Bayesian Analysis (2015) **10**, Number 2, pp. 379–410

Predictions Based on the Clustering of Heterogeneous Functions via Shape and Subject-Specific Covariates

Garritt L. Page[*] and Fernando A. Quintana[†]

Abstract. We consider a study of players employed by teams who are members of the National Basketball Association where units of observation are functional curves that are realizations of production measurements taken through the course of one's career. The observed functional output displays large amounts of between player heterogeneity in the sense that some individuals produce curves that are fairly smooth while others are (much) more erratic. We argue that this variability in curve shape is a feature that can be exploited to guide decision making, learn about processes under study and improve prediction. In this paper we develop a methodology that takes advantage of this feature when clustering functional curves. Individual curves are flexibly modeled using Bayesian penalized B-splines while a hierarchical structure allows the clustering to be guided by the smoothness of individual curves. In a sense, the hierarchical structure balances the desire to fit individual curves well while still producing meaningful clusters that are used to guide prediction. We seamlessly incorporate available covariate information to guide the clustering of curves non-parametrically through the use of a product partition model prior for a random partition of individuals. Clustering based on curve smoothness and subject-specific covariate information is particularly important in carrying out the two types of predictions that are of interest, those that complete a partially observed curve from an active player, and those that predict the entire career curve for a player yet to play in the National Basketball Association.

Keywords: Product partition models, Nonparametric Bayes, Penalized splines, Hierarchical models, Right censored data, NBA player production curves.

1 Introduction

In multi-subject studies where observations are considered to be functional realizations, it is common to observe large amounts of between-subject heterogeneity in the sense that some subjects produce curves that are quite smooth while others produce curves that are (much) more erratic. Although not often explicitly considered when modeling these types of data, the variability in curve shape can be an important feature that may help distinguish individuals and lead to better understanding of processes under study and/or better predictions. Often in these studies two types of predictions are desired: those that predict the functional output across the entire time domain for

[*] Departmento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile, page@mat.puc.cl

[†] Departmento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile, quintana@mat.puc.cl

DOI: 10.1214/14-BA919

a hypothetical new subject and those that complete a partially observed functional response for subjects currently participating in the study. Explicitly considering curve shape in modeling, in addition to considering relevant covariates, should improve both types of predictions. This is particularly true of the application motivating the present study. Decision makers of teams that belong to the National Basketball Association (NBA) are very interested in being able to group and predict future performance of basketball players that are employed or could possibly be employed by NBA teams.

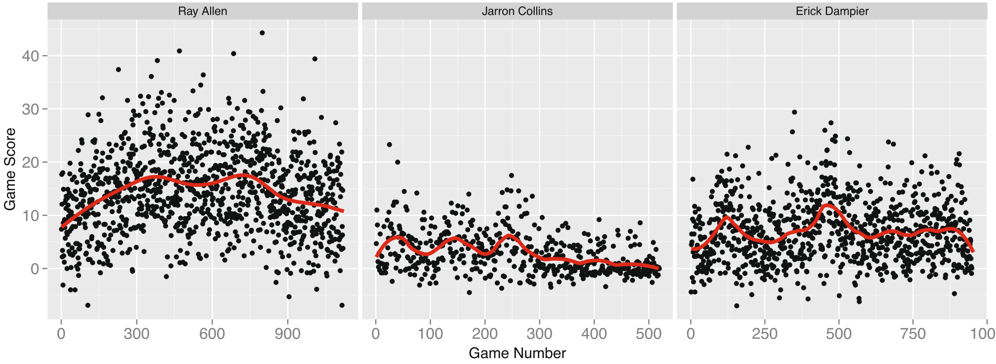

Figure 1: Career game-by-game Game Score results for three NBA players. A loess smoother with span equal to 0.3 is also provided.

The NBA is a North American professional mens basketball league that arguably employs the worlds most gifted basketball players. The league's popularity has steadily increased and as a result player salaries have exploded. Personnel decisions in the NBA (as in most professional sports leagues) are high-risk transactions. In the face of massive amounts uncertainty teams offer players guaranteed multi-year multi-million dollar contracts and as a result mistakes in player acquisition are extremely expensive. Making things even more treacherous are the abstruse rules governing player transactions found in the collective bargaining agreement (CBA). Among other things, the CBA regulates the amount of resources dedicated to player acquisition. Teams that misallocate player salary resources by over paying severely hinder a team's future flexibility and negatively impact a team's future competitiveness and profitability for years. Because of this, added value might be assigned to players who perform consistently compared to those that are more up and down.

Figure 1 displays scatter-plots and loess curves (with a span of 0.3) of game-by-game "production" for three NBA basketball players. Game-by-game "production" in Figure 1 is measured using the so called Game Score statistic (Hollinger 2002). More details are provided in Section 2 and the Appendix, but for now it suffices to know that higher values correspond with better performance and more production. Even though there is a large amount of game-to-game and player-to-player variability in Game Score it is still evident that production consistency between the three players varies. Erick Dampier appears to have two spikes of improved production, Jarron Collin's production oscillates during the first part of his career while Ray Allen's production is fairly smooth as it gradually increases and decreases with slight dip in the middle. Therefore curve shape should contain information that is valuable in distinguishing between different types of players and being able to assess their future value. Two types of predictions are used to assess future value. The first considers players who are currently members of the NBA and that will continue to participate in future games. Ray Allen in Figure 1 is an example of such a player. Although he has already played in more than 1000 games, he continues to play and predicting the performance for the remainder of his career is of considerable interest. This type of prediction will be referred to as "active player prediction". The second type of prediction considers basketball players who have yet to play in an NBA game but who have a skill set that will attract interest from NBA teams. For these players, predicting the entire career production curve is of interest. This type of prediction will be referred to as "career prediction".

It has become fairly common to consider the longitudinal curves of the type just described as discretized realizations of functional data. There is now a large literature dedicated to functional data analysis (FDA) techniques. A few popular methods that are actively being researched are functional principal components (Ramsay and Silverman 2005, chap. 6, Di et al. 2009), Gaussian process regression methods (Rasmussen and Williams 2006, Behseta et al. 2005, and Zhu and Dunson 2013) and multi-level functional basis expansion (Morris and Carroll 2006, DiMatteo et al. 2001, Biller 2000, Montagna et al. 2012, Bigelow and Dunson 2007). When considering multiple-subject studies the methods just described tend to separate individuals according to trend levels only, while ignoring the shape of the longitudinal projections. Though the idea of explicitly using shape or smoothness of curves to improve prediction is intuitively appealing there is surprisingly very little in the statistical literature dedicated to it. The one article that we are aware of is Zhu and Dunson (2012) whose focus is on estimating rate functions through a complicated system of differential equations and using covariates to explain variability in trajectories via stochastic volatility models. They applied their method to a longitudinal multi-subject blood pressure study for pregnant women and noted that blood pressure trajectories for normal women were more smooth relative to women with preeclampsia. We however, take an entirely different approach. Instead of dealing with a complicated system of differential equations we incorporate curve shape in modeling through a penalty parameter analogous to that found in penalized splines.

Our model involves an implied distribution on partitions of players. The allocation variables are treated as parameters and thus our approach may be seen as an extension of latent class analysis (LCA) (Collins and Lanza 2010) which classifies individual player curves into K pre-specificed clusters (see Dean and Raftery 2010 and Elliott et al. 2005). Unlike LCA, our methodology does not require a fixed pre-specified number of clusters as this is inferred from the corresponding posterior distribution on partitions. We briefly note that there does exist a small literature dedicated to estimating certain aspects of functional output such as dynamics (the speed of price increases and the rate at which this speed changes) that depend on covariate information (see Wang et al. 2008 and Zhu et al. 2011 and references therein). But these are not relevant to the current setting as they fail to deal with multiple-subject studies nor do they use curve shape in prediction and inference.

In sports, Berry et al. (1999) model career trajectories (or aging curves) non parametrically in order to make historical comparisons of player's abilities in baseball, hockey and golf. Connolly and Rendleman Jr. (2008) consider career paths of golf players to determine combinations of luck and skill required to win a golf tournament. Neither of these works were interested in grouping players to carry out career and active player predictions.

As noted, our principal goal is making active player and career predictions. If predictions are computed using methods that treat individual players independently, then both types of predictions would be extremely poor as they would not be data driven. One way of improving predictions is by borrowing strength (or sharing information) among players whose career production curves might be deemed similar. A straightforward way of borrowing strength is by introducing player clusters. However, if all individuals of the same cluster are restricted to have the same curve, then some individuals will invariably be poorly fit (too much borrowing of strength). Alternatively, if curves of all individuals of the same cluster are completely unrestricted, then clustering players would provide no predictive information (too little borrowing of strength). The methodology proposed in this article is able to balance very well the desire to produce good fitting individual curves while still producing clusters that allow enough borrowing of strength among similar players to guide prediction. This is carried out by employing a hierarchical model where subject-specific functions are modeled flexibly through a linear combination of basis functions whose coefficients are drawn from cluster-specific process level distributions. Doing this produces flexible subject-specific curves while still being able to produce reasonably accurate predictions by pooling together players with similar features/performances.

In our model, having covariate dependent clusters is crucial to carrying out career prediction as these are produced using the predictive distribution available from the covariate dependent clustering mechanism. Also, shape dependent clusters are useful to carrying out active player predictions as incomplete active player curves are filled in using curves of retired players that have similar career trajectories. There has been work regarding completing curves (Goldberg et al. 2014 work out Best Linear Unbiased Predictors (BLUPS) for past and future curve segments) and local borrowing of strength to fit global curves (Petrone et al. 2009 employ functional Dirichlet processes to group Gaussian process realizations), but the approaches developed and purposes are very much different from the present study.

The remainder of the article is organized as follows. Section 2 describes the data collected and employed in the analysis. Section 3 provides details regarding the development of the methodology highlighting model components associated with cluster-specific curve smoothness and active player prediction. Section 4 provides details regarding computation of posterior and predictive distributions. In Section 5 we provide details of a small simulation study. Results from the analysis of the NBA application are provided in Section 6. Finally, we provide some concluding remarks in Section 7.

2 Description of Data

We collected common game-by-game (including playoffs) modern basketball metrics for each player drafted into the NBA during the years 1992-2004 (Shaquille O'Neal to Dwight Howard) that participated in at least one game up through the 2009/2010 season. This resulted in 576 players with number of games played ranging from 2 to 1383 games. A few of the players appeared in very few games and are not representative of a typical NBA player. Because of this, and to reduce the noise introduced by the careers of players that contain little information regarding the processes of interest, we restrict our attention to players with at least three seasons of experience (the rookie contract length of the 2005 CBA). Also, to retain enough games to get a reasonable sense of a player's ability we only include players who played at least a half a season's worth of games (42). Finally, we excluded the following 8 players whose careers were cut short either by career ending injuries or untimely deaths: Bryant Reeves, Malik Sealy, Eddy Curry, Jason Collier, Eddie Griffin, Yao Ming, T. J. Ford, and Gilbert Arenas. This resulted in 408 players with number of games played through the 2009/2010 season ranging from 45 to 1383. Of the 408 players, 263 are classified as "retired" as they did not play beyond the 2009/2010 season.

Measuring game-by-game production is not straightforward as there are numerous, difficult to measure factors that influence player performance. Because of this no gold standard basketball production metric exists. That said, one that has become somewhat popular is John Hollinger's so called Game Score which is a linear combination of common variables that are recorded for each player through out the course of a game (e.g., number of baskets made and number of steals acquired. More details can be found in the Appendix and at Hollinger 2002). This metric will be used as our response variable and therefore a representation of a player's game productivity. Though Game Score has deficiencies (e.g., weighted heavily towards offensive output and doesn't account for quality of opponent), it provides a fairly accurate indicator of player production for any given game. The maximum Game Score collected is 63.5 (corresponding to Kobe Bryant's 81 point game). The minimum Game Score was -9.9 and the average Game Score among all players is 8.1. An alternative to raw Game Score is a standardized Game Score where standardization is carried out by dividing Game Score by the minutes played in each game, thus removing Games Score's dependence on minutes played. However, players whose production is not negatively impacted by increased minutes are more valuable than those who are less efficient with increased game time and distinguishing between these types of players is desirable. For this reason we opt to use raw Game Score values.

