A switching dynamic generalized linear model to detect abnormal performances in Major League Baseball

Paper Track: Baseball
Paper ID: 1427

Abstract

This paper develops a novel statistical method to detect abnormal performances in Major League Baseball. Abnormally high levels of performance may be caused by myriad factors including performance enhancing drugs (PEDs), banned equipment which offers unfair advantages, and illegal surveillance of opponents. The career trajectory of each player's yearly home run total is modeled as a dynamic process which randomly steps through a sequence of natural ability classes as the player ages. Performance levels associated with the ability classes are also modeled as dynamic processes that evolve with age. The resulting switching Dynamic Generalized Linear Model (sDGLM) models each player's natural career trajectory by borrowing information over time across a player's career and locally in time across all professional players under study. Potential structural breaks from the natural trajectory are indexed by a dynamically evolving binary status variable that flags unnaturally large changes to natural ability, possibly due to unnatural causes such as PED abuse. We develop an efficient Markov chain Monte Carlo algorithm for Bayesian parameter estimation by augmenting a forward filtering backward sampling (FFBS) algorithm commonly used in dynamic linear models with a novel Polya-Gamma parameter expansion technique. We validate the model by examining the career trajectories of several known PED users and by predicting home run totals for the 2006 season. The method is capable of identifying both Barry Bonds and Mark McGwire as players whose performance increased abnormally, and the predictive performance is competitive with a Bayesian method developed by Jensen et al. (2009) and two other widely utilized forecasting systems.

1. Introduction

For the last three decades, Major League Baseball (MLB) has been significantly impacted by anabolic steroids and human growth hormone (HGH). It is widely believed that these substances, collectively referred to as performance enhancing drugs (PEDs), allow players to build more muscle mass and recover more quickly from injuries than is naturally possible. One result of PED use in baseball is that players perform at higher levels later in their careers than they would otherwise. As physical ability diminishes with age, PEDs compensate for natural loss in ability and artificially inflate player performance. As a result, historical milestones in baseball have been artificially surpassed in the 1990s and early 2000s. Not only have PEDs led to unprecedented levels of performance, they have also interfered with fair competition and generated profits for players, team owners, and television networks. For these reasons, PED use in baseball has been the subject of investigations by journalists (Fainaru-Wada and Williams, 2007), law enforcement (Gaines, 2014), MLB itself (Mitchell, 2007), and the United States Congress (Jung, 2005).

PEDs are part of a broader set of performance enhancers that illegally provide a player with unfair

competitive advantages. We define illegal performance enhancers to be any substance, pieces of equipment, in-game strategies, or other methods which increase a player's performance and are forbidden by Major League Baseball. Corked bats, stolen pitch signals, and illegal surveillance also fall into the class of performance enhancers, all of which lead to offensive outcomes that are artificially inflated.

In this paper, we develop a novel statistical method to identify players whose offensive performance exhibits structural breaks that are inconsistent with a player's natural growth and aging process. We propose a hierarchical Bayesian model for a player's home run count trajectory over the course of his career. Our model allows for natural variation in a player's home run total across age. It also allows us to identify players whose home run total is inconsistent with the variability in performance by players of the same age and natural ability. Home run totals which are extreme outliers relative to totals from players of the same age and natural ability may be indicative of an artificial increase in performance. Our model formalizes this unexplained increase in performance with what we call an abnormal performance (AP) indicator.

Several papers in the literature have examined player performance as a function of age. Berry et al. (1999) model the effects of age on performance in baseball nonparametrically. Each player's age curve is modeled as a deviation from the mean aging curve. Albert (2002) models a player's ability to create offense with a quadratic function. Fair (2008) extends the model of Albert (2002) by allowing for two distinct quadratic functions. The model imposes smoothness in careers by requiring the quadratic functions to be equal at the peak performance. Fair (2008) finds that players reach their peak performance at approximately 28 years of age. Jensen et al. (2009) model age curves with cubic B-splines and allow for an elite ability indicator.

B-spline and quadratic models of ability assume smooth age curves. While we agree that natural age curves should change incrementally with time, it has been observed in recent years that abrupt changes sometimes occur in player performance. As an example, Nieswiadomy et al. (2012) determine that two separate structural breaks occur in the performance of Barry Bonds.

To allow for both slowly varying age curves and structural breaks, we combine a dynamic process for natural ability with jumps that are due to the abnormal performance (AP) status variable. We model performance trajectories as a function of age with sticky Markov switching between a finite and fixed number of different ability classes. Our abnormal performance indicator and the associated increase in player performance allow for large jumps in player performance.

The sequence of a player's membership in ability classes is modeled as a Markov process. The probability of membership in an ability class at time t+1 depends only on the current ability class at time t. The Markovian switches in ability class are guided by a sequence of transition matrices which are indexed by age. The transition matrices are constructed to formalize our prior belief that a player begins his career at a moderate ability level, increases in ability until the age of 28, and gradually decays in ability thereafter. This prior belief is informed by data on MLB player performance prior to 1990 and work by Fair (2008).

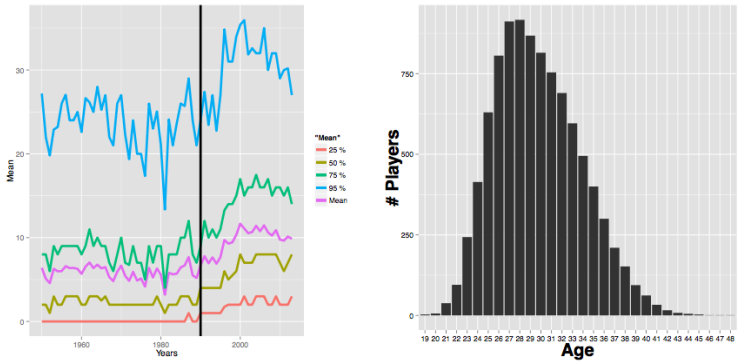
Our natural trajectory model allows incremental changes in ability every time there is a shift in the player's natural ability class. We use a switching model, instead of a fixed DGLM, because it's unreasonable to assume that a player never improves or gets worse in his ability to hit home runs. If that were the case, we would know everything about a player's future performance simply by

knowing where he ranks in the first couple of years of his professional career. Instead of comparing smooth and jumpy career trajectories, we are trying to distinguish between small shifts and large jumps in performance.

The remainder of this paper is structured as follows: Section 2 describes the nature and pre-processing of the data; Section 3 presents the model that we bring to the data; Section 4 discusses the prior distributions we elicit for model parameters and latent variables; Section 5 outlines our MCMC algorithm for model fitting; Section 6 presents inference results for abnormal performances, discusses reproducibility, one year ahead forecasting, and identifiability of ability class and AP status; Section 7 concludes.

2. Data

The data used in our analysis comes from Lahman's Baseball Database (Lahman, 2014). The database contains complete offensive and pitching statistics for every Major League Baseball player since 1871. Data from 1871 to 1949 is ignored, as its relevance to modern baseball is limited. When exploring the 1950-2014 home run data by calendar year, it appears that there are two distinct eras. Baseball is known for having different eras throughout its history. Figure 1a demonstrates that after 1990, percentiles for the distribution of home run totals increase.



(a) Percentiles (b) # Players

Figure 1: Left: Distributional summaries of home run counts of the MLB population by year. The vertical line at the year 1990 separates two distinct eras. Right: Number of players in 1990 to 2014 sample by age.

Figure 1a also demonstrates that the distribution of home run totals is fairly stable between 1950-1989. We call this the post World War II era. The data from 1990-2014 is what we call the PED era.

In this paper, we restrict our focus to players with at least 40 at bats in a season during the PED era. Players falling below 40 at bats in a season are considered reserve players and not the focus of our investigation. If the player had 60 at bats in 1991, only 30 at bats in 1992, and 150 at bats in 1993, the player would be in our sample in 1991 and 1993 only.

