from each individual to be a random, smooth function. Hence we can do statistics on the set of random curves coming from each individual we observe, rather than looking at the individual observations. For example, suppose that we observe data from $i = 1, ..., N$ players, and from each player we observe m data points $(y_{i1}, ..., y_{im})$ at time points $t_1, ..., t_n$; then the core assumption in FDA is that the observed data y_{ij} for individual i and observation j is expressed as:

$$y_{ij} = X_i(t_{ij}) + \epsilon_{ij} \tag{3.1}$$

where we assume ϵ_{ij} are independent and normally distributed with mean 0 and variance σ^2. Formula (3.1) explicitly assumes that all the observations from individual i can be modelled by a single stochastic function $X_i(t)$ after accounting for measurement error (ϵ_{ij}). The number of observed points m observed from each player can vary from player to player; this is called irregular or sparse data. In order to approximate the functions $X_i(t_{ij})$, we express them in terms of Q spline basis functions $b_1(t), .., b_Q(t)$. Spline functions are piece wise polynomials joined at specific points, called knots. The number of knots, τ, is determined by $\tau = p + Q - 2$, where p is the order of the polynomial and Q is the chosen number of basis functions. We write each function for the individual players as

$$X_i(t) - \mu(t) = \sum_{k=1}^{\infty} \alpha_{ik} b_k(t) \approx \sum_{k=1}^{Q} \alpha_{ik} b_k(t) \tag{3.2}$$

for sufficiently large Q. The basis functions $b_k(t)$ are evaluated at τ knot points which span the range of t. Hence the centered process, $X_i(t) - \mu(t)$, is specified by a linear combination of spline basis functions, and is fully determined by the Q coefficients $\alpha_{i1}, ..., \alpha_{iQ}$. Often we wish to smooth the approximated function using a parameter λ chosen using cross validation. To estimate $\alpha_{i1}, ..., \alpha_{iQ}$ our loss function is the least squares cost function.

$$\hat{\alpha}_{ik} = \underset{\alpha_{i1},...,\alpha_{iQ}}{\arg\min} \sum_{j=1}^{m} \left(y_{ij} - \mu(t) - \sum_{k=1}^{Q} \alpha_{ik} b_k(t) \right)^2 \tag{3.3}$$

Once the α_{ik} have been estimated, we can compute common summary statistics such as the mean function

$$\hat{\mu}(t) = E[X_i(t)] = \frac{\sum_{i=1}^{N} X_i(t)}{N} \tag{3.4}$$

and the variance-covariance function.

$$\hat{v}(s,t) = \frac{1}{N}\sum_{i=1}^{N}\left(X_i(t) - \mu(t)\right)\left(X_i(s) - \mu(s)\right) \tag{3.5}$$

The problem with simply using spline basis functions to model the aging curves of each player is twofold:

1. We do not have enough observation points for each player (sparsity) which would result in poor spline fits for these players.

2. We can not use the spline fits for unobserved players (i.e. prediction).

A more robust model that does not have the issues above is Functional Principal Component Analysis (FPCA).

3.2 Functional Principal Component Analysis (FPCA)

Principal Component Analysis (PCA) is a technique used when the number of variables/features (m) in a model is large compared to the number of observation points (N); or when several variables are highly correlated. A helpful review is found in [16]. We observe m-dimensional vectors $y_{i1}, .., y_{im}$ for each of the $i = 1, ..., N$ subjects. The observation points form a $N \times m$ design matrix, X. We aim to reduce the dimensionality of the model by projecting the observed values onto a smaller subspace of variables. To do this, we look for the direction within the data that explains the most variability in the data by projecting the data onto a unit vector such that the variability of points about the unit vector is less than the variability of points around the m-dimensional mean. The problem becomes an eigenvalue problem, where the eigenvectors are called the principal components of the data, and the eigenvalues are called the principal component scores.

For FPCA, the data is now in the form of random curves, $X_i(t), .., X_N(t)$; where each curve is sampled at a maximum of m points, for example, $X_1(t_1), .., X_1(t_m)$. To move to the case of functional data, we simply replace the discrete sums in the multivariate PCA formulation, with integrals which represents the continuous nature of the observed values. We reformulate the multivariate eigenvalue problem to be:

$$\int v(s,t)\xi(t)dt = \rho\xi(s) \tag{3.6}$$

where $\xi_(s)$ is the eigenfunction with corresponding eigenvalue ρ. The eigenfunction is called the Functional Principal Component (FPC) and the eigenvalue is the Functional Principal Component Score (FPC score), which is individually estimated for each subject. To find the

principal component score we solve:

$$\rho_{ij} = \int \xi_j(s) \left(X_i(s) - \mu(s) \right) ds \tag{3.7}$$

for the ith subject and the jth principal component. Equation (3.7) is solved subject to the constraints of unit norm, $\int \xi_i(t)\xi_j(t)dt =< \xi_i, \xi_i >= 1$, and orthogonality between eigenfunctions, $< \xi_i(t), \xi_j(t) >= 0$; where $< ., . >$ denotes the standard inner product between functions. Here the FPC scores are given by ρ_{ij}. FPCA seeks to maximimize the variation in these FPC scores. Equation (3.7) can be solved with numerical integration if there is no sparsity in the data. We will discuss methods to determine the FPC scores in the case of sparse data in following sections. Once we have determined the FPCs, one can show that we can express each observation by:

$$X_i(t) = \mu(t) + \sum_{j=1}^{\infty} \rho_{ij}\xi_j(t) \tag{3.8}$$

where $\mu(t) = E[X(t)]$ is the mean function for the observed curves. This formula (called the Karhunen–Loève expansion) would be ineffective, unless we could truncate the infinite sum to look at the top K eigenfunctions from the FPCA. These k eigenfunctions seek to minimize

$$\frac{1}{N} \left(\sum_{i=1}^{N} \int \left[X_i(t) - \hat{\mu}(t) - \sum_{j=1}^{k} \rho_{ij}\xi_j(t) \right]^2 dt \right) \tag{3.9}$$

which is the sum of squares loss function. As in the multivariate case, we choose the value of k by calculating the explained variance of the first K FPCA functions and by setting a cutoff at a given tolerance (for example- 90% of the explained variation). The explained variance of the mth eigenfunction, π_m, is given by

$$\pi_m = \frac{\mathrm{Var}(\xi_{im})}{\sum_{k=1}^{K} \mathrm{Var}(\xi_{ik})} \tag{3.10}$$

we can then write $\sum_{m=1}^{K} \pi_m$ to get the cumulative explained variance for the first K eigenfunctions; which allows us to select the number of PC components for our modelling.

3.3 Methods to Determine the Functional Principal Component Scores (FPCs)

3.3.1 Principal Analysis by Conditional Expectation (PACE)

The packages mentioned above (scikit-fda, fd) do not allow for the use of FDA on data sets with missing (sparse) data. The Principal Analysis by Conditional Expectation (PACE) package [5] in R allows for the functional analysis of data that has been generated by data

that is not fully observed. The algorithm uses a local regression estimator to determine the covariance structure and the variance of the measurement error, followed by eigendecomposition of the covariance function to obtain the estimates of FPCs, and calculation of the FPC scores through conditional expectation.

$$\hat{\rho}_{ij} = E[\xi_j | y_{ij}] = \hat{\sigma}_j \hat{\xi}_j \hat{\Sigma}_i^{-1} (y_{ij} - \hat{\mu}) \tag{3.11}$$

where $\hat{\sigma}_j$ is the estimated variance of the jth subject, $\hat{\mu} = (\hat{\mu}(t_1)..\hat{\mu}(t_m)))$, $\hat{\xi}_j = (\hat{\xi}_j(t_1)..\hat{\xi}_j(t_m)))$. Equation (3.11) represents the Best Linear Unbiased Predictor (BLUP) for the functional data. Because this method requires the inverse of the covariance function (through the eigendecomposition), it may be numerically unstable in some instances. The other main issue of this method is that it requires the assumption that the FPC scores are normally distributed. We will see below two alternatives to this method which attempt to correct for this.

3.3.2 Sparse Orthonormal Approximation (SOAP)

In [3], the authors seek to find the optimal empirical basis functions to approximate the centered (or de-meaned) stochastic process $X_i^\star(t) = X_i(t) - \mu(t)$; these empirical basis functions are taken to be the eigenfunctions calculated from the FPCA analysis. The main benefit of the approach the authors take is that they are able to approximate the uncentered stochastic process $X_i(t)$ through calculation of the eigenvectors of:

$$K(s,t) = E[X_i(s)X_i(t)] = \sum_i \lambda_i \xi_i(s) \xi_i(t) \tag{3.12}$$

which is the so called Mercer kernel. This function is slightly different from the usual covariance function, but the authors state a theorem showing that the uncentered process $X_i(t)$ can be approximated as a finite sum of the eigenvectors (empirical basis functions) of the estimate $\hat{K}(s,t) = \frac{1}{N} \sum_i X_i(s)X_i(t)$.

$$X_i(t) = \sum_{j=1}^{\infty} \rho_{ij} \xi_j(t) \tag{3.13}$$

where $\rho_{ij} =< X_i(t), \xi_j(t) >$. As previously stated, this method does not require the estimation of the mean function (and therefore does not require a centering step), and does not require eigendecomposition of the sample covariance function- which may be difficult when the sampling points for each process $X_i(t)$ is sparse. The eigenvalues of this mercer kernel estimate, ρ_{ij}, are called the FEC (Functional Empirical components) scores, which are equivalent to the FPC scores in the case of a zero-mean process.