For an aging (or time) variable there are various units that could be used. For example, age, number of accumulated minutes played, or simply the number of games played are all reasonable. Since each of these measurements are only available to us on a game-by-game basis, the shape (or smoothness) of the curve remains unchanged regardless of age unit employed. Thus for sake of expositional clarity we use number of games played (see Figure 1 as an example).

Through exploratory analysis we identified three covariates that in addition to being of interest in their own right are informative in grouping players. These are age during first game played (measured in years), experience before being drafted into the NBA (High School basketball only, some Collegiate basketball, or International basketball), and draft order. Draft order is the order in which players are selected in their respective drafts. For example, a player with draft order 1 implies he was the first player selected in his respective draft, and draft order 2 implies he was the second player selected etc. In Section 3 we describe how draft order is used explicitly to predict total games played for active players, but as a covariate used to influence clustering we categorize a player's draft order as being a top five pick, a first round pick (excluding the first five) and a second round pick. (Since 1989 the NBA draft has consisted of two rounds.) Table 1 provides the number of players in each of the nine categories. Other baseline covariates were considered such as position, height, and other physiological characteristics but preliminary research indicated they were not useful in partitioning players with regards to production.

Table 1: Total number of players in each of the nine categories.

	Experience		
Draft	High School	College	International
Top 5	7	51	4
1st Round	15	200	25
2nd Round	1	98	8

3 Model Description and Development

We first consider the model's clustering mechanism highlighting its dependence on subject-specific covariates. Secondly, the likelihood structure incorporating the number of games played and career length (which are right censored for active players) is detailed. Lastly, we describe the hierarchical component which balances goodness of individual fit with ability to produce clusters that are able to guide prediction.

3.1 Product Partition Model with Covariates (PPMx)

Let $i = 1, \ldots, m$ index the m players in the study. Further, let $\rho = \{S_1, \ldots, S_{k_m}\}$ denote a partitioning (or clustering) of the m individuals into k_m subsets such that $i \in S_j$ implies that individual i belongs to cluster j. Alternatively, we will denote cluster membership using $s_1, \ldots, s_m$ where $s_i = j$ implies $i \in S_j$. Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})$ denote player i's covariate vector with x_{i1} corresponding to age, x_{i2} experience and x_{i3} draft order. Let $\boldsymbol{x}_j^\star = \{\boldsymbol{x}_i : i \in S_j\}$ be the partitioned covariate vector. Our approach is to first directly model ρ with the covariate dependent product partition model of Müller et al. (2011) (which will be referred to as the PPMx model) and then construct a hierarchical model given the partition (as opposed to introducing latent variables that indirectly induce a partitioning of individuals). The PPMx prior incorporates the idea that individuals with similar covariate values are more likely *a priori* to belong to the same cluster relative to individuals with dissimilar covariate values. Additionally, this prior is very simple, highly customizable, seamlessly incorporates different types of covariates (e.g., continuous or categorical), and is particularly well suited for prediction (something that is of interest here). An alternative method not considered here can be found in Park and Dunson (2010). The PPMx prior consists of a cohesion function, $c(S_j) \geq 0$ for $S_j \subset \{1, \ldots, n\}$, and a nonnegative similarity function $g(\boldsymbol{x}_j^{\star})$. The former measures the tightness of how likely elements of S_j are clustered *a priori* and the latter formalizes the similarity of the x_i's by producing larger values of $g(\boldsymbol{x}_j^{\star})$ for x_i's that are more similar. The form of the PPMx prior is simply the following product (for more details see Müller et al. 2011)

$$P(\rho|\boldsymbol{x}) \propto \prod_{j=1}^{k_m} c(S_j) g(\boldsymbol{x}_j^{\star}). \tag{1}$$

A simple example of a cohesion function that produces a Dirichlet Process type partitioning is $c(S_j) = M \times (|S_j| - 1)!$ for some positive M and $|\cdot|$ denoting cardinality. Regarding possible similarity functions, Müller et al. (2011) provide a bit of exposition for different types of covariates (e.g., continuous, ordinal, or categorical). Generically speaking they suggest the following structure

$$g(\boldsymbol{x}_j^{\star}) = \int \prod_{i \in S_j} q(x_i|\zeta_j) q(\zeta_j) d\zeta_j. \tag{2}$$

where ζ_j is a latent variable and $q(\cdot|\zeta_j)$ and $q(\cdot)$ are (typically) conjugate probability models. This structure is not necessarily used for its probabilistic properties (indeed $\boldsymbol{x}$ is not even random), but rather as a means to measure the similarity of the covariates in cluster S_j. In reality any function that produces larger values as the entries of $\boldsymbol{x}_j^{\star}$ become more similar can be considered as a similarity function. For example $g(\boldsymbol{x}_j^{\star}) = \exp\{-s_j^2\}$ where s_j^2 is the empirical variance of $\boldsymbol{x}_j^{\star}$ is a completely reasonable similarity function for continuous covariates.

It turns out that the similarity function (2) coupled with cohesion function $c(S_j) = M \times (|S_j| - 1)!$ produces the same marginal prior distribution on partitions as that induced by using a Dirichlet process (DP). For more details see Müller et al. (2011).

Given ρ we may proceed to specify a hierarchical model that flexibly models individual curves. Before doing so, we very briefly introduce a few pieces of notation that will be used. In what follows cluster-specific and subject-specific parameters will need to be distinguished. If we let $\boldsymbol{\theta}_i$ denote some generic subject-specific parameter vector, then $\boldsymbol{\theta}_j^*$ will be used to denote a cluster-specific parameter in the sense that $i \in S_j$ implies that $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j^*$. Alternatively, cluster labels $(s_1, \ldots, s_m)$ can be used to connect subject and cluster specific parameters through $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{s_i}^*$. Lastly, vectors of subject-specific and cluster-specific parameters are denoted by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ and $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*)$.

3.2 Likelihood

To distinguish between players that play beyond the 2009/2010 season we use the following indicator variable

$$g_i = \begin{cases} 0 & \text{if player } i \text{ retired before or at the conclusion of the 2009/2010 season} \\ 1 & \text{if player } i \text{ played beyond the 2009/2010 season.} \end{cases}$$

Let n_i denote the total number of games played in player i's career. If $g_i = 0$ then n_i is observed otherwise a lower bound denoted by $\tilde{n}_i$ is observed such that $n_i \geq \tilde{n}_i$. Thus, we are dealing with right censored type observations and we will incorporate ideas developed for modeling them. We denote the response vector for players whose $g_i = 0$ with $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})$ otherwise $\boldsymbol{y}_i = (y_{i1} \ldots, y_{i\tilde{n}_i})$. The career production curve for players whose $g_i = 1$ needs to be "completed" which requires the prediction or imputation of n_i.

Predicting n_i is not trivial (even given $\tilde{n}_i$) because it is highly variable and demonstrates a strong association with very few covariates. One covariate we found that displays a strong association with n_i is career length (denoted by L_i and measured in years). Unfortunately, this variable is also right censored and for $g_i = 1$ we only observe $\tilde{L}_i$. However, we consider L_i because it displayed a stronger association with the uncensored variable draft order (denoted by d_i) than that found between n_i and d_i. Therefore we employ d_i to first impute L_i and then use L_i to predict n_i. Thus, the likelihood for the ith player is composed of the random variables $(\boldsymbol{y}_i, g_i, n_{i:g_i=0}, \tilde{n}_{i:g_i=1}, L_{i:g_i=0}, \tilde{L}_{i:g_i=1})$ which we model jointly by way of

$$\begin{aligned} p(\boldsymbol{y}_i, g_i, n_{i:g_i=0}, \tilde{n}_{i:g_i=1}, L_{i:g_i=0}, \tilde{L}_{i:g_i=1}) &= [p(\boldsymbol{y}_i|n_i, L_i)p(n_i|L_i)p(L_i)]^{1-g_i} \\ &\quad \times [p(\boldsymbol{y}_i|\tilde{n}_i, \tilde{L}_i)p(\tilde{n}_i|\tilde{L}_i)p(\tilde{L}_i)]^{g_i}. \end{aligned}$$

We now detail each of the three likelihood components.

When only considering retired players we found that the association between d_i and L_i was somewhat nonlinear. (This is reasonable considering that our pool of players consists of only those who played at least one NBA game thus retaining only the "good" 2nd round picks.) Because of this, we assume $L_i \sim N(\nu_i, \psi^2)$ where $\nu_i = \gamma_0 + \gamma_1 d_i + \gamma_2 d_i^2$. However, the association between n_i and L_i was fairly linear so we assume $n_i | L_i \sim N(\eta_i, \delta^2)$ where $\eta_i = \alpha_0 + \alpha_1 L_i$. Thus, $(n_{i:g_i=0}, \tilde{n}_{i:g_i=1}, L_{i:g_i=0}, \tilde{L}_{i:g_i=1})$'s contribution to the likelihood is

$$\begin{aligned} [p(n_i|L_i)p(L_i)]^{1-g_i}[p(\tilde{n}_i|\tilde{L}_i)p(\tilde{L}_i)]^{g_i} &= \left[N(n_i; \eta_i, \delta^2) N(L_i; \nu_i, \psi^2)\right]^{1-g_i} \\ &\quad \times \left[\left\{1 - \Phi\left(\frac{\tilde{n}_i - \eta_i}{\delta}\right)\right\}\left\{1 - \Phi\left(\frac{\tilde{L}_i - \nu_i}{\psi}\right)\right\}\right]^{g_i} \end{aligned}$$

where $N(\cdot; m, s^2)$ denotes a Gaussian density function with mean m and variance s^2 and $\Phi(\cdot)$ denotes a standard normal cdf. As a result, imputing n_i for active players is carried out by first imputing L_i using a quadratic model with d_i.

We briefly note that although a Poisson model for n_i might seem natural, it is not appropriate in the current context as the simultaneous increasing of the mean and variance of the Poisson distribution seems to contradict what is empirically observed. Thus, for simplicity, we elected to employ a Gaussian to model n_i and round the predictions (something that is not uncommon, see page 458 of Gelman et al. 2013). Also, modeling

n_i non-parametrically could potentially improve prediction but we elected to employ the simpler parametric model as its predictions were satisfactory for our purposes. Nonetheless, predicting n_i is of considerable interest in its own right to NBA decision makers and could be an interesting future research project.

Finally, given n_i and letting $f_i(z_{it})$ denote the ith player's underlying production curve value for the tth game played (denoted by z_{it}), we model measurements y_{it} as

$$y_{it} = \beta_{0i} + f_i(z_{it}) + \epsilon_{it} \text{ for } t = 1, \ldots, n_i \ (\tilde{n}_i \text{ for } g_1 = 1) \tag{3}$$

where $\epsilon_{it} \sim N(0, \sigma_i^2)$ independently across i. It is possible that incorporating a more sophisticated error model (such as autoregressive errors) could prove to be beneficial, but for simplicity we maintain independence. A fairly popular method of characterizing $f_i(\cdot)$ is to define a collection of basis functions (e.g., wavelet, polynomial) and assume that y_{it} lies in their span. We adopt this method and employ a B-spline basis as it has a number of attractive computational properties and facilitates active player prediction as will be detailed shortly. Therefore, $f_i(\cdot)$ can be written as the following linear combination

$$f_i(z_{it}) = \sum_{\ell=1}^{P_i} \beta_{i\ell} h_\ell(z_{it}; \boldsymbol{\xi}_i) \tag{4}$$

where $h_\ell(z; \xi_i)$ denotes the ℓth B-spline basis function evaluated at knots contained in $\boldsymbol{\xi}_i$. If p_i denotes the number of inner knots and q the spline degree, then $P_i = p_i + q + 1$. Now define $\boldsymbol{H}_i$ as the $n_i \times P_i$ matrix with rows $\{h_1(z_{it}), \ldots, h_{P_i}(z_{it})\}$ for $t = 1, \ldots, n_i$ ($\tilde{n}_i$ for $g_i = 1$), and $\boldsymbol{\beta}_i = \{\beta_{i1}, \ldots, \beta_{iP_i}\}$. Combining (3) and (4) produces

$$\boldsymbol{y}_i = \beta_{0i}\mathbf{1}_i + \boldsymbol{H}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \text{ for } \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma_i^2 \boldsymbol{I}_{n_i}), \tag{5}$$

where $\mathbf{1}_i$ denotes a vector of ones and $\boldsymbol{I}_{n_i}$ an identity matrix.