In order to share information across player careers that occur in different calendar years, we align players by age. The primary assumption behind this alignment is that from 1990 to 2014, the factors that influence the length and productivity of a career are constant. In other words, it is not important whether a player was 25 years old in 1992 or 2002. The only factor that matters in a

player's home run production is that the player is 25 years old. In this way, we re-index the temporal dimension of the data. Figure 1b presents the total number of players in the 1990-2014 sample by age. We exclude from the analysis ages which have fewer than 50 players in the sample. This prevents us from having many empty or sparsely populated ability classes. With this threshold, the youngest player in our sample is 21 years old. The oldest player in our sample is 40 years old. Our data set includes players at all ages in between.

Table 1 demonstrates the age alignment process for Ken Griffey Jr. and Albert Pujols. The left panel of the table presents their home run totals by year. Note that their careers overlap but do not occur in the exact same years. The right panel presents the same home run totals aligned by player age.

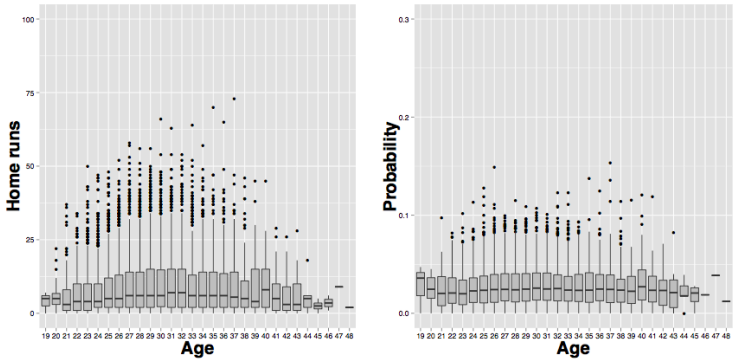
Table 1: Home run totals for Ken Griffey Jr. and Albert Pujols. Left table: home runs by year. Right table: home runs by age.

Year	Griffey	Pujols
1990	22	
1991	22	
1992	27	
1993	45	
1994	40	
1995	17	
1996	49	
1997	56	
1998	56	
1999	48	
2000	40	
2001	22	37
2002	8	34
2003	13	43
2004	20	46
2005	35	41
2006	27	49
2007	30	32
2008	3	37
2009	19	47
2010	0	42
2011		37
2012		30
2013		17

Age	Griffey	Pujols
18		
19		
20		
21	22	37
22	22	34
23	27	43
24	45	46
25	40	41
26	17	49
27	49	32
28	56	37
29	56	47
30	48	42
31	40	37
32	22	30
33	8	17
34	13	
35	20	
36	35	
37	27	
38	30	
39	3	
40	19	
41	0	

The home run data from the PED era is summarized in Figure 2a. The maximum home run total for a single season in MLB history is 73. It was recorded by Barry Bonds in 2001 when he was 36 years old. Bonds also holds the career home run total record at 762 career home runs. Note that while the mean of the distribution increases very slightly in the late 20s, the distribution for the number of home runs is fairly flat across age. Although the number of outliers in the distribution does increase with time, this increase corresponds to the growing number of players in the sample as shown in Figure 1b. We believe the flatness in Figure 2a is due to the selection bias in the sample. Only very talented baseball players reach MLB at young ages. Similarly, only very talented players continue to play in their late 30s and early 40s. For this reason, there is no distinguishable rise or fall in the first and third quartiles of the distribution for home run totals.

One challenge with examining the distribution of home run totals as a function of age is that players in the peak of their career are given more at bats. We examine the empirical distribution of home run rates by age in Figure 2b. The home run rate is defined as the number of home runs hit in a season divided by the number of at bats recorded. Note that, while the number of outliers in the distribution increases slightly, this distribution is also roughly flat during the peak years of the career. Elite players are present in this sample at both early and late ages.



(a) Home runs (b) Rate

Figure 2: Left: Marginal distribution of observed home runs by age in 1990-2014 sample. Right: Marginal distribution of observed rate by age in 1990-2014 sample.

3. Model

Our hierarchical model consists of a binomial sampling model for player home run totals and three sub-models for the dynamics of (1) the ability class levels; (2) the ability class transitions; and (3) the abnormal/natural performance transitions. This hierarchical framework is summarized graphically in Figure 3. The top nodes in the graph denoted by $y_{i,t}$ correspond to home run totals for a specific player. The nodes in the second level of the hierarchy, $\eta_{i,t}$, are the player-specific log odds ratios of hitting a home run. The log odds depends on the three sub-models for ability level (θ_t), ability class ($\gamma_{i,t}$), and the player level abnormal performance indicator ($\zeta_{i,t}$).

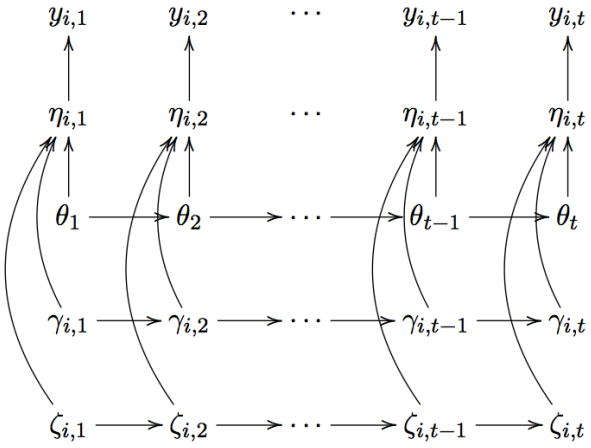


Figure 3: Graphical model representation of our switching dynamic generalized linear model for abnormal performance detection. The temporal dependence of Θ_t, $\gamma_{i,t}$, and $\zeta_{i,t}$ are shown through horizontal directed edges.

The hierarchical approach allows us to jointly model complex dynamic patterns in home run totals for a diverse population of players. For the ability class levels, we leverage state-space models to approximate smooth ability curves. The probability models which govern ability class and

abnormal/natural performance transitions are informed by existing literature.

In our framework, the sequence of a player's home run total is indexed by time points t = 1,...,20 which correspond to ages 21,...,40. At each time t, there are n_t players in the sample. The set of players which are active at time t is indexed by $i \in \{1,...,n_t\}$. The sequence of n_t values is graphically depicted in Figure 1b. $N_{i,t}$ denotes the number of at bats for player i at time t. In those $N_{i,t}$ at bats, player i hits $y_{i,t}$ home runs.

3.1. Sampling model and ability level dynamics

Associated with each player i at time t is an unobserved probability of hitting a home run in a single offensive trial. The probability of hitting a home run depends on the player's latent natural ability class, which is denoted by $\gamma_{i,t}$, and abnormal performance (AP) status, which is denoted by $\zeta_{i,t}$. Because the probability of hitting a home run is a function of both ability class and AP status, we denote it by $\mu_{i,k,z,t}$, where k is a realization of random variable $\gamma_{i,t}$, the latent ability class, and z is a realization of the random variable $\zeta_{i,t}$, the AP indicator. The $\gamma_{i,t}$ class membership variable takes value $k \in \{1,...,K\}$. The natural ability classes are ordered, with players in class k hitting home runs at a lower rate than players in class k + 1. The $\zeta_{i,t}$ AP indicator is a binary random variable. If a player's performance is abnormally inflated in some way, $\zeta_{i,t} = 1$. If a player's performance is natural, $\zeta_{i,t} = 0$.