The process by which the FEC scores are estimated is outlined as follows (where n is the number of subjects and m is the number of observed samples for the ith subject):

1. Choose an initial value for the FEC $\xi_1(t)$ which satisfies the usual constraints (orthornormalility)
2. Obtain an estimate to the FEC score $\vec{\rho}_1 = (\rho_{11}...\rho_{1n})^T$ by minimizing

$$\frac{1}{N}\sum_{i}^{n}\frac{1}{m}\sum_{j=1}^{m}(y_{ij} - \rho_{i1}\hat{\xi}_1(t_{ij})) \tag{3.14}$$

   (this is equivalent to least squares estimation)
3. Given the curent estimate of the FEC score $\vec{\rho}_1$, update the estimate of $\xi_1(t)$ by minimizing the same loss function above
4. Continue the iterations until the desired threshold is achieved

The subsequent (2nd) FEC scores are approximated by replacing y_{ij} in the loss function above with the residual $\hat{r}^1_{ij} = y_{ij} - \hat{\rho}_{i1}\hat{\xi}_1(t_{ij})$. The subsequent J-2 FEC vectors are then determined by appending to this residual with the previously estimated FEC scores.

3.3.3 Informatively Missing Functional Principal Component Analysis (imFunPCA)

Methods (like PACE) used to calculate the FPCs from sparse data often assume the data is missing at random. For the case of the aging curves, as we have previously explained this is not the case; for example it is well known that the majority of players see a decline in performance and an increase of injury risk as they age. In [2] the authors adjust for this bias by proposing a likelihood approach to the imputation of missing data. By assuming the data is normally distributed, the authors show that the first eigenfunction can be calculated by maximizing

$$\prod_{i=1}^{n}\prod_{j=1}^{n_i}\phi(y_{ij};\hat{\mu}(t_{ij}) + \rho_{i1}\xi_1(t_{ij}),\sigma^2)^{\frac{1-\delta_{ij}}{n_i}}\Phi(c_{ij};\hat{\mu}(t_{ij}) + \rho_{i1}\xi_1(t_{ij}),\sigma^2)^{\frac{\delta_{ij}}{n_i}} \tag{3.15}$$

(A similar equation is used to find the estimate mean function $\hat{\mu}(t)$) subject to $|\xi_1|^2 = 1$. Here δ_{ij} is an indicator for whether the jth out of n_ith observation of individual i is missing ($\delta_{ij} = 1$) or not. ϕ is the Gaussian pdf and Φ the Gaussian cdf as per the usual convention. c_{ij} is chosen such that the missing data is assumed to smaller than c_{ij}. The eigenfunction $\xi_1(t)$ is expressed as a sum of B-spline basis functions:

$$\xi_1(t) = \sum_{i=1}^{S}\beta_{1,s}b_s(t) = \beta_1^T b(t) \tag{3.16}$$

The full algorithm is described as:

1. Choose an initial estimate for β_1, $\beta_1^{(0)}$ (and hence through the equation above we have an initial estimate for ψ_1)

2. Obtain an estimate for ρ_{i1} through maximization of:

$$\prod_{j=1}^{n_i} \phi(y_{ij}; \hat{\mu}(t_{ij}) + \rho_{i1}\xi_1^{(0)}(t_{ij}), \sigma^2)^{1-\delta_{ij}} \Phi(c_{ij}; \hat{\mu}(t_{ij}) + \rho_{i1}\xi_1^{(0)}(t_{ij}), \sigma^2)^{\delta_{ij}} \tag{3.17}$$

3. Conditional on the current estimate of ρ_{i1}, update the estimate for ξ_1 using the loss function above

4. Repeat 2-3 until desired convergence tolerance is reached

The subsequent J FPCs are estimated sequentially conditioned on the previous J-1 FPCs calculated through the algorithm supplied above.