The dimension of $\boldsymbol{H}_i$ depends on n_i (and $\tilde{n}_i$ for active players). This coupled with the fact that B-splines form a local basis in that each basis function is non-negative only on an interval formed by $q + 2$ adjacent knots can be exploited to carry out active player prediction. Since for any fixed z_{it} at most $q + 1$ basis functions are positive, the predicted value of n_i for active players will determine the number of zero columns in $\boldsymbol{H}_i$. Thus, the section of an active players curve corresponding to the z_{it} values between n_i and $\tilde{n}_i$ are completely informed by the cluster specific curve or in the case that the player belongs to a singleton, the grand mean curve (more details are in Section 3.3). Using $\tilde{\boldsymbol{H}}_i$ to denote the design matrix that incorporates the predicted value of n_i based on $\tilde{n}_i$, the full likelihood for $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\sigma}^2, \boldsymbol{\eta}, \boldsymbol{\nu}, \psi^2, \delta^2)$ is

$$\begin{aligned}
&\ell(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m, \boldsymbol{n}, \boldsymbol{L}, \tilde{\boldsymbol{n}}, \tilde{\boldsymbol{L}}, \boldsymbol{g} | \boldsymbol{\Theta}) \\
&= \prod_{i=1}^{n} \left[N_{n_i}(\boldsymbol{y}_i; \beta_{0i}\mathbf{1}_i + \boldsymbol{H}_i\boldsymbol{\beta}_i, \sigma_i^2 \boldsymbol{I}_{n_i}) N(n_i; \eta_i, \delta^2) N(L_i; \nu_i, \delta^2) \right]^{1-g_i} \times \\
&\times \left[N_{\tilde{n}_i}(\boldsymbol{y}_i; \beta_{0i}\mathbf{1}_i + \tilde{\boldsymbol{H}}_i\boldsymbol{\beta}_i, \sigma_i^2 \boldsymbol{I}_{\tilde{n}_i}) \left\{ 1 - \Phi\left(\frac{\tilde{n}_i - \eta_i}{\delta} \right) \right\} \left\{ 1 - \Phi\left(\frac{\tilde{L}_i - \nu_i}{\psi} \right) \right\} \right]^{g_i}.
\end{aligned} \tag{6}$$

3.3 Hierarchical Model

The number and location of the inner-knots that make up $\boldsymbol{\xi}_i$ are rarely known. Their selection is crucial to producing an attractive curve without over-fitting. So called free-knot splines is a very flexible method that treats $\boldsymbol{\xi}_i$ as an unknown and has proved to be quite parsimonious in knot selection. (DiMatteo et al. 2001 and Denison et al. 2002 provide a nice overview.) Therefore, a possible direction to incorporating shape variability in prediction as desired would be to base clustering on the number and location of knots. However, to fit a free-knot spline some type of transdimensional Markov Chain Monte Carlo (MCMC) algorithm is often employed and this coupled with the PPMx prior for ρ would result in a doubly transdimensional MCMC algorithm that would become prohibitively expensive. To avoid these computational issues and to make the methodology more readily accessible, for each subject we instead select a moderate number of equally spaced knots within the knot domain and employ the Bayesian P-spline technology of Lang and Brezger (2004). Now shape variability can influence clustering through the penalty parameter of the P-splines. However, to retain flexible subject-specific fits, we use P-splines as a prior distribution of process level parameters and allow subject-specific coefficients to vary around a cluster-specific mean. That is, we assume the following process level structure for the $\boldsymbol{\beta}$'s:

$$\boldsymbol{\beta}_i | \boldsymbol{\theta}^*_{s_i}, \lambda^{2*}_{s_i} \sim N(\boldsymbol{\theta}^*_{s_i}, \lambda^{2*}_{s_i}\boldsymbol{I}) \text{ with } \sqrt{\lambda^{2*}_j} \sim UN(0, A), \tag{7}$$

and use a Bayesian P-spline prior for the $\boldsymbol{\theta}^*_j$'s (with $UN(\cdot,\cdot)$ denoting a Uniform distribution). A particularly nice feature of the methodology is the explicit ability to control the similarity between individual curves and their group counterparts through the hyper-parameter A.

In order to highlight two departures from the Bayesian P-splines of Lang and Brezger (2004) required by the present modeling we very briefly introduce them here. For more details see Lang and Brezger (2004) and Fahrmeir and Kneib (2005). Bayesian P-splines are the Bayesian analogue to splines penalized by d-order differences and are constructed around d-order Gaussian random walks. For example, for $d = 1$

$$\theta^*_{j\ell} = \theta^*_{j,\ell-1} + u_{j\ell} \quad \ell = 2, \dots, n \tag{8}$$

with $u_{j\ell} \sim N(0, \tau^{2*}_j)$. Typically $p(\theta^*_{j1}) \propto v$, but an improper prior is not an appropriate probability model for the Polya urn representation used in the PPMx. Thus, similar to what was done in Telesca and Inoue (2008) we assume $\theta^*_{j1} \sim N(0, \tau^{2*}_j/v^2)$ (with analogous extensions for $d > 1$). The value v can be assigned a prior distribution or be set to a fixed value. Equation (8) together with the $\theta^*_{j1} \sim N(0, \tau^{2*}_j/v^2)$ produce $\boldsymbol{\theta}^*_j \sim N(\boldsymbol{0}, \tau^{2*}_j \boldsymbol{K}^{-1})$ where $\boldsymbol{K}$ is a banded penalty matrix with v incorporated. τ^{2*}_j is the smoothing parameter associated with Bayesian P-splines and is crucial in being able to distinguish between individuals based on the smoothness of their respective curves. As suggested by Lang and Brezger (2004) we adopt $\tau^{2*}_j \sim IG(a_\tau, b_\tau)$ where $IG(\cdot,\cdot)$ denotes an inverse Gamma distribution and a_τ and b_τ are user supplied.

Recall that active player prediction is carried out by borrowing strength among players in a cluster. If player i belongs to a singleton or all members of his cluster are active players, then at least part of his prediction is completely guided by the prior on $\boldsymbol{\theta}^*_j$. Since the prior is centered at $\mathbf{0}$ this would produce poor active player predictions. To improve prediction in these situations, we introduce $\boldsymbol{\mu}$ as a vector of global curve coefficients such that $\boldsymbol{\theta}^*_j \sim N(\boldsymbol{\mu}, \tau_j^{2*}\boldsymbol{K}^{-1})$ with $\boldsymbol{\mu} \sim N(\mathbf{0}, s^2_\mu \boldsymbol{I})$. Including $\boldsymbol{\mu}$ potentially influences the values of $\boldsymbol{\theta}^*_j$ in that smaller magnitudes achieve the same amount of smoothing as when $\boldsymbol{\mu} = \mathbf{0}$. This should be taken into account when selecting values for a_τ and b_τ. Also, apart from improving prediction, $\boldsymbol{\mu}$ is of interest in its own right as it provides information regarding an average career curve among all players.

We end the description of our Bayesian P-spline approach with details regarding knot selection. A complicating factor of knot selection in modeling these data is the massive misalignment associated with the number of games played for each of the players. Making things worse is the inherent discontinuities in games played through out the course of one's career (e.g., the offseason, injuries, etc.) that we are not considering. There is a "curve registration" literature dedicated to better aligning functional domains in multi-subject studies (Telesca and Inoue 2008). However, we align career paths by matching the percentile number of career games played. This is carried out by transforming "time" to the unit interval which greatly simplifies the process of selecting $\boldsymbol{\xi}_i$. Therefore $z^*_{it} = z_{it}/n_i$ is used instead of z_{it}. (For retired players n_i is the observed number of games played and the predicted for active players.) Thus for retired players $z^*_{in_i} = 1$ while for active players $z^*_{in_i} < 1$. Now $\boldsymbol{\xi}_i$ can be a knot set that partitions the unit interval into equal subintervals and since it does not depend on n_i it can be the same for all players. We do note that aligning career paths in this way is imperfect as the 95th percentile of games played for one player might be during his third season while for another player during his fifteenth season. Even so, we believe that matching curves by way of percentile of games played produces coherent comparisons and valid borrowing of strength. With an enriched data set we could attempt to take into account possible discontinuities in career paths. (This would actually be very interesting as many players improve during the off season.) It would be fairly straightforward to expand the model in a variety of ways, but the base model proposed would continue being the work horse even as other more idiosyncratic aspects of the data are considered.

With regards to modeling β_{0i} and σ^2_i there are any number of ways one might proceed. It seems plausible that σ^2_i might depend on z_{it}. That said, for sake of simplicity we utilize the common prior structure for variance components $\sigma^2_i \sim IG(a_\sigma, b_\sigma)$ with a_σ and b_σ being user supplied. For the subject-specific random intercepts, we use a Gaussian-inverse-Gamma hierarchy such that $\beta_{0i} \sim N(\mu_{b_0}, \sigma^2_{b_0})$ with $\mu_{b_0} \sim N(0, s^2_{b_0})$ and $\sigma^2_{b_0} \sim IG(a_{b_0}, b_{b_0})$. Finally, typical conjugate priors are used for $\boldsymbol{\alpha} = (\alpha_0, \alpha_1) \sim N(\boldsymbol{m}_a, s^2_a \boldsymbol{I})$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2) \sim N(\boldsymbol{m}_\gamma, s^2_\gamma \boldsymbol{I})$, $\delta^2 \sim IG(a_\delta, b_\delta)$, and $\psi^2 \sim IG(a_\psi, b_\psi)$.

Equation (9) is provided to aid in visualizing how all the moving parts of the hierarchical model are connected. Through out the remainder of the paper we will refer to the entire hierarchical model as HPPMx.

$$
\begin{aligned}
\boldsymbol{y}_i, n_i, L_i, \tilde{n}_i, \tilde{L}_i | g_i, \boldsymbol{\beta}_i, \rho, \sigma_i^2, \beta_{0i}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \delta^2, \psi^2 & \\
\sim & \left[N_{n_i}(\boldsymbol{y}_i; \beta_{0i}\boldsymbol{1}_i + \boldsymbol{H}_i\boldsymbol{\beta}_i, \sigma_i^2\boldsymbol{I}_{n_i}) N(n_i; \eta_i, \delta^2) N(L_i; \nu_i, \psi^2)\right]^{1-g_i} \\
\times & \Bigg[N_{\tilde{n}_i}(\boldsymbol{y}_i; \beta_{0i}\boldsymbol{1}_i + \tilde{\boldsymbol{H}}_i\boldsymbol{\beta}_i, \sigma_i^2\boldsymbol{I}_{\tilde{n}_i}) \\
\times & \left\{1 - \Phi\left(\frac{\tilde{n}_i - \eta_i}{\delta}\right)\right\} \left\{1 - \Phi\left(\frac{\tilde{L}_i - \nu_i}{\psi}\right)\right\} \Bigg]^{g_i} \\
\sigma_i^2 | a_\sigma, b_\sigma \sim & \; IG(a_\sigma, b_\sigma) \\
\beta_{0i} | \mu_{b_0}, \sigma_{b_0}^2 \sim & \; N(\mu_{b_0}, \sigma_{b_0}^2) \;\text{ with }\; \mu_{b_0} \sim N(0, s_{b_0}^2) \;\text{ and }\; \sigma_{b_0}^2 \sim IG(a_{b_0}, b_{b_0}) \\
\boldsymbol{\alpha} \sim & \; N(\boldsymbol{m}_a, s_a^2\boldsymbol{I}) \;\text{ and }\; \delta^2 \sim IG(a_\delta, b_\delta) \\
\boldsymbol{\gamma} \sim & \; N(\boldsymbol{m}_\gamma, s_\gamma^2\boldsymbol{I}) \;\text{ and }\; \psi^2 \sim IG(a_\psi, b_\psi) \qquad (9) \\
\boldsymbol{\beta}_i | \rho, \theta_{s_i}^*, \lambda_{s_i}^* \sim & \; N(\boldsymbol{\beta}_i; \boldsymbol{\theta}_{s_i}^*, \lambda_{s_i}^{2*}) \;\text{ with }\; \sqrt{\lambda_h^{2*}} \sim UN(0, A) \\
\boldsymbol{\theta}_h^* | \rho, \boldsymbol{\mu}, \tau_h^{2*}, \boldsymbol{K} \sim & \; N(\boldsymbol{\mu}, \tau_h^{2*}\boldsymbol{K}^{-1}) \;\text{ with }\; \tau_h^{2*} \sim IG(a_\tau, b_\tau) \\
\boldsymbol{\mu} \sim & \; N(\boldsymbol{0}, s_\mu^2\boldsymbol{I}) \\
Pr(\rho) \propto & \prod_{h=1}^{k_m} c(S_j) g(\boldsymbol{x}_j^\star),
\end{aligned}
$$

for $i = 1, \ldots, m$ and $h = 1, \ldots, k_m$.