The total number of home runs hit by player i in year t is modeled with a binomial probability distribution. Observations for players of the same age, ability class, and AP status are independent and identically distributed. We follow Jensen et al. (2009) and condition on each player's yearly at bat total, $N_{i,t}$, throughout the paper.

$$y_{i,t}|N_{i,t}, \mu_{i,k,z,t} \overset{iid}{\sim} Binomial(N_{i,t}, \mu_{i,k,z,t})$$

The unobserved probability, $\mu_{i,k,z,t}$, is the logistic transformation of a parameter, $\eta_{i,k,z,t}$.

$$\mu_{i,k,z,t} = \frac{e^{\eta_{i,k,z,t}}}{1 + e^{\eta_{i,k,z,t}}}$$

The $\eta_{i,k,z,t}$ parameter is the log odds of hitting a home run in a single at bat for a player belonging to ability class k with AP status z. We induce autocorrelation in the marginal distribution for $\eta_{i,k,z,t}$ across age with an underlying state-space representation. For each ability class, k, there is a random ability level, $\theta_{k,t}$. $\theta_{k,t}$ is the state variable for class k at time t. Similarly, for players with abnormally increased performance, there is a random increase in log odds, denoted by $\theta_{AP,t}$. The log odds $\eta_{i,k,z,t}$ is the sum of the level for ability class k and the increase in log odds associated with an abnormal performance.

$$\eta_{i,k,z,t} = \theta_{k,t} + z\theta_{AP,t}$$

The parameter $\theta_{k,t}$ should be interpreted as the log odds of hitting a home run in a single at bat for a player that belongs to natural ability class k without performance inflation (i.e. $\zeta_{i,t} = 0$). $\theta_{AP,t}$ should be interpreted as the increase in log odds associated with an abnormal performance (i.e. $\zeta_{i,t} = 1$). Realization z of the indicator variable $\zeta_{i,t}$ takes values on {0, 1} so that only players whose performance is abnormally enhanced receive the additive increase to the log odds of hitting a home run.

For computational purposes, it is convenient to represent $\eta_{i,k,z,t}$ as the vector product of a regression vector, $F_{i,k,z,t}$ and a state variable, Θ_t. The state variable, Θ_t is a $(K + 1) \times 1$ vector containing the log odds parameters for each respective ability class and $\theta_{AP,t}$.

$$\Theta_t = (\theta_{1,t}, \ldots, \theta_{K,t}, \theta_{AP,t})'$$

The dynamic regression vector, $F_{i,k,z,t}$ will select from Θ_t the particular components that impact the overall log odds of player i hitting a home run at age t. As an example, if $K = 3$, $\gamma_{i,t} = 2$, and $\zeta_{i,t} = 1$, then $F_{i,2,1,t} = (0, 1, 0, 1)'$. That is, the second entry of $F_{i,2,1,t}$ will be unity to encode the membership of player i at time t in class 2. Similarly, the last entry of $F_{i,2,1,t}$ will be unity since $\zeta_{i,t} = 1$. If $\gamma_{i,t} = 3$ and $\zeta_{i,t} = 0$, then $F_{i,3,0,t} = (0,0,1,0)'$. While it is possible to include covariates such as a player's position and ballpark in the state-space representation, we assume that these effects are absorbed into a player's latent ability class membership.

The dynamics of the concatenated Θ_t follow a random walk. As noted in Ferreira et al. (2011), the random walk model for Θ_t should be interpreted as a discretized first-order Taylor series approximation of smooth time-varying functions for ability class levels. One consequence of this model is that players having the same latent ability class and identical AP status are stochastically equivalent.

$$\begin{aligned} \eta_{i,k,z,t} &= F'_{i,k,z,t}\Theta_t \\ \Theta_t &= \Theta_{t-1} + w_t, \qquad w_t \sim N(\Theta_{t-1}, W_t) \\ \Theta_0 &\sim N(m_0, C_0) \end{aligned}$$

In our analysis, we let $W_t^{(K+1)\times(K+1)} = W = .5 \times C_0$, where C_0 is the prior covariance matrix for Θ_0.

3.2. Ability class and abnormal performance transitions

We model player-level AP status, $\zeta_{i,t}$, and latent class membership, $\gamma_{i,t}$, with two independent Markov chains with transition matrices Q^{γ}_t and Q^{ζ}_t. $Q^{\gamma}_{t,k,j}$ represents the element in the k^{th} row and j^{th} column of Q^{γ}_t. An identical notation is used for matrix Q^{ζ}_t.

$$Pr(\gamma_{i,t} = k|\gamma_{i,t-1} = k') = Q^{\gamma}_{t,k,k'}$$
$$Pr(\zeta_{i,t} = z|\zeta_{i,t-1} = z') = Q^{\zeta}_{t,z,z'}$$

The matrices Q^{γ}_t and Q^{ζ}_t are assumed known at all times. While specific construction of these transition matrices will be discussed in detail in Section 4, it is worth noting a few important properties of these Markov chains. The most important property is that Q^{γ}_t and Q^{ζ}_t are constructed so that player-level latent class membership and AP transitions are sticky. If a player belongs to an elite ability class at time t, it is likely he will also belong to an elite class at time t + 1. The same thinking applies to modeling abnormal performance increases. If a player's performance is abnormally inflated at time t, it is more likely his performance will also be abnormally inflated at time t + 1. A second important property is that Q^{γ}_t is constructed so that transitions between neighboring latent classes representing small changes in ability are more likely than transitions between latent classes representing drastic changes in ability. In Section 6, we consider how sensitive our inference is to different choices of Q^{γ}_t.

3.3. Likelihood

Figure 3 presents the graphical model corresponding to a single player across all times t = 1, . . . , T . The joint likelihood can be computed by utilizing the conditional independencies encoded in the graphical model. For notational simplicity, we let the full collection of player-level class membership variables at time t be denoted by an $n_t \times 1$ vector $\Gamma_t = \gamma_{\cdot,t}$. Similarly, we let the full collection of player-level AP indicator variables be denoted by the $n_t \times 1$ vector $Z_t = \zeta_{\cdot,t}$. Let $y_{\cdot,t} = \{y_{1,t},\ldots,y_{n_t,t}\}$. Also let $N_{\cdot,t} = \{N_{1,t},\ldots,N_{n_t,t}\}$.

$$P(y_{\cdot,1},\ldots,y_{\cdot,T}|N_{\cdot,1},\ldots,N_{\cdot,T},\Theta_{1:T},\Gamma_{1:T},Z_{1:T}) = \prod_{t=1}^{T}\prod_{i=1}^{n_t} p(y_{i,t}|N_{i,t},\gamma_{i,t},\zeta_{i,t},\Theta_t)$$
$$= \prod_{t=1}^{T}\prod_{i=1}^{n_t}\binom{N_{i,t}}{y_{i,t}}\left(\mu_{i,\gamma_{i,t},\zeta_{i,t},t}\right)^{y_{i,t}}\left(1-\mu_{i,\gamma_{i,t},\zeta_{i,t},t}\right)^{N_{i,t}-y_{i,t}}$$

The structure of the binomial likelihood will be important in developing our Polya-Gamma Gibbs sampler. We will discuss the data augmentation strategy in more detail in Section 5.

4. Prior distributions

As noted in Section 2, the observed sample of players exhibits selection bias. Only elite players reach the professional league at early ages. Similarly, only elite players are able to continue to play in their late 30s and early 40s. Despite the presence of selection bias, the prior distributions that we elicit are for the ability of an arbitrarily chosen player. We elicit priors for an arbitrary player because we aim to model performance curves which are consistent with physiological aging patterns.

In this section, we first elicit priors for the ability class levels $\theta_{k,t}$. In Section 4.2, we construct the

age-specific probability distributions which govern transitions between ability classes $\gamma_{i,t}$ and $\gamma_{i,t+1}$. The probability distributions which govern transitions between abnormal performance variables $\zeta_{i,t}$ and $\zeta_{i,t+1}$ are developed in Section 4.3. In Section 4.4, we put these different components together to examine the implied marginal prior distribution for the probability of hitting a home run as a function of age. This marginal prior has a realistic age curve because the component prior distributions are chosen to be consistent with information from MLB's Mitchell Report (Mitchell, 2007), physiological aging patterns, and data from the pre-1990 sample.