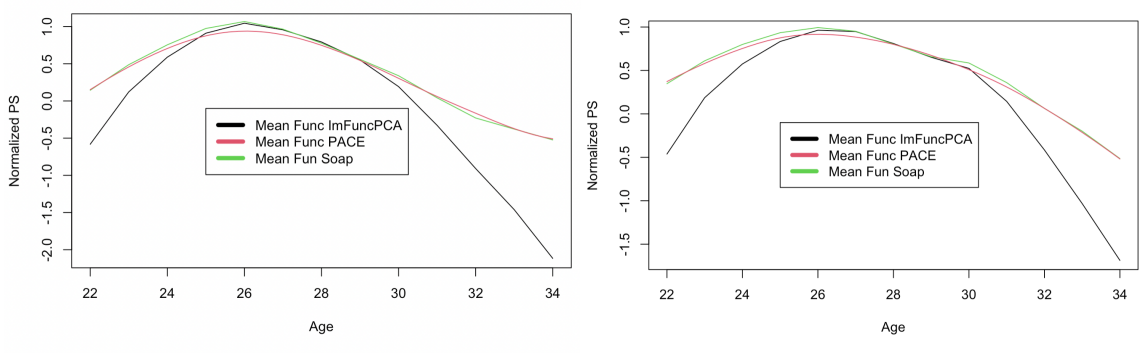
Chapter 4

Results

4.1 Functional Principal Component Analysis

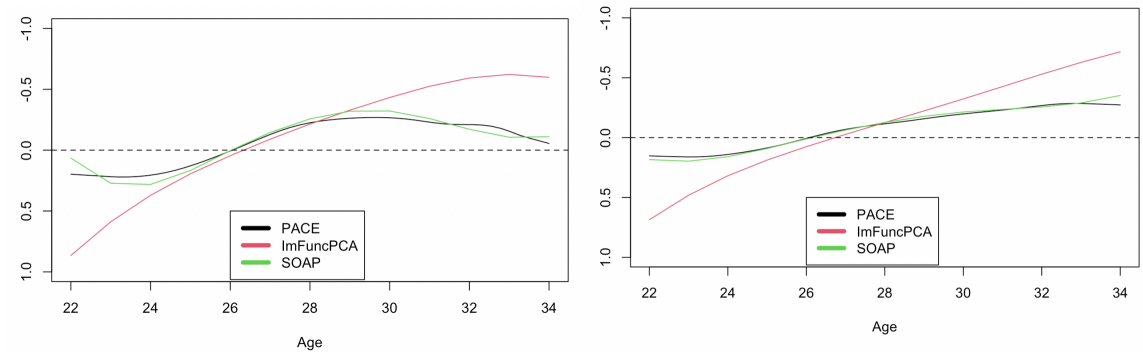
Following the approach described in Section 3.2 we subset the data to include only players ages 22 years to 34 years with a minimum threshold of 30 games played in a season. We also limit the data to players who had a career length of at least seven seasons. This provides us with 438 unique defensemen and 873 unique forwards. We further randomly partitioned the data into training (750 forwards, 370 defensemen) and testing (123 forwards, 68 defensemen) sets.

First, we compare the estimated mean function using the three methods outlined in Chapter 3 (PACE, SOAP, imFunPCA) on the training data in Figure 4.1. We use six spline basis functions of order four, resulting in $\tau = 8$ knot points. We normalize the target variable (PS) by subtracting the mean from each player. The normalization helps with the estimation of the mean function and FPC eigenfunctions but does not affect prediction of a player's career PS trajectory. The three methods all lead to aging curves (for both forwards and defensemen) which peak at 26 years and monotonically decline until 34 years. This point is further emphasized by Figure 4.2, which presents the derivative of the mean function for the three methods. In Figure 4.2, we see that each method has a derivative function which is zero at age 26 years - meaning each method agrees that age 26 years corresponds to the peak age of performance for players. The mean functions for the SOAP and PACE methods are very similar, although the SOAP method predicts a slightly higher peak compared to the PACE method. The imFunPCA method predicts a drastic decline, especially after age 30 years. The imFunPCA method is more susceptible to the decline in performance seen in most players. This is because the imFunPCA method uses a likelihood approach to impute the missing data at the tails of a career, and enforces the condition that the unobserved seasons should have a lower PS value than the minimum observed by the player.



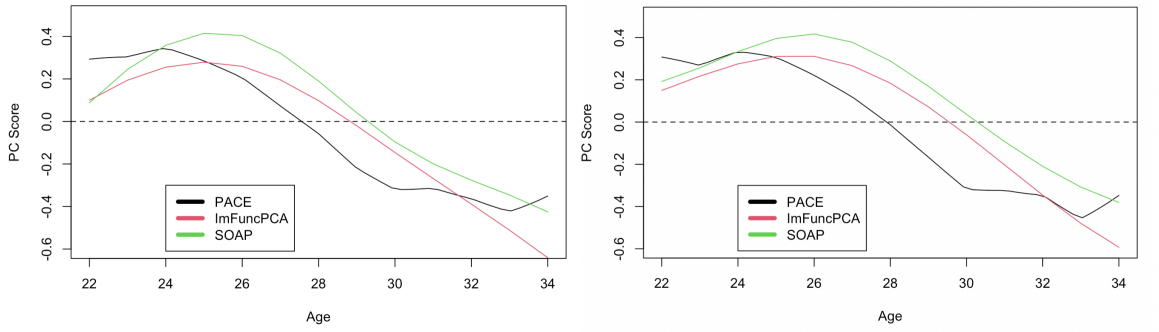
(a) Forwards (b) Defensemen

Figure 4.1: Comparison of the estimated mean function from three methods by position.