Before proceeding we make a brief comment regarding some specific model components. Since the Bayesian P-splines are used at the prior level rather than the process level of the hierarchical model, individual curves are not directly influenced by its smoothing penalization. The wiggliness of individual curves is a function of both τ_j^{2*} and A. As A increases the influence that τ_j^{2*} has on individual curves decreases. This is investigated further in the simulation study of Section 5. Therefore, if smooth individual curves are desired together with large within group variability it may be necessary to use 10-15 knots instead of the 20-30 knots recommended by Lang and Brezger (2004).

4 Posterior Computation

4.1 MCMC Implementation

We fit the proposed model to data by simulating a Markov chain whose equilibrium distribution is the desired posterior distribution. The algorithm employed is similar to Neal (2000)'s algorithm number 8 in that it can be divided into two basic pieces. The first updates the partition ρ via the Polya urn scheme of Blackwell and MacQueen (1973) (and further developed by Quintana 2006) and the other updates the hierarchical model parameters using a Gibbs sampler (Geman and Geman 1984 and Gelfand and Smith 1990) and Metropolis steps (Metropolis et al. 1953).

To update the cluster membership for subject i, cluster weights are created by comparing the unnormalized posterior for the hth cluster when subject i is excluded to that when subject i is included. In addition to weights for existing clusters, algorithm 8 of Neal (2000) requires calculating weights for p empty clusters whose cluster specific parameters are auxiliary variables generated from the prior. To make this more concrete, let S_h^{-i} denote the hth cluster and k_m^{-i} the number of clusters when subject i is not considered. Similarly $\boldsymbol{x}_h^{\star-i}$ will denote the vector of covariates corresponding to cluster h when subject i has been removed. Then the multinomial weights associated with the k_m^{-i} existing clusters and the p empty clusters are

$$Pr(s_i = h|-) \propto \begin{cases} N(\boldsymbol{\beta}_i; \theta_h^*, \lambda_h^{2*}\boldsymbol{I})\frac{c(S_h^{-i}\cup\{i\})g(\boldsymbol{x}_h^{\star-i}\cup\{\boldsymbol{x}_i\})}{c(S_h^{-i})g(\boldsymbol{x}_h^{\star-i})} \text{ for } h = 1,\dots,k_m^{-i} \\ N(\boldsymbol{\beta}_i; \boldsymbol{\theta}_{new,h}, \lambda^2_{new,h}\boldsymbol{I})c(\{i\})g(\{\boldsymbol{x}_i\})p^{-1} \text{ for } h = k_m^{-i}+1,\dots,k_m^{-i}+p. \end{cases}$$

Values for $\boldsymbol{\theta}_{new,h}$, $\lambda^2_{new,h}$ are auxiliary variables drawn from their respective priors as required by algorithm 8. Care must be taken when subject i belongs to a singleton cluster as removing the ith subject produces an empty cluster. This in turn requires relabeling the existing cluster specific components to avoid gaps in the cluster labeling.

The full conditional distributions of $(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma_i^2, \theta_j^*, \tau_j^{2*})$ are fairly common derivations and are provided in the Appendix. To update $(\lambda_j^*, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \delta^2, \psi^2)$ we employed a random walk Metropolis step with a Gaussian proposal distribution. A Markov chain can be constructed by employing a Gibbs sampler that first updates ρ and then on an individual basis updates model parameters by cycling through each full conditional and using a Metropolis step for the non conjugate parameters.

4.2 Posterior Prediction Distributions

A particularly nice feature of using PPMx is the availability of career prediction through covariate dependent predictive distributions. Using PPMx, posterior predictive distributions are readily available and can be obtained online in the sense that draws from this distribution can be collected within the MCMC algorithm. The posterior predictive distributions depend on covariate values through the allocation of a new individual to one of the k_m existing clusters or to a new cluster using the following multinomial weights

$$Pr(s_{n+1} = h|-) \propto \begin{cases} \frac{c(S_h\cup\{n+1\})g(\boldsymbol{x}_h^\star\cup\{\boldsymbol{x}_{n+1}\})}{c(S_h)g(\boldsymbol{x}_h^\star)} \text{ for } h = 1,\dots,k_m \\ c(\{n+1\})g(\{\boldsymbol{x}_{n+1}\}) \text{ for } h = k_m+1. \end{cases} \quad (10)$$

Once the future player has been allocated to a cluster, one carries out the typical Monte Carlo integration to sample from the posterior predictive.

4.3 Predicting n_i

Predicted values of (n_i, L_i) for retired players are produced at each MCMC iteration. We essentially employ the multiple imputation ideas of Little and Rubin (1987) but with the exception that we are very much interested in the values being imputed. Predictions are fairly straight forward as they only depend on the full conditionals of n_i and L_i which turn out to be truncated normal with $\tilde{n}_i$ and $\tilde{L}_i$ acting as lower bounds (the full conditionals are provided in the Appendix).

5 Simulation Studies

We conduct two small simulation studies to investigate the behavior of the HPPMx model (9). Recall that the principal motivation in incorporating a hierarchy is to balance goodness of individual curve fits with the production of meaningful clusters which facilitate prediction. Therefore apart from showing improved prediction performance, the simulation study explores just how much goodness of individual fit is sacrificed in the name of prediction (which is very little as will be seen). The first simulation study demonstrates the method's superior predictive performance by comparing out of sample mean integrated prediction rates to that of two competitors (which are detailed shortly). The second explores how certain model components influence subject-specific fits, curve smoothness, and clustering. The two competitors selected represent the extremes HPPMx attempts to balance, namely fitting each player independently versus assigning players cluster-specific curves. The first competitor is a semi-parametric regression model (henceforth SP) that fits individual curves independently and flexibly. The second is a semi-parametric regression model with a Dirichlet process prior (henceforth SPDP) which produces individual curves that are cluster specific. More precisely we consider

$$y_{it} = \boldsymbol{x}_i'\boldsymbol{\beta} + f_i(z_{it}) + \epsilon_{it} \text{ with } \epsilon_{it} \sim N(0, \sigma_i^2) \text{ for } i = 1, \ldots, m \text{ and } t = 1, \ldots, n \qquad (11)$$

where $(f_i(z_{i1}), \ldots, f_i(z_{in}))' = \boldsymbol{H}\boldsymbol{\theta}_i$ is modeled using subject-specific linear combinations of B-spline basis functions, $z_{it} \in [0, 1]$, $\boldsymbol{x}_i$ is a vector of covariates that will be described shortly and

$$\begin{array}{ll} \underline{\text{SP}} & \underline{\text{SPDP}} \\ \boldsymbol{\theta}_i \sim N(\mathbf{0}, \tau^2\boldsymbol{K}^{-1}) & \boldsymbol{\theta}_i | G \sim G \\ & G \sim DP(M, G_0) \text{ with } G_0 = N(\mathbf{0}, \tau^2\boldsymbol{K}^{-1}). \end{array}$$

Notice that under SP a P-spline prior is used for $\boldsymbol{\theta}_i$ while under SPDP the base measure of the DP is a P-spline prior. As with HPPMx, $\tau^2 \sim IG(a_\tau, b_\tau)$. The competitors selected, though reasonable, aren't capable of providing active player predictions. Therefore, prediction assessment in the simulation study is only carried out for career prediction. Finally, we investigate the influence that covariates have on clustering by considering the HPPMx model with a PPM prior rather than a PPMx prior (which will be hence forth referred to as HPPM).

Since both simulation studies employ the same general data generating mechanism we provide details here. A response vector is generated using

$$y_{it} = b_{0i} + f_{group_i}(z_{it}) + \epsilon_{it} \text{ with } \epsilon_{it} \sim N(0, w^2) \qquad (12)$$

where $f_{group_i}(\cdot)$ corresponds to $group_i = 1, \ldots, 6$ possible mean curves which were created using the NBA data as a guide (see Figure 2). The six mean curves are made to depend on covariates by creating two categorical covariates that when crossed produce six categories, one for each mean curve. A continuous covariate was generated by $x_i^* \sim N(group_i, 0.1)$. Since x_i^* depends on the two categorical covariates, an interaction

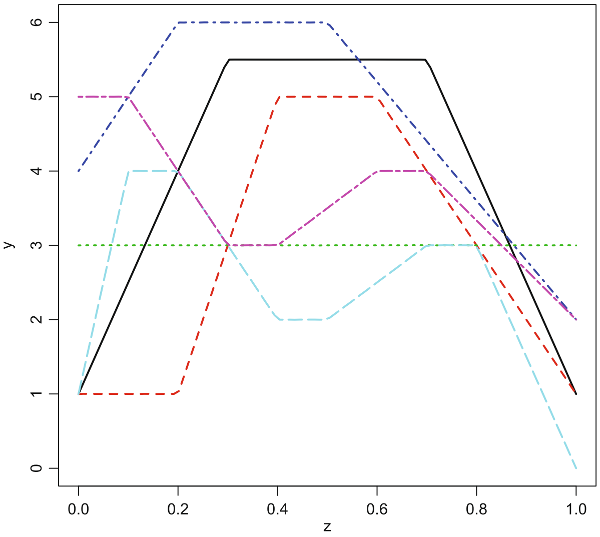

Figure 2: The six mean curves used in the simulation study.

between them is created. The three covariates were included in all model fits. Lastly, the random intercept is generated using $b_{0i} \sim N(10, 2)$.

The factors we explore in the simulation study with their respective levels are

- value for hyper parameter A (0.1, 1, 10)
- number of knots (5,15,30)
- variance of (12) ($w^2 = 0.1$, $w^2 = 1$)
- number of observations per subject ($n = 50$, $n = 100$).

A and the number of knots were selected to investigate how P-splines function as a prior at the process level instead of at the observation level of a hierarchical model. With n and w^2 we see how the methodology performs as more information becomes available relative to noise. For each combination of the factor levels 100 data sets with $m = 60$ (10 players per group) are generated and for each data set SP, SPDP, HPPM, and HPPMx are fit.

For all four procedures we set $a_\tau = 1$, $b_\tau = 0.05$, $a_\sigma = b_\sigma = 1.0$ and $v = 1$. For PPMx and PPM $s^2_{b_0} = s^2_\mu = 100^2$, and for SP and SPDP $\boldsymbol{\beta} \sim N(\mathbf{0}, 100^2\boldsymbol{I})$. Finally for HPPMx, the cohesion and similarity functions employed are those that match the marginal prior on partitions implied by a DP prior (see Section 6.1 for more details). Each of the four procedures were fit to each synthetic data set using 1000 MCMC iterates after discarding the first 5000 as burn-in. Empirically based starting values were employed which accelerated convergence making the 5000 iterate burn-in sufficient.

5.1 Out of Sample (Career) Prediction

To assess out of sample (career) prediction, for each of the 100 generated data sets 100 additional out of sample subjects were generated. The $f_{group}(\cdot)$ associated with each new subject is known and therefore out of sample prediction can be assessed by comparing $\hat{f}(\cdot)$ from the four procedures to $f_{group}(\cdot)$. After centering both $f_{group_j}(\cdot)$ and $\hat{f}_j(\cdot)$ (i.e. subtract off the empirical mean) for the jth out of sample subject, we measure prediction accuracy using the mean integrated squared prediction error

$$MISPE_j = E\int[\hat{f}_j(z) - f_{group_j}(z)]^2dz \approx \sum_t \Delta_t E[\hat{f}(z_{jt}) - f_{group}(z_{jt})]^2 \tag{13}$$

where $\Delta_t = (z_{jt+1} - z_{jt})$. Equation (13) essentially measures the average squared area between $f_{group_j}(\cdot)$ and $\hat{f}_j(\cdot)$ for the jth out of sample player over z's domain. The values in Table 2 correspond to

$$\frac{1}{D}\sum_{d=1}^{D}\frac{1}{100}\sum_{j=1}^{100} MISPE_{dj} \tag{14}$$

where d indexes the $D = 100$ generated data sets.