4.1. Prior distributions for ability class levels

We begin by eliciting prior distributions on the levels associated with each of the K ability classes at t = 0, $\theta_{1,0}, \ldots, \theta_{K,0}$. Each $\theta_{1,0}, \ldots, \theta_{K,0}$ is assumed to be Gaussian distributed with parameters $m_{k,0}$ and σ_k^2. The prior means associated with each ability class, $m_{k,0}$, are chosen to be equally spaced on the log odds interval [−4.5, −2.25]. On the probability scale, this interval corresponds to [0.01, 0.095]. In this analysis, we let K = 15. In Table 2, the expectations of the ability class levels, their respective expected probabilities of hitting a home run, and the expected home run totals for a player with 500 at bats are presented. Sensitivity of inference to difference choices of K is discussed in Section 6.

The variance of the Gaussian distributions for $\theta_{1,0},\ldots,\theta_{K,0}$ is chosen so that neighboring densities intersect at a fraction, β, of the density's maximum value, which implies that $\sigma_k^2 = -\frac{1}{8}\frac{(m_{k+1,0}-m_{k,0})^2}{\log(\beta)}$. In the analysis presented here, $\beta = \frac{2}{10}$. Setting the prior variance σ_k^2 and innovation variance W_t to be low makes it unlikely that label switching between neighboring classes becomes a problem in our MCMC simulation. In the event that it does and the order of Θ_t is shuffled, we re-label the Θ_t in a post-processing step to ensure proper ordering and consistency with the transition matrices Q^{γ}_t.

Recall that $W_t = .5C_0$, so that the k^{th} element of the diagonal matrix W_t is $\frac{1}{2}\sigma_k^2$. We choose to fix this hyperparameter rather than eliciting a hyper prior or using a discount factor method (West and Harrison, 1997) because of the structure in the data. When there are few players in the sample in the younger and older ages, there are many (near) empty ability classes. Hyper prior and discount factor approaches to modeling the dynamic variance term are computationally unstable when the ability classes are empty. With this instability, all hope of reliably imposing an ordering on the classes is lost. For this reason, we fix W_t.

As k increases, the distance between the means on the probability scale, denoted by $\pi_{k,0}$ in Table 2, increases. While the equally spaced Gaussian components lead to distributions on the probability scale that are not equally spaced, this is somewhat compensated for by the associated increase in variance in the binomial distributions. Figures 4c and 4f demonstrate that, despite the apparent gaps in the distribution on the probability scale evident in Figures 4b and 4e, home run counts ranging from zero to those exceeding 75 are well supported by the prior. Designing priors that place (small) mass on high home run counts even without an abnormal performance increase is important. It is not desirable for the model to flag all high home run counts as being abnormal performances.

Table 2: The prior mean of the $K = 15$ ability levels associated with each ability class and the implied means on the probability and home run scale. The HR quantity is $E[y_{i,0}|N_{i,0} = 500, \theta_{k,0} = m_{k,0}, \gamma_{i,0} = k, \zeta_{i,0} = 0]$. The HR Variance quantity is $500 \times \pi_{k,0} \times (1 - \pi_{k,0})$

k	$m_{k,0}$	$\pi_{k,0}$	HR	HR Variance
1	-4.500	0.011	5.493	5.433
2	-4.339	0.013	6.439	6.356
3	-4.179	0.015	7.545	7.431
4	-4.018	0.018	8.837	8.681
5	-3.857	0.021	10.346	10.131
6	-3.696	0.024	12.106	11.813
7	-3.536	0.028	14.156	13.756
8	-3.375	0.033	16.543	15.996
9	-3.214	0.039	19.316	18.570
10	-3.054	0.045	22.532	21.516
11	-2.893	0.053	26.254	24.875
12	-2.732	0.061	30.552	28.685
13	-2.571	0.071	35.500	32.980
14	-2.411	0.082	41.180	37.788
15	-2.250	0.095	47.675	43.129

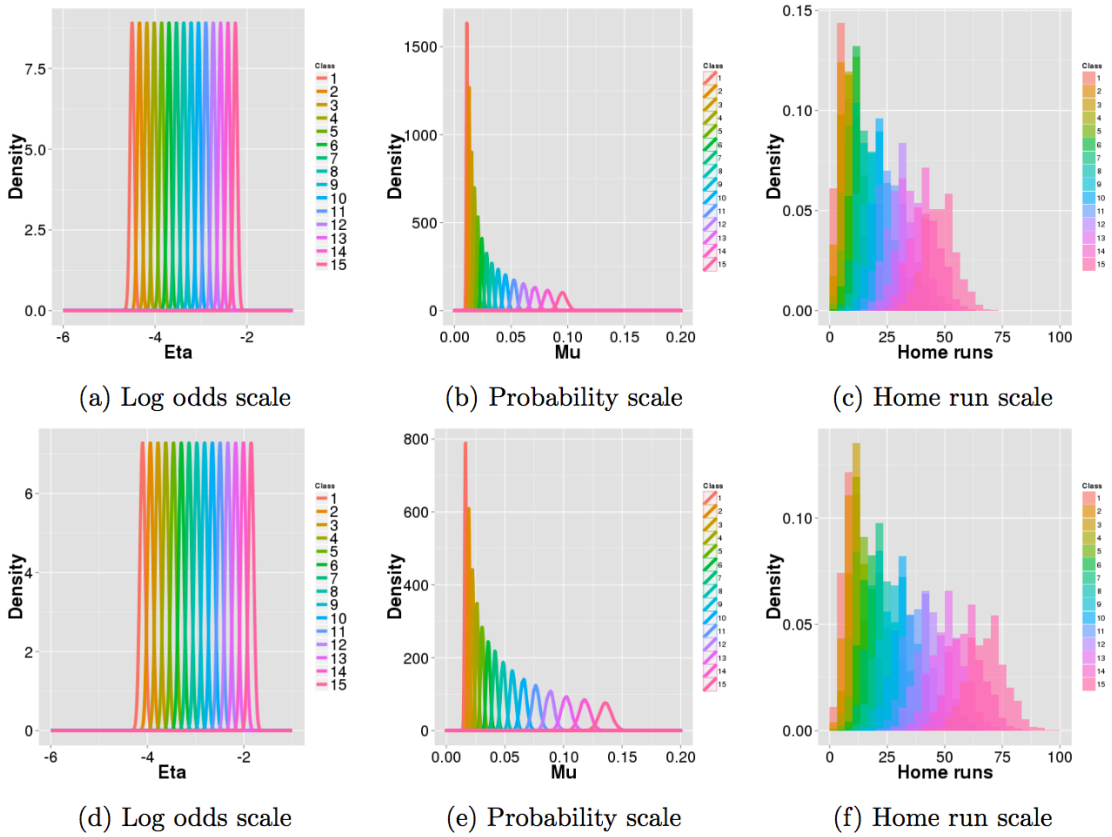
The top row of Figure 4 presents the Gaussian log ability level component distributions, the implied distributions on the probability scale, and the marginal distributions for home run totals in each class. The bottom row of Figure 4 shows the same figures but shifted to the right by the stochastic abnormal inflation. The prior distribution for $\theta_{AP,0}$ is N(0.4,.001). We want the Gaussian prior to be concentrated tightly on significant increases in home run hitting ability.

4.2. Ability class transitions

Transitions between two classes at consecutive time points, $\gamma_{i,t}$ and $\gamma_{i,t+1}$, are modeled with Markov switching. The Markov transition matrices depend on the age of the player. In all transition kernels, it is likely that a player remains in the same ability class from one year to the next. Parameter α captures this stickiness. The higher α, the more likely it is for a player to remain in his current ability class. In our analysis, we let α = 5.

We believe that a player's ability class is more likely to increase than decrease until he reaches 28 years of age. The age 28 is chosen to be consistent with the findings of Fair (2008). We encode this belief in our prior by designing the transition kernel to be asymmetric about the current ability class. If a player is 28 years old or younger, transitions to higher ability classes are more likely than transitions to lower ability classes. For ages less than 28:

$$Q^{\gamma}_{t,k,k'} = Prob(\gamma_{i,t} = k|\gamma_{i,t-1} = k') \propto \begin{cases} e^{-2\times\alpha|m_{k,0}-m_{k',0}|}, & \text{if } k < k' \\ e^{-\alpha|m_{k,0}-m_{k',0}|}, & \text{if } k \geq k' \end{cases}.$$



(a) Log odds scale (b) Probability scale (c) Home run scale

(d) Log odds scale (e) Probability scale (f) Home run scale

Figure 4: The rows correspond to different AP status. All figures in the first row exclude AP effect. All figures in the second row include AP effect. Left column: Prior distributions for the log odds of hitting a home run for each ability class. Middle column: Induced prior distributions for probability of hitting a home run for each ability class. Right column: Histogram of prior distributions for home run totals in 500 at bats for each ability class.