(a) Forwards (b) Defensemen

Figure 4.2: Comparison of the first derivative of the estimated mean function from three methods by position.



(a) Forwards

(b) Defensemen

Figure 4.3: Comparison of the first estimated FPCA eigenfunction from three methods by position.

Figures 4.3 and 4.4 show the calculated FPCA eigenfunctions for the three methods by position. Figure 4.3 is the first estimated FPC eigenfunction for the two position groups studied. We notice similar trends between the two positions. The first FPC eigenfunction is the weighted average of the data over the time interval. Figure 4.3 can be interpreted as the change in performance between the early and late career of a player. Each method predicts a different decline; the PACE method is positive over the interval (22 years, 28 years) and negative between (28 years, 34 years), versus the SOAP and imFunPCA methods which are positive between (22 years, 30 years) and negative otherwise. Hence, a player with a large positive ρ_{i1} would correspond to a player who performs extremely well in their early career (players who perform poorly in their early career would have negative ρ_{i1}).

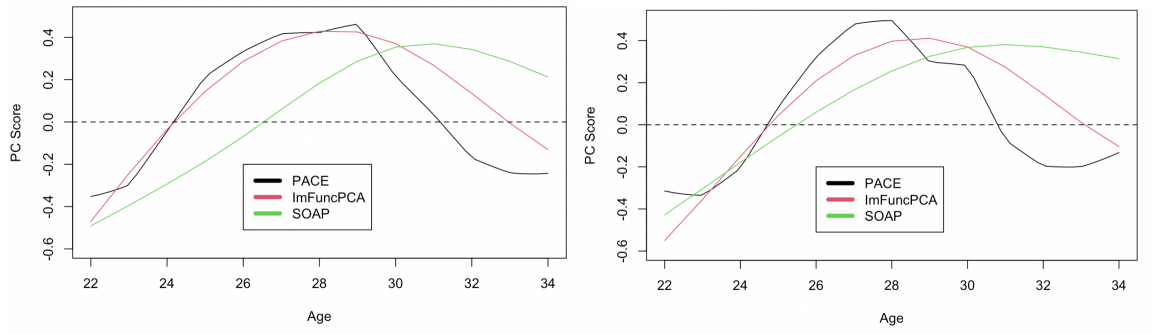
Figure 4.4 plots the second estimated FPC eigenfunction. For the PACE method the eigenfunction is positive between ages 24-30 years, and negative otherwise. The second FPC eigenfunction is interpreted as the difference between a player's peak age, and their early and late stages in their careers. Hence, players who peak late in their careers would have large ρ_{i2} versus players who peak early. The SOAP method is positive over the entire interval considered, and so the method would overestimate players with late career peaks.

4.2 Using the FPC Scores for Prediction

The prediction error for the ith player is given by the mean absolute error (MAE)

$$\text{MAE}_i = \sum_{j \in D} |y_{ij} - \hat{X}_i(t_{ij})| \tag{4.1}$$

where $j \in D$ are the observed ages, y_{ij} is the true performance (point shares for age j years) and $\hat{X}_i(t_{ij})$ is the predicted value at age j years. While the mean function and



(a) Forwards (b) Defensemen

Figure 4.4: Comparison of the second estimated FPCA eigenfunction from three methods by position.

FPC eigenfunctions are estimated with the normalized PS, prediction is done on the raw PS values for the player. We measure the performance of each FDA model as the average prediction error for the n players in the testing set

$$\text{Test Error} = \frac{1}{n}\sum_{i=1}^{n} \text{MAE}_i \tag{4.2}$$

To make predictions on the test set we implement the following procedure:

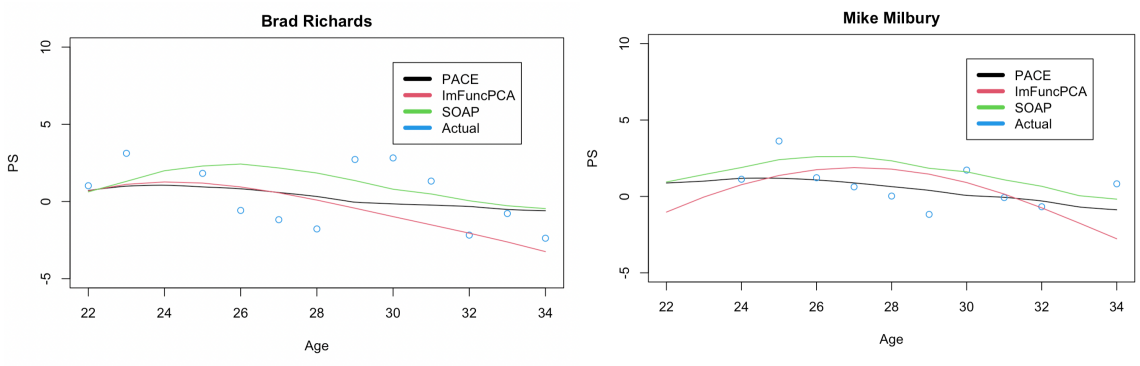
1. Estimate the mean function, FPC eigenfunctions and FPC scores from the training data.