From Table 2 we see that HPPM and SPDP provide similar predictions which is to be expected as both employ a DP prior (although not at the same level of a hierarchy). What should be very obvious is that HPPMx does a much better job in out of sample prediction relative to the other three procedures for all data generating scenarios.

5.2 Goodness of Individual Fits, Curve Smoothness, and Clustering

To assess goodness of individual fits we employ the following R^2 type goodness-of-fit statistic from Ramsay and Silverman (2005):

$$R_i^2 = 1 - \frac{\sum_t(\hat{f}_i(z_{it}) - y_{it})^2}{\sum_t(y_{it} - \bar{y}_i)^2}. \tag{15}$$

R_i^2 can be loosely interpreted as a coefficient of determination in that as R_i^2 approaches 1, individual fits improve. Negative values of R_i^2 indicate that $\bar{y}_i$ predicts better than $\hat{f}_i(\cdot)$. The values in Table 3 correspond to

$$\frac{1}{D}\sum_{d=1}^{D}\frac{1}{m}\sum_{i=1}^{m} R_{di}^2. \tag{16}$$

From Table 3 we see that SP tends to produce the best individual fits and SPDP the worst. This is of course expected as all individuals are fit independently by SP while SPDP provides cluster specific curves. However, HPPMx does remarkably well in producing good individual fits as HPPMx is very close to SP particularly as n increases. Thus, HPPMx's meaningful cluster production doesn't require sacrificing much goodness of individual fit.

Table 2: Results from the simulation study investigating out of sample prediction. Table entries are calculated using (14) with $m = 60$ players.

			$n = 50$			$n = 100$		
			Number of knots			Number of knots		
		Model	5	15	30	5	15	30
$w^2 = 0.1$	$A = 0.1$	HPPMx	0.131	0.119	0.123	0.149	0.144	0.122
		HPPM	0.891	0.868	0.847	0.873	0.855	0.848
		SP	1.310	1.301	1.319	1.297	1.288	1.277
		SPDP	0.782	0.784	0.789	0.780	0.779	0.775
	$A = 1$	HPPMx	0.034	0.026	0.095	0.026	0.025	0.029
		HPPM	0.791	0.795	0.799	0.779	0.784	0.777
		SP	1.324	1.319	1.306	1.285	1.300	1.279
		SPDP	0.785	0.789	0.782	0.779	0.782	0.776
	$A = 10$	HPPMx	0.022	0.025	0.108	0.023	0.041	0.037
		HPPM	0.786	0.791	0.800	0.774	0.776	0.784
		SP	1.321	1.307	1.292	1.282	1.278	1.289
		SPDP	0.785	0.783	0.779	0.775	0.777	0.780
$w^2 = 1$	$A = 0.1$	HPMMx	0.155	0.203	0.254	0.142	0.151	0.242
		HPMM	0.882	0.848	0.852	0.854	0.828	0.837
		SP	1.315	1.312	1.326	1.277	1.287	1.283
		SPDP	0.783	0.786	0.787	0.771	0.776	0.777
	$A = 1$	HPPMx	0.100	0.102	0.170	0.045	0.069	0.102
		HPPM	0.796	0.807	0.829	0.778	0.789	0.795
		SP	1.312	1.316	1.306	1.257	1.279	1.288
		SPDP	0.783	0.785	0.785	0.766	0.776	0.776
	$A = 10$	HPPMx	0.080	0.095	0.157	0.047	0.066	0.113
		HPPM	0.823	0.817	0.836	0.798	0.794	0.809
		SP	1.331	1.305	1.327	1.291	1.269	1.284
		SPDP	0.789	0.780	0.787	0.778	0.772	0.776

To assess curve smoothness we calculate the standard deviation of the lag one differences from the estimated curve

$$\ell SD_i = \sqrt{\frac{1}{n-3}\sum_{t=1}^{n-1}(lag_{it} - \overline{lag}_i)^2}, \tag{17}$$

where $lag_{it} = \hat{f}_i(z_{it+1}) - \hat{f}_i(z_{it})$ for $t = 1, \ldots, n-1$ and $\overline{lag}_i = 1/(n-1)\sum_{t=1}^{n-1} lag_{it}$. Large values of ℓSD_i generally indicate more wiggliness relative to small values. Values provided in Table 4 correspond to

$$\frac{1}{D}\sum_{d=1}^{D}\frac{1}{m}\sum_{i=1}^{m}\ell SD_{di}. \tag{18}$$

From Table 4 it appears that curves become less wiggly as n increases relative to

Table 3: Results from the simulation study investigating goodness-of-fit. Table entries are calculated using (16) with $m = 60$ players.

			$n = 50$			$n = 100$		
			Number of knots			Number of knots		
		Model	5	15	30	5	15	30
$w^2 = 0.1$	$A = 0.1$	HPPMx	0.979	0.990	0.950	0.981	0.989	0.984
		HPPM	0.813	0.908	0.919	0.812	0.939	0.952
		SP	0.985	0.993	0.994	0.984	0.992	0.992
		SPDP	0.802	0.794	0.774	0.764	0.759	0.743
	$A = 1$	HPPMx	0.981	0.965	0.864	0.984	0.991	0.979
		HPPM	0.983	0.987	0.950	0.984	0.991	0.989
		SP	0.985	0.993	0.994	0.984	0.992	0.992
		SPDP	0.808	0.811	0.798	0.771	0.743	0.746
	$A = 10$	HPPMx	0.984	0.975	0.853	0.984	0.990	0.979
		HPPM	0.985	0.987	0.937	0.984	0.991	0.992
		SP	0.985	0.993	0.994	0.984	0.992	0.992
		SPDP	0.802	0.793	0.787	0.765	0.747	0.750
$w^2 = 1$	$A = 0.1$	HPPMx	0.583	0.593	0.572	0.574	0.593	0.575
		HPPM	0.485	0.537	0.497	0.476	0.552	0.541
		SP	0.613	0.663	0.713	0.570	0.575	0.533
		SPDP	0.538	0.554	0.565	0.515	0.521	0.520
	$A = 1$	HPPMx	0.585	0.608	0.630	0.572	0.589	0.605
		HPPM	0.589	0.618	0.663	0.574	0.596	0.616
		SP	0.613	0.662	0.711	0.566	0.575	0.531
		SPDP	0.541	0.555	0.562	0.511	0.522	0.523
	$A = 10$	HPPMx	0.587	0.610	0.626	0.574	0.589	0.607
		HPPM	0.593	0.624	0.667	0.577	0.597	0.617
		SP	0.613	0.662	0.712	0.568	0.574	0.535
		SPDP	0.539	0.554	0.561	0.514	0.519	0.522

the number of knots. This is expected. Also unsurprising is that HPPMx and HPPM produce similar values of (18) with the biggest differences occurring when w^2 (within player variability) and A (within cluster variability) are small. What is a bit surprising is that the value of A doesn't much alter curve smoothness for HPPMx. It appears that w^2 is more influential. Overall, since the values of (18) for HPPMx are fairly similar to those for SP and SPDP (recall that SP and SPDP are not influenced by A), penalizing curves directly with a P-spline prior produces curves with similar smoothness as those produced through the hierarchical model.

To see how the PPMx prior improves clustering relative to the PPM prior, Table 5 provides the number of estimated clusters (k_m) averaged over all $D = 100$ data sets. For each data set ρ was estimated using Dahl (2006)'s method which is based on least-squares distance from the matrix of posterior pairwise co-clustering probabilities (note that an estimate of ρ also provides an estimate of k_m).

Table 4: Results from the simulation study investigating smoothness. Table entries are calculated using (18) with $m = 60$ players.

			$n = 50$			$n = 100$		
			Number of knots			Number of knots		
		Model	5	15	30	5	15	30
$w^2 = 0.1$	$A = 0.1$	HPPMx	0.162	0.165	0.171	0.082	0.082	0.087
		HPPM	0.143	0.154	0.165	0.073	0.078	0.084
		SP	0.162	0.162	0.166	0.081	0.080	0.083
		SPDP	0.154	0.153	0.157	0.077	0.076	0.079
	$A = 1$	HPPMx	0.163	0.165	0.170	0.081	0.081	0.084
		HPPM	0.163	0.164	0.168	0.081	0.081	0.085
		SP	0.162	0.162	0.166	0.081	0.080	0.083
		SPDP	0.154	0.154	0.158	0.077	0.076	0.079
	$A = 10$	HPPMx	0.163	0.165	0.169	0.082	0.081	0.085
		HPPM	0.162	0.164	0.168	0.082	0.081	0.086
		SP	0.161	0.162	0.166	0.081	0.080	0.083
		SPDP	0.153	0.154	0.158	0.077	0.076	0.079
$w^2 = 1$	$A = 0.1$	HPPMx	0.179	0.208	0.274	0.096	0.107	0.132
		HPPM	0.155	0.182	0.250	0.082	0.096	0.120
		SP	0.195	0.206	0.258	0.088	0.071	0.062
		SPDP	0.172	0.174	0.191	0.095	0.090	0.095
	$A = 1$	HPPMx	0.184	0.210	0.288	0.097	0.109	0.136
		HPPM	0.183	0.209	0.298	0.097	0.108	0.140
		SP	0.198	0.205	0.258	0.088	0.071	0.061
		SPDP	0.174	0.175	0.190	0.094	0.090	0.095
	$A = 10$	HPPMx	0.184	0.212	0.286	0.097	0.110	0.137
		HPPM	0.185	0.210	0.296	0.100	0.108	0.141
		SP	0.197	0.206	0.258	0.088	0.071	0.062
		SPDP	0.173	0.175	0.189	0.094	0.090	0.095

The true value of k_m in Table 5 is six for all scenarios. It appears that as n increases, the PPMx prior tends to converge to the six clusters faster than the PPM prior. It also appears that the clustering mechanisms of the HPPMx and HPPM depend on A. This is to be expected however, because as A increases curves are allowed to deviate further from cluster specific means, thus creating more clusters.

6 Analysis and Results

In this we section provide results of fitting HPPMx to the NBA data set.

6.1 Model Details and Prior Selection

We first provide a bit of detail regarding cohesion and similarity functions used and then on prior values. The cohesion and similarity functions employed match the PPMx prior

Table 5: Results from the simulation study investigating cluster estimation. Table entries correspond to the number of estimated clusters averaged over 100 simulated data sets.