Because the exponent decays twice as fast when $k < k'$, transitions to classes with the same or higher levels of ability are more likely.

We believe that a player reaches his physical peak around 28 years of age and remains near that peak until 32 years of age. Between these ages, transitions are symmetric about the current ability class. It is equally likely that a player's ability class will increase as it is that his class will decrease. If age is between 29 and 32,

$$Q^{\gamma}_{t,k,k'} = Prob(\gamma_{i,t} = k | \gamma_{i,t-1} = k') \propto e^{-\alpha |m_{k,0} - m_{k',0}|}.$$

Once a player reaches 33 years old, we believe that his ability begins to decay. We encode this belief in our model by favoring transitions to a lower ability class. After 32 years of age,

$$Q^{\gamma}_{t,k,k'} = Prob(\gamma_{i,t} = k | \gamma_{i,t-1} = k') \propto \begin{cases} e^{-2\times\alpha|m_{k,0} - m_{k',0}|}, & \text{if } k > k' \\ e^{-\frac{1}{2}\times\alpha|m_{k,0} - m_{k',0}|}, & \text{if } k \leq k' \end{cases}.$$

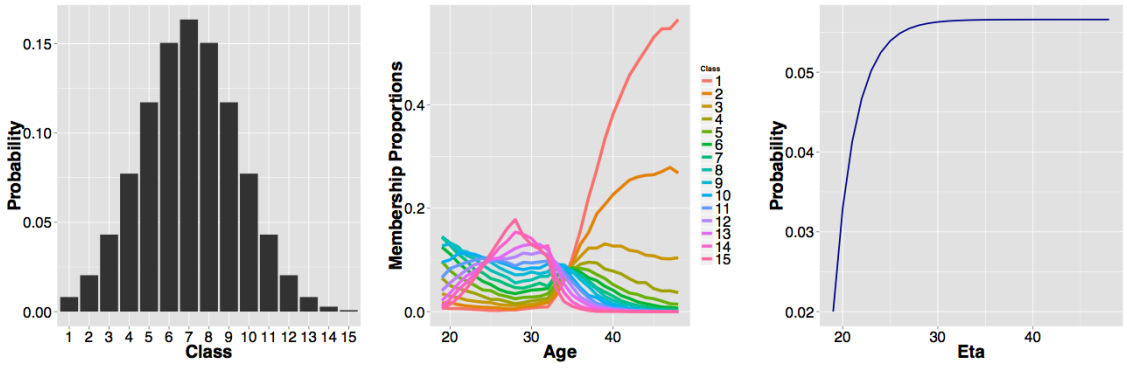
Because the exponent decays four times faster when $k > k'$, transitions to lower ability classes are heavily favored.

We suppose that at age 18, players begin their career in a randomly chosen class. The prior distribution for the initial class at age 18 is proportional to a squared exponential: $P(\gamma_{i,0} = k) \propto e^{-\frac{1}{2c}(k-k_0)^2}$. Figure 5a presents the prior distribution for a player's ability class at age 18 with $c = 3$ and $k_0 = 7$. For players whose first year in the professional league occurs after age 18, the prior distribution for a player's initial ability class at the age when he first enters the sample is the marginal probability distribution of classes implied by the Markov switching at that age of first entry. Figure 5b presents the prior probability of class membership across ages.

4.3. Abnormal and natural performance transitions

In addition to the Markov switching for ability class, we also model the switching of AP status with a Markov process. Figure 5c shows the marginal probability that $\zeta_{i,t} = 1$ across age. The stationary distribution of the Markov chain is chosen to be between 5-7%. In 2003, an anonymous round of PED testing was performed by MLB. They found that 5-7% of the random sample of players chosen tested positive for PEDs (Mitchell, 2007). We use this 5-7% as a conservative estimate of the number of players whose performances are abnormally inflated.

The AP transition kernel is constant in time. We specify the transition kernel to implement a prior belief that if a player's current performance is abnormal, it will likely be abnormal in subsequent time points as well. If a player is currently following a natural growth trajectory, he will likely follow a natural aging pattern in the future. We also choose the transition kernel to reflect the desired stationary probability of 5-7%.



(a) Age 18 class (b) Class membership (c) AP

Figure 5: Left: Prior probability of class membership at age 18. Middle: Marginal probability of class membership over time. Right: marginal probability of AP status over time.

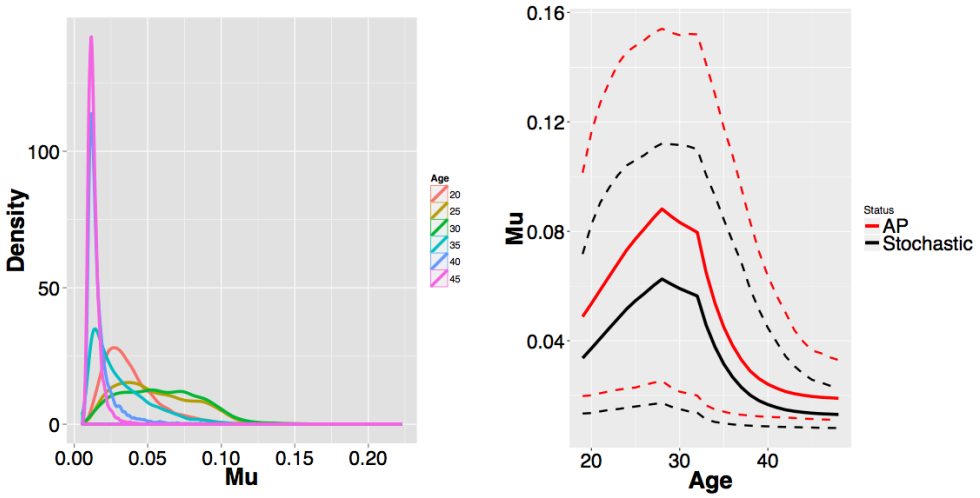
$$Prob(\zeta_{i,t} = 1|\zeta_{i,t-1} = 1) = \frac{2}{3}$$
$$Prob(\zeta_{i,t} = 0|\zeta_{i,t-1} = 0) = \frac{49}{50}$$

Note that we do not have information about the use of performance enhancers by age. As a result, we want the Markov chain to reach its stationary distribution quickly. This is demonstrated in Figure 5c.

4.4. Implied marginal prior distributions

One way of assessing the sensibility of the prior distributions we have elicited is to examine the implied marginal prior distributions for home run counts and probabilities as a function of age. Figure 6a presents the densities for the marginal probability of hitting a home run at ages 20, 25, 30, 35, 40, and 45. We denote the marginal probability of player i hitting a home run at time t as $\mu_{i,t}$. The prior for $\mu_{i,t}$ favors moderate probabilities at ages 20 and 35. For ages 25 and 30, the priors are very diffuse. The densities are quite peaked at low probabilities for ages 40 and 45.

Figure 6b presents the marginal distribution for $\mu_{i,t}$ plotted against the age of the player. The solid black line is the prior mean for $\mu_{i,t}$. The dashed black lines represent the 95% prior credible interval. The solid red line is the prior mean for $\mu_{i,t,\zeta_{i,t}=1}$. That is, it is the prior probability of hitting a home run for a player who always delivers abnormal performances. Again, the red dashed lines represent the 95% prior interval of uncertainty. Figure 6b demonstrates two important features of the marginal prior distribution for $\mu_{i,t}$. The first feature is that the prior distribution encodes our belief that performance rises in the early twenties, peaks at 28, and diminishes in the thirties. The second important feature is the wide uncertainty intervals. We have placed prior mass over very reasonable ranges of the rate at which players hit home runs.