2. Regress the $j = 2$ eigenfunctions, $\hat{\xi}_j$ on the observed performance of the player minimizing the sum of squared errors; the coefficients from the linear regression are the FPC score estimates, $\hat{\rho}_{ij}$, for the test player. See Section 3.2 for a review of FPCA.

3. Predict the future performance, $\hat{X}_i(t)$, of the player using the mean function and Karhunen–Loève expansion

$$\hat{X}_i(t) = \hat{\mu}(t) + \sum_{j=1}^{k} \hat{\rho}_{ij}\hat{\xi}_j. \tag{4.3}$$

Table 4.1 below compares the prediction error per player for the three methods we have explored. The error corresponds to the difference in career point shares between the prediction and actual results. Hence a test error of 10 corresponds to a model being 10 career point shares off from the true performance of the player (since we use mean absolute error, 10 could be an overstimate or an underestimate from the model). For a player who plays for 10 years in the NHL, this error would correspond to a difference in one PS per year between the predicted and actual performance. To compare to the baseline prediction, we include the testing error from using the delta method [20] calculated from the same training



(a) Forwards (b) Defensemen

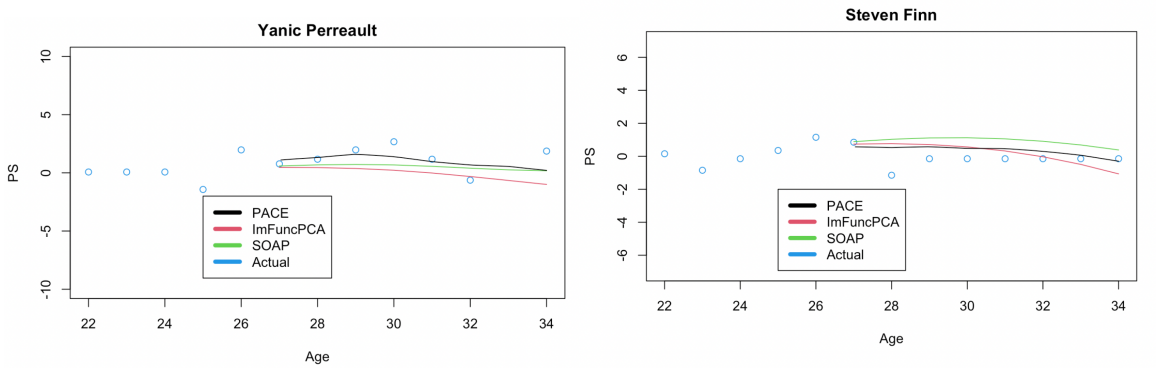
Figure 4.5: Comparison of training set prediction from the three methods by position.

and testing datasets. Five out of six of the FDA models outperform the baseline prediction. Figures 4.5 demonstrates the training prediction on a random player from each position group.

Figure 4.6 and Figure 4.7 present prediction on the test set, where by we use the first six seasons of a player's performance to predict their aging curve until their age 34 season. This is one of the major benefits of the FPCA analysis, we can approximate the FPC scores from a player's first six seasons, and then use the estimated mean curve and estimated FPC eigenfunctions to forecast the player's future performance. In Figure 4.6 the players Yanic Perreault and Steven Finn display little variation in their career performance in the first six seasons; and the FDA models correctly forecast a moderate decline in performance as they age. In Figure 4.7, Teemu Selanne and Ryan Whitney are both high performing players, and their projected decline is more extreme compared to the players in Figure 4.6.