			$n = 50$			$n = 100$		
			Number of knots			Number of knots		
		Model	5	15	30	5	15	30
$w^2 = 0.1$	$A = 0.1$	HPPMx	5.98	5.98	5.99	5.97	5.98	5.96
		HPPM	3.93	4.26	4.49	4.32	4.37	4.40
	$A = 1$	HPPMx	9.24	8.53	6.52	10.93	10.26	9.02
		HPPM	9.49	9.18	7.89	10.85	10.77	10.02
	$A = 10$	HPPMx	10.32	8.55	6.42	13.39	10.34	8.86
		HPPM	10.45	8.31	7.14	11.84	9.67	8.94
$w^2 = 1$	$A = 0.1$	HPPMx	5.97	5.96	5.99	5.98	5.97	5.99
		HPPM	4.09	4.36	4.72	4.02	4.54	4.66
	$A = 1$	HPPMx	7.25	6.87	6.11	8.20	7.94	6.82
		HPPM	7.08	6.74	6.15	7.94	7.77	7.40
	$A = 10$	HPPMx	7.24	6.89	6.28	8.90	8.06	7.08
		HPPM	5.84	6.03	5.36	7.14	6.88	6.94

to the marginal prior on partitions implied by the DP prior. This results in an a priori clustering of a few large clusters that represent typical player production and a few smaller clusters of "abnormal" players. Thus, we set $c(S_j) = M(|S_j| - 1)!$ with $M = 1$ favoring a small number of clusters. The similarity functions used are typical conjugate models for continuous and categorical variables with parameter values suggested by Müller et al. (2011) resulting in

$$\begin{aligned} g(\boldsymbol{x}_j^\star) &= g_1(\boldsymbol{x}_{j1}^\star)g_2(\boldsymbol{x}_{j2}^\star)g_3(\boldsymbol{x}_{j3}^\star) \\ &= \int \prod_{i \in S_j} N(x_{i1}; m_j, 1)N(m_j; 0, 10)\pi_{i,x_{i2}} \\ &\times Dir(\pi_{i,x_{i2}}; 0.1, 0.1, 0.1)\pi_{i,x_{i3}} Dir(\pi_{i,x_{i3}}; 0.1, 0.1, 0.1) dm_j d\boldsymbol{\pi}_{1j} d\boldsymbol{\pi}_{2j} \\ &= \frac{N_{n_j}(\boldsymbol{x}_j^\star; \mathbf{0}, \boldsymbol{I})N(0; 0, 10)}{N(\hat{m}; 0, \hat{s}^2)} \frac{\Gamma(\sum_{c=1}^3 n_{1jc} + 0.1)}{\prod_{c=1}^3 \Gamma(n_{1jc} + 0.1)} \frac{\Gamma(\sum_{c=1}^3 n_{2jc} + 0.1)}{\prod_{c=1}^3 \Gamma(n_{2jc} + 0.1)}. \end{aligned}$$

$\pi_{i,x_{i2}}$ and $\pi_{i,x_{i3}}$ denote x_{i2} and x_{i3}'s probability vector, n_{1jc} are the number of players in cluster j that have covariate value c for x_{i2} and n_{2jc} the number of players for x_{i3} (as a reminder, x_{i1} corresponds to age, x_{i2} experience and x_{i3} draft order). In addition, $\hat{s}^2 = (n_j + 1/10)^{-1}$ and $\hat{m} = \hat{s}^2 \mathbf{1}' \boldsymbol{x}_{1j}^\star$.

A first-order Bayesian P-spline prior was used and following suggestions in Lang and Brezger (2004), we set $a = 1$ and $b = 0.05$. We found that results were fairly robust to reasonable prior specifications of τ_h^2. From the simulation study setting $A = 1$ seemed reasonable so that individual curves are fairly similar to their cluster-specific counterparts. Preliminary investigations indicated that methodology is robust to variance prior specifications so with hopes of being diffuse we set $a_\sigma = b_\sigma = a_\delta = b_\delta = a_\psi = b_\psi = 1$ and $s_{b_0}^2 = 100^2$. To produce a flat prior for $\boldsymbol{\gamma}$ we use $\boldsymbol{m}_\gamma = \mathbf{0}$ and $s_\gamma^2 = 100^2$. Since there are 82 games in an NBA season we set $m_{a_1} = 76$ (taking into account missed games due to injury) with $s_a^2 = 10^2$.

The MCMC algorithm was run until 1000 iterates from a Markov chain were collected after discarding the first 25,000 as burn in and thinning by 25. Convergence was monitored using MCMC iterate trace plots.

6.2 Fits of Individual Player Production Curves

Figure 3 displays the posterior mean curves with 95% credible and prediction bands for the three players introduced in Figure 1. Notice that even though the fits are fairly flexible they are smoothed relative to the loess fits provided in Figure 1.

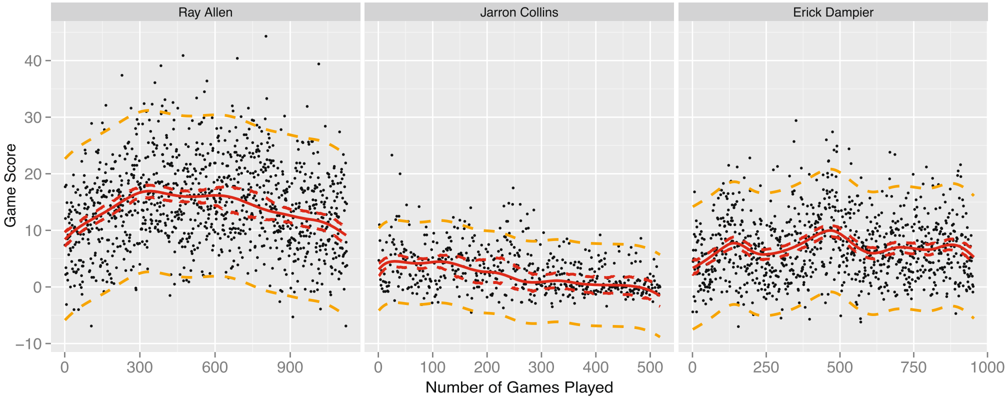

Figure 3: Posterior fits for three NBA players. The solid red lines are point-wise posterior mean curves, the dashed red lines are 95% mean point-wise credible intervals, and the dashed orange lines are 95% point-wise prediction intervals.

6.3 Active Player Prediction

Displaying the results of active player prediction in and of itself is not trivial as curves depend completely on the predicted values of n_i. To simplify the process we display the active player prediction curves conditioned on $E(n_i|\boldsymbol{y}_i)$. This requires producing a curve conditioned on $E(n_i|\boldsymbol{y}_i)$ for each MCMC iterate of $\boldsymbol{\beta}$. From these MCMC iterates, we estimate an average prediction curve with point-wise 95% credible bands and prediction bands. Figure 4 contains the estimated mean curve with credible bands and prediction bands corresponding to four players in varying stages of their career. Shaquille O'Neal played beyond the 2009/2010 season but has since retired. His average number of predicted games played turned out to be 1545 and the actual number of games played is 1423 (including playoffs). Ray Allen continues to play but is nearing the end of his career and the predicted sharp decrease in production mirrors reality. Dwight Howard and Luke Ridnour are two completely different types of players and

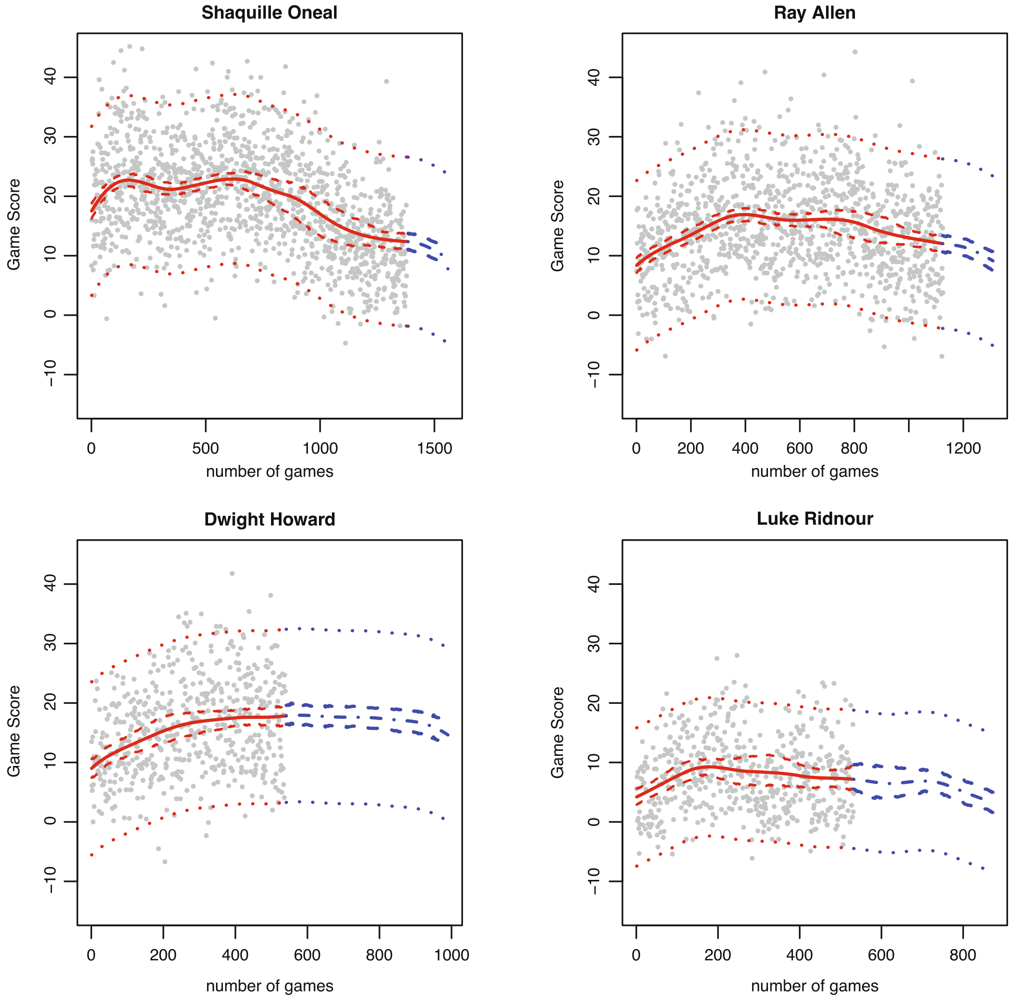

Figure 4: Active player predictions for four NBA players. Lines associated with active player prediction are blue. The dashed lines represent point-wise 95% mean credible bands, and the dotted lines 95% point-wise prediction bands.

are provided as a means to demonstrate the flexibility in the predictions. $E(n_i|\boldsymbol{y}_i)$ for Dwight Howard is quite conservative and barring injury should under estimate his career game total, while $E(n_i|\boldsymbol{y}_i)$ for Beno Udrih is quite reasonable. Regardless, the four predictions display completely plausible decreases in production as the players approach retirement.

6.4 Career Prediction Analysis

For career prediction we employ the predictive distributions as described in Section 4.2

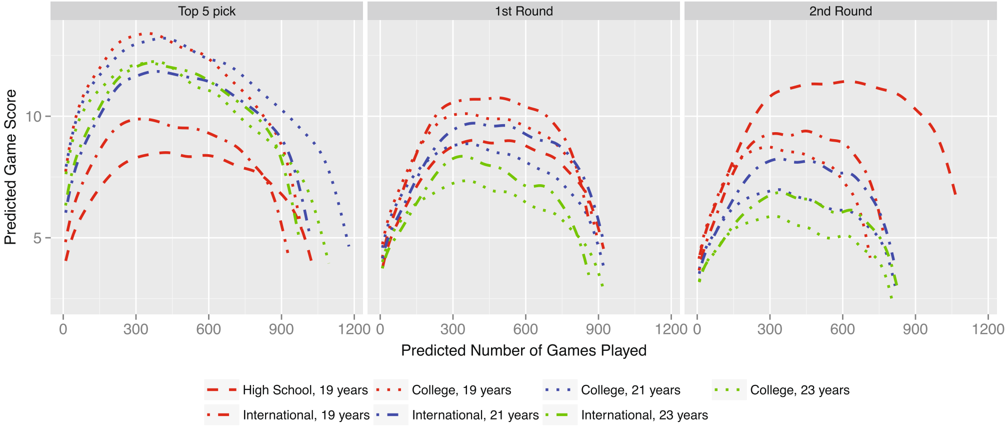

Figure 5: Career prediction curves for different levels of draft order, playing experience and age during first game played.

resulting in the curves found in Figure 5. We include the High School level of experience even though the latest CBA requires at least one year post high school experience before being drafted. For age during first game played, we considered 19, 21, and 23 years old. (We do not consider ages 21 and 23 for High School level of experience as those scenarios are practically impossible.) The curves are presented conditioned on the predicted number of games played averaged over all active players that belong to each respective group.

Before describing results it is important to keep in mind that only players who played at least three years were included in the analysis. This explains the seemingly high predictions for second round picks. Also, from Table 1 it can be seen that only one player (considered in the analysis) was drafted in the second round straight from high school (Rashard Lewis). So you will notice that the predicted curve for this group follows a trajectory similar to that of Rashard Lewis's career. Even with that being the case, a few interesting trends emerge. It appears that there is much more variability in curve location for Top Five Picks. Also the players that are Top Five Picks tend to reach their max production earlier in their career. Age clearly influences a player's production as players that start their career at a younger age tend to have higher production rates. It appears that the shapes of curves vary by experience with international players decreasing slightly earlier relative to college or high school players.