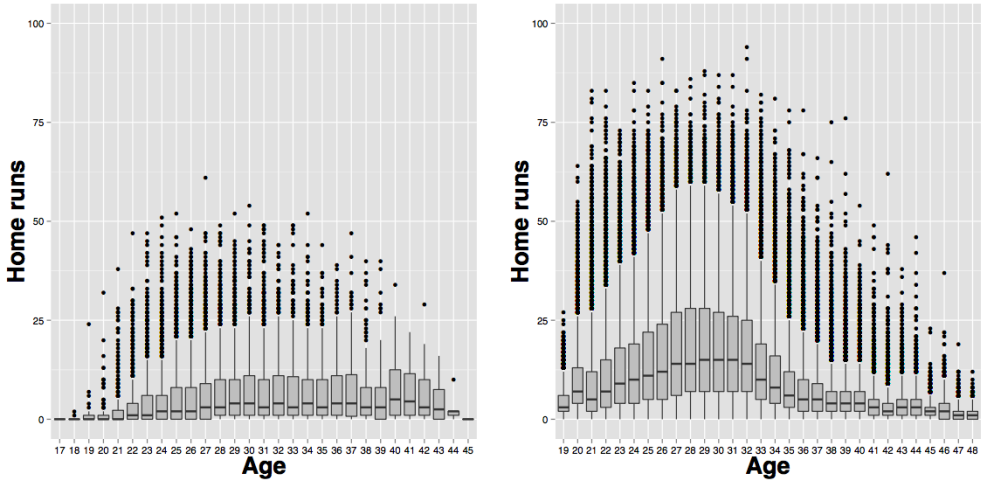


(a) $\mu_{i,t}$ Densities (b) $\mu_{i,t}$ vs Age

Figure 6: Left: Mixture prior densities for home run probability over time. Right: Implied age curve for players who always have AP indicator on and players with uncertain AP status.

In addition to examining the marginal distribution for the probability of hitting a home run as a function of age, we can examine the implied marginal distribution for home run counts as a function of age. Figure 7a presents the empirical distribution of home run counts by age in the post World War II sample, which spans 1950-1989. In this sample, outliers for home run counts reach the middle 50s and low 60s. Figure 7b presents the marginal prior distribution for home run counts excluding the abnormal inflation effect: $p(y_{i,t}|\zeta_{i,t} = 0)$. Outliers in this distribution are in excess of 75 home runs.

Note that the marginal prior distribution is for an arbitrarily chosen player and is not intended to account for selection bias. Also note that we have favored higher home run counts than those exhibited in Figure 7a for two reasons. First, we believe modern medicine, nutrition, and physical training have increased the home run production of players when compared to those in the post World War II sample. Second, we want our method to be conservative in detecting home run anomalies. We have placed enough prior mass in elite natural ability classes that it is possible with our prior specification for a player to be a historically elite home run hitter without the benefit of abnormal inflation. We don't want all elite home run hitters to be automatically flagged as abnormal by our method. By placing prior mass on natural home run totals which are historically high, we preclude such a scenario.



(a) Empirical HR 1950-1989 (b) Marginal $p(y_{i,t} = j|\zeta_{i,t} = 0)$

Figure 7: Left: Observed marginal distribution of home runs in data from 1950-1989. Right: Marginal prior distribution for home runs in data from 1990-2014.

5. Markov chain Monte Carlo

The sDGLM is a switching state-space model. Switching state-space models have a long history in the time series literature (Shumway and Stoffer, 1991; Kim, 1994; Fox, 2009). Traditionally, switching state-space models have been utilized to model dynamic processes in which the observed data is continuous. Model fitting has relied on Markov chain Monte Carlo (Fruhwirth-Schnatter, 2001), variational inference (Ghahramani and Hinton, 2000), and particle based methods (Whiteley et al., 2010). For discrete data, Gamerman (1998) developed a Metropolis-Hastings algorithm for the DGLM. We develop a Gibbs sampling algorithm that utilizes Polya-Gamma data augmentation (Polson et al., 2013) for posterior inference. Our data augmentation strategy allows

us to use a forward filtering backward sampling algorithm commonly used for dynamic linear models (Carter and Kohn, 1994; Frhwirth-Schnatter, 1994).

Before outlining the MCMC strategy, it is worth summarizing the parameters and state variables that we seek to estimate. Our target posterior distribution is $p(\Theta_{1:T}, \Gamma_{1:T}, Z_{1:T} | y_{\cdot,1:T})$. Informally, we are trying to estimate one set of parameters and two sets of latent variables. At each time point, we estimate the parameters $\theta_{k,t}$ for $k \in \{1, \ldots, K\}$ and $\theta_{AP,t}$. In terms of latent variables, for each player and age, we estimate his latent natural ability class membership, $\gamma_{i,t}$, and his AP status, $\zeta_{i,t}$.

Because our observed data has a binomial sampling distribution, we are able to take advantage of recent advances in data augmentation methods for Bayesian logistic regression. Polson et al. (2013) developed a data augmentation method for logistic regression with a Polya-Gamma random variable. We augment the likelihood for $y_{i,t}$ with a Polya-Gamma random variable $\omega_{i,t}$. A player's home run count $y_{i,t}$ and at bat total $N_{i,t}$ enter the likelihood through $\kappa_{i,t} = y_{i,t} - \frac{N_{i,t}}{2}$.

$$p(y_{i,t}|\Theta_t, \gamma_{i,t} = k, \zeta_{i,t} = z, \omega_{i,t}) \propto e^{\kappa_{i,t} F'_{i,k,z,t} \Theta_t - \frac{\omega_{i,t}}{2}\left(F'_{i,k,z,t}\Theta_t\right)^2}$$

The Polya-Gamma augmentation strategy allows for conditionally Gaussian updates of state variables, Θ_t. Because of the conditionally Gaussian structure, it is easy to implement a forward filtering backward sampling (FFBS) algorithm. MCMC samples are drawn by iteratively sampling from the following full conditionals in a Gibbs sampler.

1. Sample $\Theta_{1:T}|Z_{1:T}, \Gamma_{1:T}, y_{\cdot,1:T}$.
2. For each player, jointly sample $\gamma_{i,1:T}, \zeta_{i,1:T}|\Theta_{1:T}, y_{i,1:T}$. It is possible to parallelize this sampling step across players.
3. For each player and time, sample $\omega_{i,t}|\gamma_{i,t} = k, \zeta_{i,t} = z, \Theta_t$. It is possible to parallelize this sampling step across players and ages.

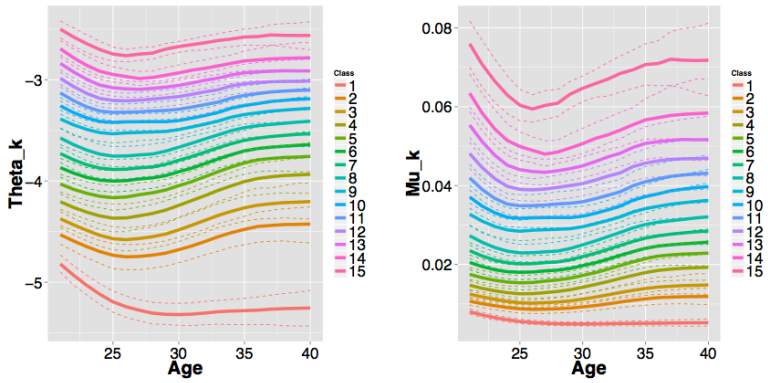
We run the Gibbs sampler for 20, 000 iterations and discard the first 10,000 as a burn-in. In addition, we thin the samples by only recording every tenth sample. This leaves us with 1, 000 post burn-in samples. The computation required approximately five hours on a single core. The R code which implements our MCMC algorithm can be downloaded from the GitHub page https://github.com/G-Lynn/sDGLM.