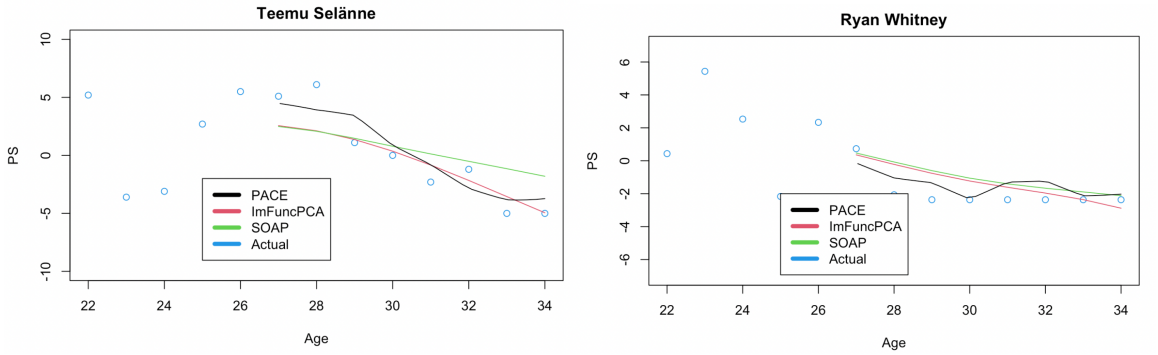
Position	Method	Error
FWD	PACE	18.65
	SOAP	16.29
	imFunPCA	21.51
	Delta	19.24
DEF	PACE	16.95
	SOAP	17.59
	imFunPCA	14.73
	Delta	18.27

Table 4.1: Performance of the different models.



(a) Forwards (b) Defensemen

Figure 4.6: Comparison of test set prediction on player one from the three methods by position.



(a) Forwards (b) Defensemen

Figure 4.7: Comparison of test set prediction on player two from the three methods by position.

4.3 Cluster Analysis of the FPC scores

Cluster analysis is an unsupervised learning technique where by we attempt to form groups of subjects that share common characteristics. If the clustering algorithm is effective, then one would see significant variation in between groups. We use the FPC score estimates from the PACE method for each position in order to cluster different players. We chose the KMeans clustering algorithm from python's scikit-learn library [18] using the first two FPC scores as features, to partition players into three groups. The KMeans algorithm endeavours to subset N samples into K groups of equal variance such that the within-cluster sum of squares loss function, WCSS, for each cluster C

$$\text{WCSS} = \sum_{i=1}^{N} \min_{\mu_j \in C}(||\rho_i - \vec{\mu}_j||^2) \tag{4.4}$$

is minimized [18]. Here ρ_i corresponds to a single observation with the two estimated FPC scores as features; μ_j is the cluster mean for cluster j, also called the centroid. The cluster averages for each position are shown in Table 4.2.

The FPC scores represent the only two features in the KMeans clustering model, and yet they do a good job of forming separable clusters. For the forward position group in particular, we notice that players from cluster three play longer, retire at a later age and produce more career point shares as compared with the players in cluster one and two. There is a similar differentiation between the defensemen clusters, but the between cluster variation is smaller. Accurately predicting the future performance of a NHL player is a difficult problem; by assigning players to different clusters it is easier for the front office staff to estimate the trajectory of a player. For forwards, a player assigned cluster two would be expected to play 600 games and retire with 22 career PS, versus a player assigned to cluster three who is estimated to play 900 career games and retire with 62 career PS. By clustering the players we can get a rough estimate of how their careers will play out without having to accurately forecast their performance. This is especially relevant for players who are just

Position	Class	Career Games Played	Career PS	Retirement Age
FWD	1	814	53.5	33.2
	2	633	21.7	31.9
	3	923	62.9	35.6
DEF	1	628	36.9	32.1
	2	854	64.7	35.7
	3	810	61.4	33.8

Table 4.2: Cluster averages by position.

(a) Forwards (b) Defensemen

Figure 4.8: Clustering the FPC scores by position.

starting out in their career or players who have missed a number of seasons due to injuries.

Figure 4.8 is a scatter plot demonstrating the different clusters graphically, plotted against the two features (the two FPC scores estimated for each player). For the forward position group, players in the top right of the plot (cluster one) have larger FPC scores and correspond to players who decline at a slower rate relative to players in cluster two (bottom left of the plot).

Chapter 5

Discussion

The first conclusion we can draw from Chapter 4 is that all three FDA models have relatively similar accuracy when it comes to making predictions. Each model estimates the same peak, age 26 years, for performance and the estimated mean functions have similar shapes (although as previously stated, the imFunPCA methods predicts a more drastic decline in performance after age 30 years relative to the two other methods). The eigenfunctions estimated by the three models are also quite similar other than the second eigenfunction estimated by SOAP. In general, we can say that the FDA models perform better predicting the performance of the defensemen position group relative to the forward position group.

The three methods are each useful when tasked with different problems. The SOAP method tends to excel when the sparsity of the data is large. In our application, sparsity would occur if a player misses seasons due to injury; or if a player starts his career later than age 22 years or ends his career earlier than age 34 years. The PACE method is the simplest model to use (since it is executable from a simple R package and can be implemented with relatively few lines of code). The imFunPCA method predicts drastic declines in performance, which is useful for worst case analysis. All three mthods do not require us to assume a given polynomial degree and allow us to fit quite different curves to different players; this is a testement to the flexibility of the FDA models. In terms of runtime, the imFunPCA method was the slowest to run, followed by SOAP; PACE was the fastest model to run.