Table 6 provides estimates of the number of games need to reach peak performance. Generally speaking players drafted straight out of high school take longer to reach maximum performance while those with college experience are the quickest. That said, any conclusions drawn from Table 6 or Figure 5 should be made with care as some of the curves are accompanied with a moderate to substantial amount of variability.

Table 6: Predicted game at which max performance is attained (prediction errors are in parenthesis).

		Age During First Game Played		
Draft	Experience	19	21	23
Top 5	High School	477(169.8)	-	-
	College	385(162.2)	437(184.9)	386(178.9)
	International	399(155.6)	448(160.7)	423(147.1)
1st Round	High School	472(147.4)	-	-
	College	436(185.5)	417(193.5)	403(200.2)
	International	446(151.3)	473(172.4)	413(167.3)
2nd Round	High School	568(184.7)	-	-
	College	344(152.6)	386(179.1)	369(179.7)
	International	409(144.6)	416(151.1)	419(169.6)

Cluster Analysis

A nice property of the model is the ability to postulate what characteristics guide clustering. To do this it is necessary to obtain a point estimate using cluster MCMC iterates. Since posterior summaries of cluster specific parameters are arbitrary, using typical posterior summaries (mean and median) makes little sense. We employ Dahl (2006)'s method which is based on least-squares distance from the matrix of posterior pairwise co-clustering probabilities. Using this method produces a partitioning of the 408 players into 18 clusters with cluster membership ranging from 3 to 63 players. Figure 6 provides player-specific posterior mean curves for nine clusters. Except for cluster 6, these clusters represent those that contain the highest number of players (approximately 80% of players). Cluster 6 was selected as it contains curves that are in our opinion more interesting than clusters not shown. The remaining nine clusters for the most part are comprised of role players. Although each cluster contains curves that are slightly unique, they are relatively flat. The dashed segments at the end of some curves are active player predictions. To facilitate comparisons we maintain the x-axis on the percentile number of games played scale. We highlight a few of the clusters by pointing out some of the well known players. Cluster 1 is comprised of role players (e.g., Tony Allen and Matt Bonner) whose production is constant. Cluster 2's key member is LeBron James. Players in this cluster begin careers close to peak level and appear to maintain production. Cluster 3 contains Carlos Boozer and Ron Artest who had sharp increase in production but maintained peak performance for a short time with a gradual decrease in performance. Cluster 4 is comprised of role players who showed an increase in production right before retiring (e.g. James Posey). Kobe Bryant is the key player of Cluster 5 (along with Chauncey Billups, Steve Nash). In this cluster, players started slow but experienced large sharp increases of production and maintained it for much of their career. Clusters 6 and 7 are comprised of players who start at peak performance and gradually decline through out their career with players in Cluster 6 showing a brief increase towards the end of their career. Grant Hill is a member of Cluster 6 and Shaquille O'Neal is a member of Cluster 7. Clusters 8 and 9 are primarily comprised of role players with Cluster 8 showing gradual increase until the later stages of a career.

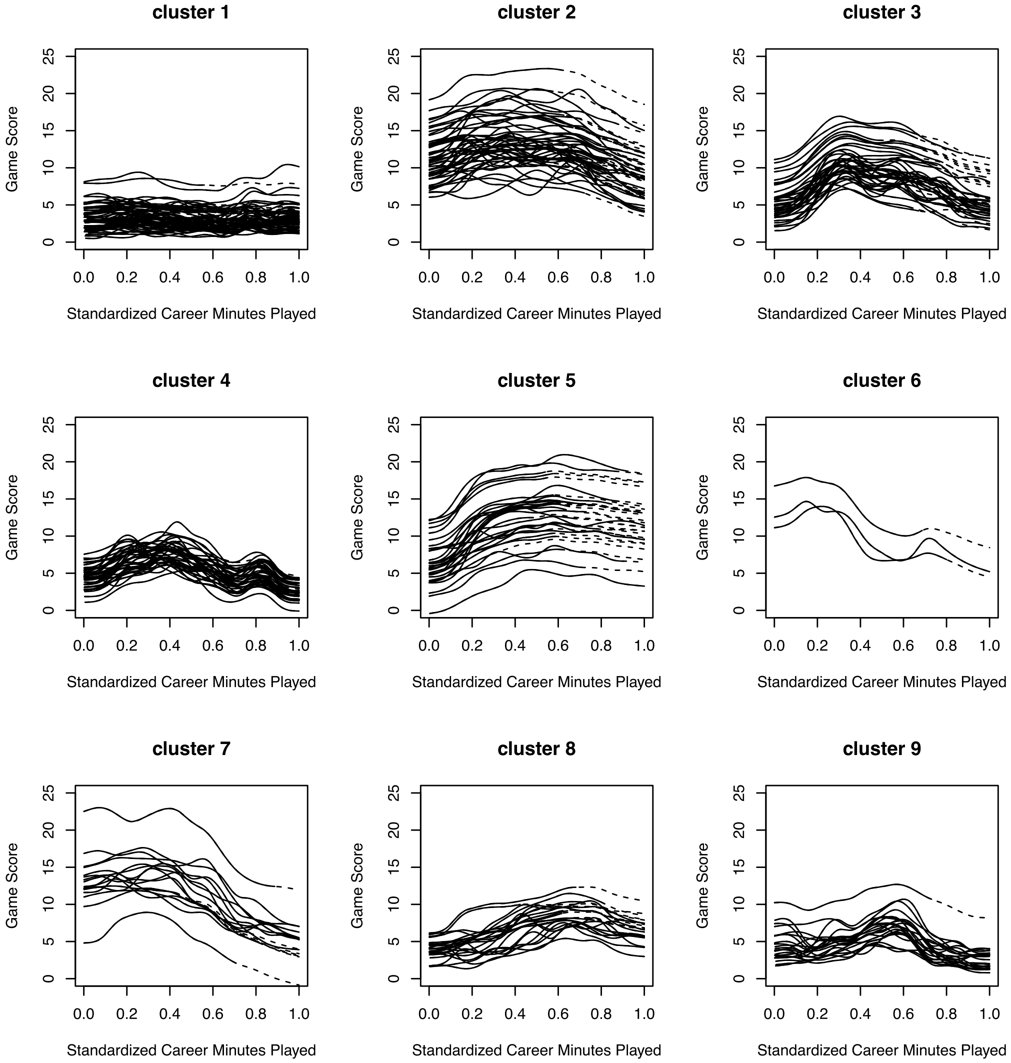

Figure 6: Player specific posterior mean production curves divided into the 9 clusters corresponding to the partition estimated using Dahl (2006)'s cluster estimate method. Active player prediction for active players is displayed by a dashed line.

An example is Matt Barnes. Marcus Camby is member of Cluster 9 and the decrease in production towards the end of the career is more sharp relative to Cluster 8. Overall, the clusters contain curves that have distinct shapes. Information provided in Figure 6 could potentially be used to guide NBA personnel decisions regarding contract length and amount. For example, production for players in 3 begins to decrease earlier than Cluster 2.

Finally, Figure 7 displays the average age during first game played vs. the percent of players in each of the six categories for each of the 18 clusters. Apart from demonstrating the presence of an interaction between the three covariates, these plots confirm what is already widely known. That is, on average, the age of players increases as draft order increases and players that play college tend to begin NBA careers at an older age relative to high school and international players.

6.5 Assessing Trade-off between Individual Player Fits and Prediction

As mentioned previously, incorporating the PPMx in the hierarchical model improves predictions at the cost of a small loss in individual fits. To show that the cost is minimal relative to gains in prediction, we randomly selected 50 retired players and removed the final 25% of games played (essentially treating them as active players). We then proceeded to fit four models (details follow) to these partitioned data and assess model fit through Mean Square Error (MSE). To assess prediction accuracy, active player prediction was carried out for each of the 50 randomly selected players and Mean Squared Prediction Error (MSPE) was computed. The four models considered were the SP model (12), an extension of the SP model that improves prediction, the HPPMx, and the following 5th degree polynomial regression model:

$$\begin{aligned}
&y_{it} = \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{j=0}^{5}\gamma_{ji}t^j + \epsilon_{it} \ \text{ with } \ \epsilon_{it} \overset{iid}{\sim} N(0,\sigma_i^2) \ \text{ and } \ \boldsymbol{\beta} \sim N(\mathbf{0}, s^2\boldsymbol{I})\\
&\boldsymbol{\gamma}_i \sim N(\boldsymbol{\mu}, \boldsymbol{T}) \ \text{ where } \ \boldsymbol{T} = \text{diag}(\tau_0^2, \dots, \tau_5^2)\\
&\boldsymbol{\mu} \sim N(0, s^2\boldsymbol{I}).
\end{aligned}$$

The SP model (12) was extended in the following way

$$\begin{aligned}
\boldsymbol{\theta}_i &\sim N(\boldsymbol{\mu}, \tau^2\boldsymbol{K}^{-1})\\
\boldsymbol{\mu} &\sim N(\mathbf{0}, s^2\boldsymbol{I}).
\end{aligned}$$

We refer to this model as hSP (hierarchical semi-parametric). Predicting (or extrapolating) the last 25% of games played using the SP model requires drawing θ's associated with knots for removed games from its prior distribution. Therefore, centering the prior on a vector of global spline coefficients should improve prediction relative to a prior centered at 0.

The MSE averaged over the 408 players was calculated for each of the four models resulting in Polynomial (30.72), SP (29.66), hSP (29.71), HPPMx (31.71). As expected the flexible penalized splines produce the smallest MSE and HPPMx has the highest MSE illustrating the surrender of a bit of individual player fit. The MSPE averaged over the 50 randomly selected players turned out to be Polynomial (476.32), SP (34.00), hSP (31.10), and HPPMx (26.90). As expected HPPMx gained quite a bit in terms of prediction (extrapolation) accuracy at a fairly minimal individual fit cost.

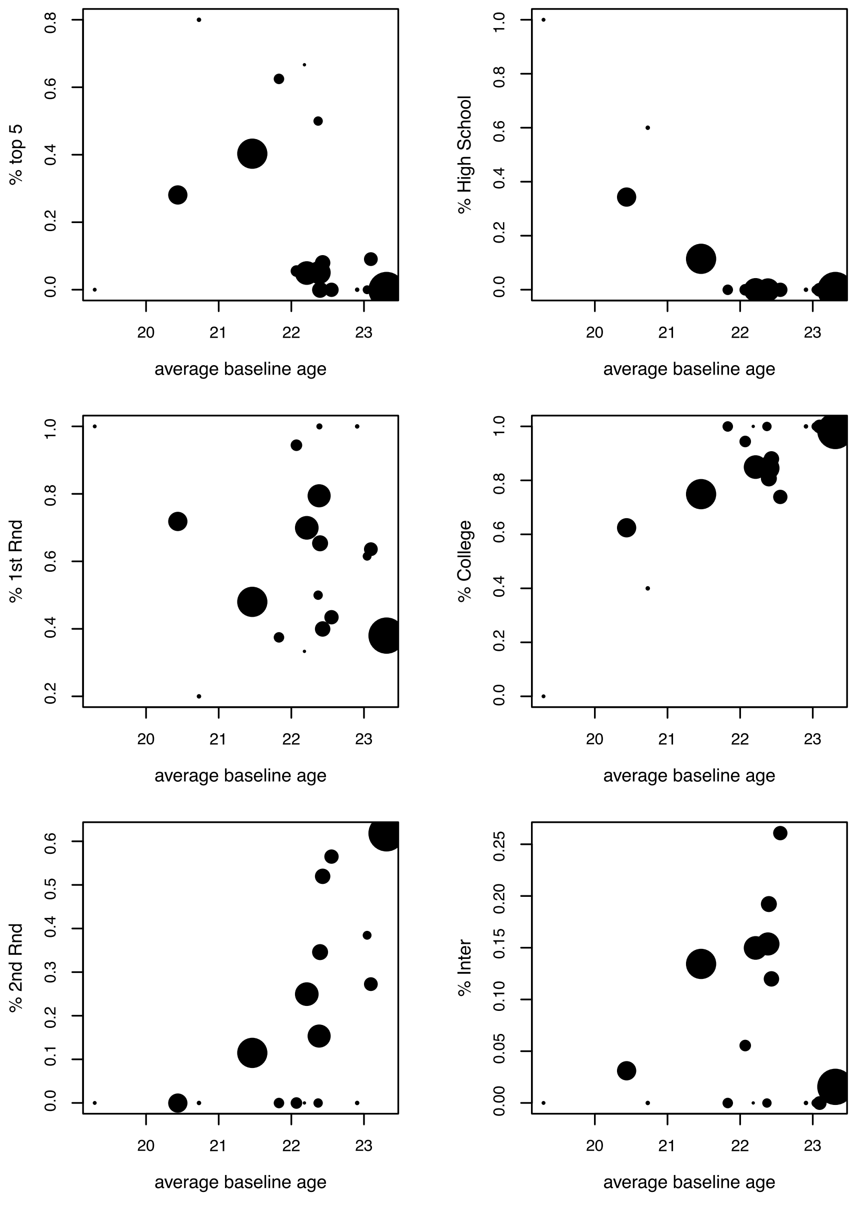

Figure 7: Average baseline age and percent of players for each of the 18 clusters for the six categories corresponding to experience and draft order. The size of the dot is proportional to the number of players allocated to the cluster.