6. Results

Our Markov chain Monte Carlo inference procedure provides us with a full set of posterior distributions for model parameters and latent variables. In Section 6.1, we examine the posterior distributions for ability class levels. We proceed to present player-specific inferences for Derek Jeter, Mark McGwire, Alex Rodriguez, and Barry Bonds in Section 6.2. In addition to validating the model's ability to flag suspicious performances for known cases of PED use, we assess its predictive performance in Section 6.3. Further, we demonstrate that our MCMC algorithm generates reproducible inferences in Section 6.4 and discuss the sensitivity of inference to parameter choices in Section 6.5.

6.1. Posterior distribution for ability class levels

We begin our discussion by presenting posterior distributions for each of the K ability class levels and the increase in performance associated with the AP indicator. Figure 8a shows the posterior distribution for each of $\theta_{k,t}$. The dashed lines of the same color as the solid line represent the 95% posterior credible interval. At the earliest and oldest ages, the ability class levels are generally highest. This finding is consistent with selection bias in the sample. Elite players form our sample at the earliest and latest ages. As players with a wider range of natural abilities enter the sample in their early twenties, ability levels drop. When players of modest ability gradually exit the sample beginning in their late twenties, ability levels gradually rise again.



(a) Θ_t (b) μ_t

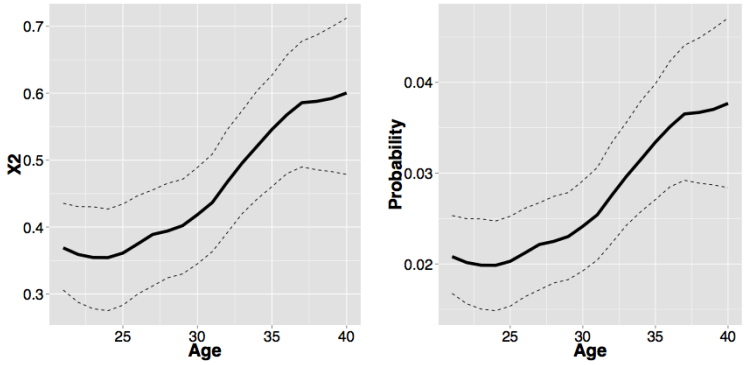
Figure 8: Posterior summaries for Θ_t and μ_t.

While the dynamics of $\theta_{k,t}$ are not dramatic, the evolution of ability levels on the probability scale suggests more dynamic behavior in each ability class. This is particularly true of classes 13 through 15. Class 15 drops by almost 2% before increasing by 1%. It makes sense that the highest classes are the most dynamic on the probability scale because the logistic transformation is nonlinear.

The natural ability classes presented thus far ignore the impact of the AP indicator. Figure 9a presents the dynamic variation in the posterior distribution for $\theta_{AP,t}$. The dashed lines represent the 95% posterior credible interval. For players in their twenties, the mean of the posterior distribution for $\theta_{AP,t}$ is lower than the t = 0 prior mean of 0.4. For players in their thirties, the increase in log odds associated with the AP indicator exceeds the expected value of the prior.

One of the open questions in the literature is how much does PED use inflate home run totals. Schmotzer et al. (2008) estimate that steroid use increased the metric adjusted runs created in 27 outs by approximately 12%. In Figure 9b, we present the increase in probability of hitting a home run due to our abnormal performance indicator. We assume an ordinary baseline player who hits home runs in 5% of his at bats. For players in their late 30s, having the AP indicator turned on increases the probability of hitting a home run by approximately 3.7% for an overall home run probability of 8.7%. Another way of stating this is that, for the natural 5% home run hitter, having the AP indicator on increases his home run rate by approximately 1.75 times. For the same 5% home run hitter, having the AP indicator on increases his home run total by approximately 18 home

runs in every 500 at bats.



(a) $\theta_{AP,t}$ (b) Increase

Figure 9: Posterior summaries for $\theta_{AP,t}$ and increase in μ_t for a player who hits home runs at 5%.

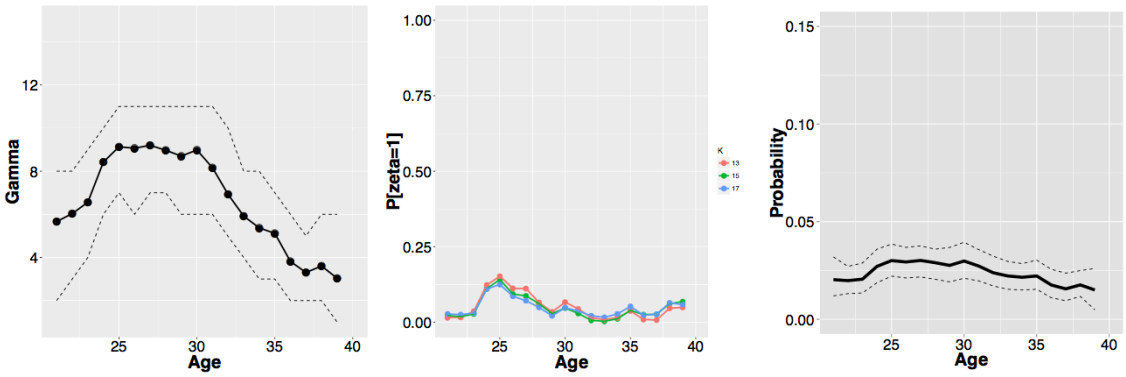
6.2. Player level inference

An important feature of our model and computational method is the ability to make player specific inferences for ability class and AP status. Below we present the inferences for Derek Jeter, Mark McGwire, Alex Rodriguez, and Barry Bonds. It is widely believed that Jeter abstained from PEDs and other illegal performance enhancers. Mark McGwire (Kepner, 2010) and Alex Rodriguez (Weaver, 2014) have admitted steroid use, and Bonds has admitted that he unknowingly used steroids (Washington Post, 2011). While not presented here, we have estimated ability class membership, AP status, and ability curves for every player in our sample.

In Figure 10, we present the ability class membership, probability that the AP indicator is unity, and latent ability curve for Derek Jeter. We present the results for Derek Jeter to establish the capability of our method to infer a traditional age curve. Figure 10a demonstrates that Jeter's ability class membership rises and falls as human aging suggests it should. In Figure 10b, the posterior mean of $\zeta_{i,t}$ is presented. Since $\zeta_{i,t}$ is binary, one way of interpreting $E[\zeta_{i,t}|y_{\cdot,1:T}]$ is the posterior point estimate of the probability that the performance of player i in year t is abnormally inflated. Also note that in Figure 10b, there are three sets of probabilities corresponding to different choices for K. Observe that all three choices for K demonstrate the same qualitative behavior. In each of these choices, Derek Jeter has a very low probability of abnormal performance across his career. In addition, because we do not impose any thresholding or sparsity in the probability, our model does not support probabilities that are exactly zero. The probabilities may be low, but not zero. This indicates that all low probabilities should be treated as providing no evidence of abnormal performance. In Figure 10c, we present the posterior inference for the ability curve of Jeter. This ability curve includes the increase in ability a player might receive from performance enhancers. For each player in the sample, it is possible for us to estimate when he reached his peak performance. Our analysis demonstrates that Derek Jeter maintained his peak home run hitting ability from approximately 25 to 30 years of age before his ability gradually diminished.