The FDA models we have considered in this work are useful for front office staff to make decisions on acquiring players. When considering signing a veteran player in Free Agency, for example, it is useful to know how his performance might decline in future seasons. This might influence the length of the contract you offer him, or the role you expect him to play on your team (star player versus supporting cast, for example). For a player coming off of a major injury or outlier season, the imFunPCA method would help project the worst case forecast for the player. We saw in the results section that we can use the first s seasons of a player's career to predict his performance in future seasons. This is important for a

front office staff since a player drafted by a team is under team control in the form of an entry level contract or restricted free agency (RFA). A player may only declare himself to be an unrestricted free agent if he is over the age of 27 or has played in the league for a minimum of 7 years. Using the FDA models we have considered, a team can forecast the future performance of their young players to make decisions such as extensions for young players or matching RFA offer sheets from other teams.

When trading younger players for older players a front office staff might take into account the fact that the younger player is likely to perform better in future seasons, while the veteran (older) player is likely to see a decline in performance. In general, teams are likely to use aging curve models when making roster decisions. The front office staff must balance short term decisions (optimizing their roster in the current season) versus long term decisions (what players do they want to keep for future seasons).

The FDA models studied here are unique in the fact that we are able to fit separate aging curves for each player using Functional Principal Component Analysis. This means that not all veteran players are projected to see the same decline in performance in future seasons, and not all younger players are likely to see the same increase in performance over time. While projecting a player's future performance solely using age effects is not the most accurate method for prediction as opposed to other modelling techniques; these age effects could be combined with other features in a regression model. For example, one might use the previous three years of a player's performance along with age effects from the FPCA analysis in a projection system.

Future work to extend the modelling done in this research include adding more covariates to the model; for example dividing the training data in different ways (separating the forward group into different positions - right wing, center, left wing; or creating different aging curves for players binned by specific height or weight thresholds). We could also study the effect of using different target variables (for example minutes played, goals, points per game, etc.), and considering different sports and position groups (goalies, for example).

Bibliography

[1] Berry, S. M., Reese, C. S., & Larkey, P. D. (1999). *Bridging Different Eras in Sports.* Journal of the American Statistical Association, 94(447), 661–676. https://doi.org/10.2307/2669973

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

[3] Nie, Y., Yang, Y., Wang, L., & Cao, J. (2022). *Recovering the underlying trajectory from sparse and irregular longitudinal data.* Canadian Journal of Statistics, 50(1), 122-141.

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

[5] Chen, K., Zhang, X., Petersen, A., & Müller, H. G. (2017). *Quantifying infinite-dimensional data: Functional data analysis in action.* Statistics in Biosciences, 9(2), 582-604.

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

[7] Wakim, A., & Jin, J. (2014). *Functional data analysis of aging curves in sports.* arXiv preprint arXiv:1403.7548.

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

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

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

[11] Wang, J. L., Chiou, J. M., & Müller, H. G. (2016). *Functional data analysis.* Annual Review of Statistics and its application, 3, 257-295.

[12] Schuckers, M., Lopez, M., & Macdonald, B. (2021). *What does not get observed can be used to make age curves stronger: estimating player age curves using regression and imputation.* arXiv preprint arXiv:2110.14017.

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

[14] Albert, J. (2002). *Smoothing career trajectories of baseball hitters.* Manuscript, Bowling Green State University.

[15] Distefano, G., & Goodpaster, B. H. (2018). *Effects of exercise and aging on skeletal muscle.* Cold Spring Harbor perspectives in medicine, 8(3), a029785.

[16] Jolliffe, I. T., & Cadima, J. (2016). *Principal component analysis: a review and recent developments.* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[17] Wise, B. M., & Gallagher, N. B. (1998). *An introduction to linear algebra.* Critical reviews in analytical chemistry, 28(1), 1-20.

[18] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Vanderplas, J. (2011). *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 12, 2825-2830.

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

[20] Lichtman, M. (2009). *How do baseball players age?* Fangraphs. Retrieved November 3, 2022, from https://tht.fangraphs.com/how-do-baseball-players-age-part-2/

[21] EvolvingWild (2017). *A New Look at Aging Curves for NHL Skaters.* Hockey Graphs. Retrieved November 8, 2022, from https://hockey-graphs.com/2017/04/10/a-new-look-at-aging-curves-for-nhl-skaters-part-2/

[22] James B. & Henzler J. (2002). *Win shares (1st ed.)*. STATS Pub.

[23] Kubatko, J. (2010). *Calculating Point Shares.* Hockey-Reference.com - Hockey Statistics and History. Retrieved October 21, 2022, from https://www.hockey-reference.com/about/point_shares.html

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