7 Conclusions

We have proposed a completely novel methodology that incorporates information regarding the shape of longitudinal curves in predicting future NBA player production based on age, experience, and draft order of player. Clearly, curve shape provides information beyond available covariates and the inclusion of this information in modeling efforts should improve inferences. In addition, the methodology does well in balancing individual fits and producing clusters that provide adequate borrowing of strength among players. The PPMx prior employed does a nice job of being able to incorporate both covariate and curve shape information when forming clusters and ultimately borrowing strength among subjects to improve active player and career predictions. From a basketball perspective, individual player production clearly depends on many omitted variables (such as team strength, injury history and coaching philosophy) and these variables can be easily incorporated in the model using the PPMx prior when they become available. Finally, though the methodology was demonstrated using production curves of NBA basketball players, the idea of incorporating curve shape in inferences should be applicable in a wide variety of settings (e.g., biomedical, finance, and environmental studies).

Acknowledgments

The authors wish to thank the reviewers and editors for their comments, which improved the paper. The first author was partially funded by grant FONDECYT 11121131 and the second author was partially funded by grant FONDECYT 1141057.

References

Behseta, S., Kass, R. E., and Wallstrom, G. L. (2005). "Hierarchical Models for Assessing Variability Among Functions." *Biometrika*, 92(2): 419–434. 381

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. 382

Bigelow, J. L. and Dunson, D. B. (2007). "Bayesian Adaptive Regression Splines for Hierarchical Data." *Biometrics*, 63: 724–732. 381

Biller, C. (2000). "Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models." *Journal of Computational and Graphical Statistics*, 9: 122–140. 381

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions via Pólya Urn Schemes." *The Annals of Statistics*, 1: 353–355. 390

Collins, L. M. and Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis*. Hoboken, New Jersey: John Wiley and Sons, first edition. 381

Connolly, R. A. and Rendleman Jr., R. J. (2008). "Skill, Luck, and Streaky Play on the PGA Tour." *Journal of the American Statistical Association*, 103: 74–88. 382

Dahl, D. B. (2006). "Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model." In Vannucci, M., Do, K. A., and Müller, P. (eds.), *Bayesian Inference* *for Gene Expression and Proteomics*, 201–218. Cambridge University Press. 396, 402, 403

Dean, N. and Raftery, A. E. (2010). "Latent Class Analysis Variable Selection." *Annals of the Institute of Statistical Mathematics*, 62: 11–35. 381

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. New York: John Wiley & Sons, first edition. 388

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). "Multilevel Functional Principal Component Analysis." *The Annals of Applied Statistics*, 3: 458–488. 381

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). "Bayesian Curve-Fitting with Free-Knot Splines." *Biometrika*, 88: 1055–1071. 381, 388

Elliott, M. R., Gallo, J. J., Ten Have, T. R., Bogner, H. R., and Katz, I. R. (2005). "Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction." *Biostatistics*, 6: 119–143. 381

Fahrmeir, L. and Kneib, T. (2005). *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. New York: Oxford University Press, 1st edition. 388

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, 85: 398–409. 390

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Boca Raton Florida: CRC Press, third edition. 386

Geman, S. and Geman, D. (1984). "Stochastic Relaxation, Gibbs Distribution and Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis of Machine Intelligence*, 6: 721–741. 390

Goldberg, Y., Ritov, Y., and Mandelbaum, A. (2014). "Predicting the Continuation of a Function with Applications to Call Center Data." *Journal of Statistical Planning and Inference*, 147: 53–65. 382

Hollinger, J. (2002). *Pro Basketball Prospectus 2002*. Pro Basketball Forecast. Free Press. 380, 383

Lang, S. and Brezger, A. (2004). "Bayesian P-Splines." *Journal of Computational and Graphical Statistics*, 13(1): 183–212. 388, 390, 398

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: J. Wiley & Sons, 1st edition. 391

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). "Equations of State Calculations by Fast Computing Machines." *Journal of Chemical Physics*, 21: 1087–1091. 390

Montagna, S., Tokdar, S. T., Neelon, B., and Dunson, D. B. (2012). "Bayesian Latent Factor Regression for Functional and Longitudinal Data." *Biometrics*, 68: 1064–1073. 381

Morris, J. S. and Carroll, R. J. (2006). "Wavelet-Based Functional Mixed Models." *Journal of the Royal Statistical Society, Series B*, 68: 179–199. 381

Müller, P., Quintana, F., and Rosner, G. L. (2011). "A Product Partition Model With Regression on Covariates." *Journal of Computational and Graphical Statistics*, 20(1): 260–277. 384, 385, 398

Neal, R. M. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9: 249–265. 390, 391

Park, J.-H. and Dunson, D. B. (2010). "Bayesian Generalized Product Partition Model." *Statistica Sinica*, 20: 1203–1226. 385

Petrone, S., Guindani, M., and Gelfand, A. (2009). "Hybrid Dirichlet Mixture Models for Functional Data." *Journal or the Royal Statistical Society Series B*, 71: 755–782. 382

Quintana, F. A. (2006). "A Predictive View of Bayesian Clustering." *Journal of Statistical Planning and Inference*, 136: 2407–2429. 390

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. New York: Springer, second edition. 381, 394

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press. 381

Telesca, D. and Inoue, L. Y. T. (2008). "Bayesian Hierarchical Curve Registration." *Journal of the American Statistical Association*, 103: 328–339. 388, 389

Wang, S., Jank, W., Shmueli, G., and Smith, P. (2008). "Modeling Price Dynamics in eBay Auctions Using Differential Equations." *Journal of the American Statistical Association*, 103: 1100–1118. 381

Zhu, B. and Dunson, D. B. (2012). "Stochastic Volatility Regression for Functional Data Dynamics." arXiv:1212.0181v1 [stat.AP]. 381

— (2013). "Locally Adaptive Bayes Nonparametric Regression via Nested Gaussian Processes." *Journal of the American Statistical Association*, (504): 1445–1456. 381

Zhu, B., Taylor, J. M. G., and Song, P. X. K. (2011). "Semiparametric Stochastic Modeling of the Rate Function in Longitudinal Studies." *Journal of the American Statistical Association*, 106: 1485–1495. 381

Appendix

1 John Hollinger's Game Score

The Hollinger game score is the following linear combination of statistics appearing in a typical box-score summary of each players game statistics:

- PTS = total points scored by player in game
- FGM = number of shots that player made in game
- FGA = number of shots that player attempted in game
- FTM = number of free throws made in game
- FTA = number of free throw attempts in game
- OREB = number of offensive rebounds
- DREB = number of defensive rebounds
- STL = number of steals
- AST = number of assists recorded
- BLK = number of blocked shots recorded
- TO = number of turn overs
- PF = personal fouls.

$$\begin{aligned} Game\ Score = {} & PTS + FGM \times 0.4 - FGA \times 0.7 - (FTA - FTM) \times 0.4 + OREB \\ & \times 0.7 + DREB \times 0.3 + STL + AST \times 0.7 + BLK \times 0.7 - PF \times 0.4 \\ & - TO. \end{aligned}$$

2 Full Conditionals

We list the full conditionals used in the Gibbs sampler. In what follows we use $[\theta|-]$ to denote the distribution of θ conditioned on all other parameters and data and n_h denotes the number of subjects belonging to cluster h. Also for notational convience, $\boldsymbol{H}_i$ denotes the B-spline basis design matrix for both $g_i = 1$ and $g_1 = 0$.

$$\begin{aligned} [\boldsymbol{\beta}_i|-] &\sim N_m\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right), \\ & \qquad \boldsymbol{\mu}_\beta = \left[\sigma_i^{-2}\boldsymbol{H}_i'\boldsymbol{H}_i + \lambda_{s_i}^{-2*}\boldsymbol{I}\right]^{-1}\left[\sigma_i^{-2}\boldsymbol{H}_i'(\boldsymbol{y}_i - \mathbf{1}_i\beta_{0i}) + \lambda_{s_i}^{-2*}\boldsymbol{\theta}_{s_i}^*\right] \\ & \qquad \boldsymbol{\Sigma}_\beta = \left[\sigma_i^{-2}\boldsymbol{H}_i'\boldsymbol{H}_i + \lambda_{s_i}^{-2*}\boldsymbol{I}\right]^{-1} \\ [\beta_{0i}|-] &\sim N\left(\frac{\sigma_{b_0}^2\sum_{t=1}^{n_i}[y_{it} - g_i\boldsymbol{h}_{it}'\boldsymbol{\beta}_i - (1-g_i)\tilde{\boldsymbol{h}}_{it}'\boldsymbol{\beta}_i] + \sigma_i^2\mu_{b_0}}{n_i\sigma_{b_0}^2 + \sigma_i^2}, \frac{\sigma_{b_0}^2\sigma_i^2}{n_i\sigma_{b_0}^2 + \sigma_i^2}\right), \\ [\sigma_i^2|-] &\sim IG\left(0.5n_i + a_0, 0.5\sum_{t=1}^{i}[y_{it} - \beta_{0i} - g_i\boldsymbol{h}_{it}'\boldsymbol{\beta}_i - (1-g_i)\tilde{\boldsymbol{h}}_{it}'\boldsymbol{\beta}_i]^2 + 1/b_0\right), \\ [\boldsymbol{\theta}_h^*|-] &\sim N_m\Bigg(\left[n_h\lambda_h^{-2*}\boldsymbol{I} + \tau_h^{-2*}\boldsymbol{K}\right]^{-1}\left[\sum_{i\in S_h}\lambda_h^{-2*}\boldsymbol{\beta}_i + \tau_h^{-2*}\boldsymbol{\mu}'\boldsymbol{K}\right], \Big[n_k\lambda_h^{-2*}\boldsymbol{I} \\ & \qquad + \tau_h^{-2*}\boldsymbol{K}\Big]^{-1}\Bigg), \end{aligned}$$

$$\begin{aligned}
[\tau_h^{2*}|-] &\sim IG\left(0.5m + a_\tau, 0.5(\boldsymbol{\theta}_h^* - \boldsymbol{\mu})'\boldsymbol{K}(\boldsymbol{\theta}_h^* - \boldsymbol{\mu}) + 1/b_\tau\right) \\
[\boldsymbol{\mu}|-] &\sim N([\boldsymbol{K}\sum \tau_h^{-2*} + s_\mu^{-2}]^{-1}[\sum \tau_h^{-2*}\boldsymbol{\theta}_h^*\boldsymbol{K}], [\boldsymbol{K}\sum \tau_h^{-2*} + s_\mu^{-2}]^{-1}) \\
[\mu_{b_0}|-] &\sim N\left(\left[m\sigma_{b_0}^{-2} + s_{b_0}^{-2}\right]^{-1}[\sigma_{b_0}^{-2}\sum_i \beta_{0i}], \left[m\sigma_{b_0}^{-2} + s_{b_0}^{-2}\right]^{-1}\right) \\
[\sigma_{b_0}^2|-] &\sim IG(0.5m + a_{b_0}, 0.5\sum(\beta_{0i} - \mu_{b_0})^2 + 1/b_{b_0}) \\
[n_i|-] &\sim TN(\eta_i, \delta^2, \tilde{n}_i, \infty) \\
[L_i|-] &\sim TN(\nu_i, \psi^2, \tilde{L}_i, \infty).
\end{aligned}$$