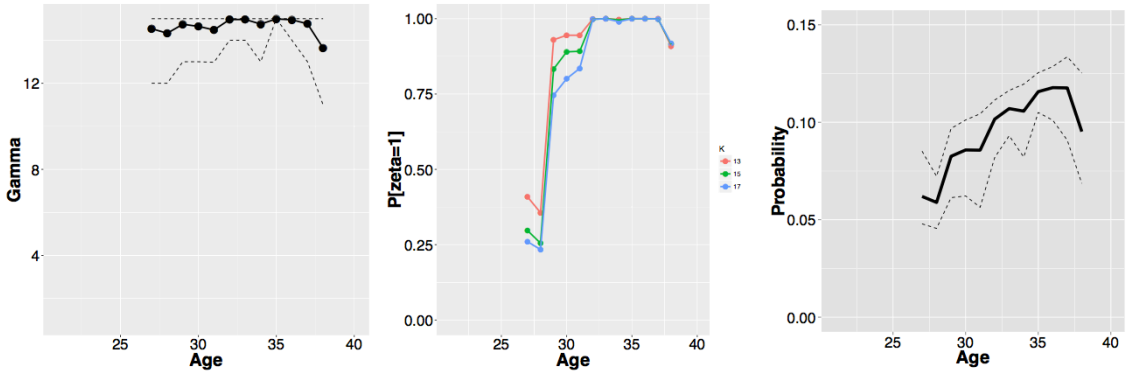
We present the results for Mark McGwire because he set the single season home run record by hitting 70 home runs in 1998. Figure 11a demonstrates that, with a high degree of certainty, Mark McGwire is a member of the highest ability class. In addition to this membership in the highest ability class, Figure 11b demonstrates that his offensive output is extremely abnormal. McGwire

was 34 years old in his record setting season of 1998. The ability curve for McGwire, which is presented in Figure 11c, demonstrates an unusual aging pattern where his ability to hit home runs continues to increase late into his career.



(a) Class Membership (b) $E[\zeta_{i,t}|y_{\cdot,1:T}]$ (c) Ability Curve

Figure 10: Ability class membership, AP inference, and ability curves for Derek Jeter. The points are the estimates, but lines are included for visual clarity.

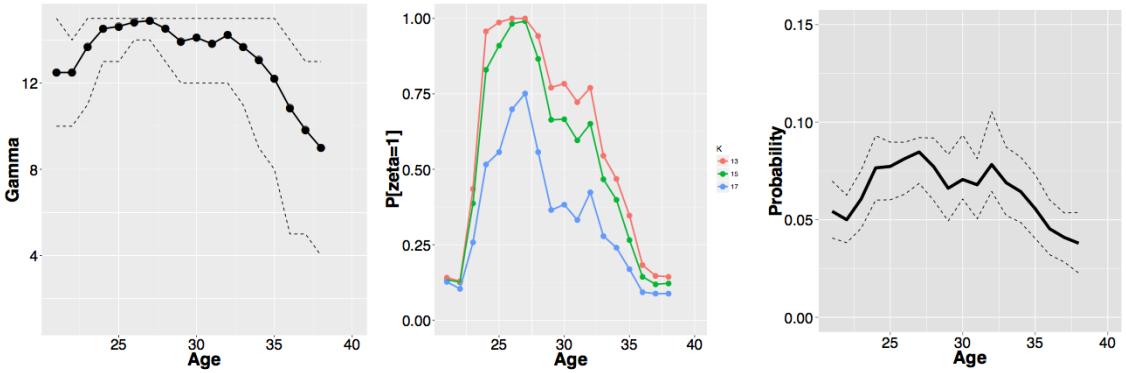


(a) Class Membership (b) $E[\zeta_{i,t}|y_{\cdot,1:T}]$ (c) Ability Curve

Figure 11: Ability class membership, AP inference, and ability curves for Mark McGwire. The points are the estimates, but lines are included for visual clarity.

One of the challenges of detecting abnormal performances is the entanglement of ability class and AP status. Is a player an elite home run hitter because he is a member of an elite ability class? Or is he an elite home run hitter because he uses performance enhancers? Resolving this identifiability problem is critical to reliably inferring both ability class and AP status. In this paper, we have attempted to resolve the identifiability issue through the prior distribution. Figures 12 and 13 present the ability class membership, probability that the AP indicator is unity, and ability curve for Alex Rodriguez and Barry Bonds. Both players were elite home run hitters over the course of their careers, and both players have been tied to PED use. At the age of 32, Bonds hit 42 home runs; however, the method only identifies Bonds' performances as likely abnormal after the age of 35. The method only identifies Rodriguez as a likely beneficiary from an abnormal increase until the

age of 28, yet in 2007, at the age of 31, Rodriguez hit 54 home runs. Both of these performances demonstrate that it is possible for players to record high home run totals without being definitively flagged by our method as abnormal.

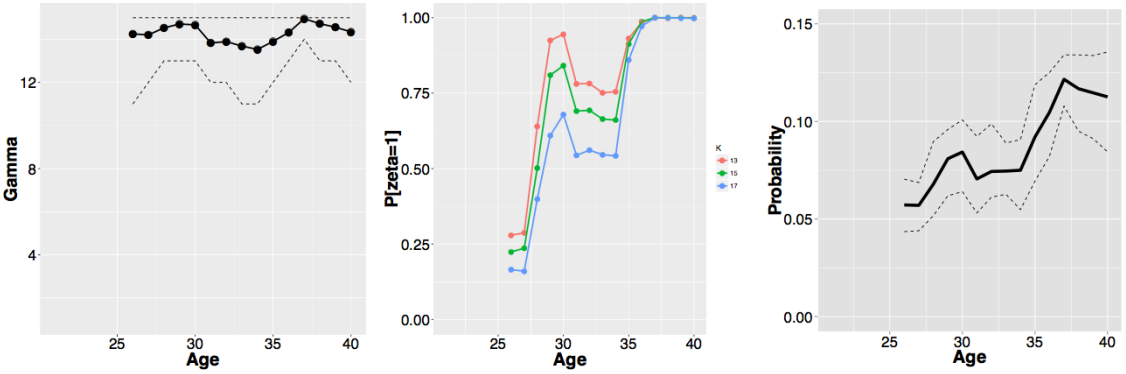


(a) Class Membership (b) $E[\zeta_{i,t}|y_{\cdot,1:T}]$ (c) Ability Curve

Figure 12: Ability class membership, AP inference, and ability curves for Alex Rodriguez. The points are the estimates, but lines are included for visual clarity.

Figures 12a and 13a confirm the elite natural ability of both Rodriguez and Bonds. Both players are members of the highest natural ability classes. Figures 12c and 13c provide an important contrast in their career trajectories. While Rodriguez's ability level decreases in his late twenties and early thirties, Bonds ability to hit home runs continues to increase.

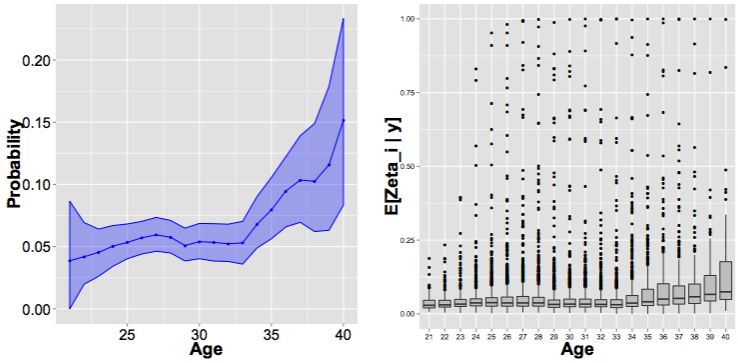
When aggregating AP status across players, we get a sense of the proportion of the population whose performance is abnormally inflated. Figure 14a presents the posterior distribution of the expected proportion of the population with abnormally inflated performances. Formally, it is the distribution of the posterior mean across the sample: $E[Z_t|y_{\cdot,1:T}]$. The prior distribution for AP status elicited in Figure 5c demonstrated that the prior probability of abnormal performance quickly reached a stationary distribution around 5%. The posterior expectation presented in 14a hovers around 5% in the early and mid twenties and then significantly increases in the late thirties.



(a) Class Membership (b) $E[\zeta_{i,t}|y_{\cdot,1:T}]$ (c) Ability Curve

Figure 13: Ability class membership, AP inference, and ability curves for Barry Bonds. The points are the estimates, but lines are included for visual clarity.

At age 40, the expected proportion of players whose performance is abnormally inflated is 15%. The posterior distribution of $E[Z_t|y_{\cdot,1:T}]$ indicates that as players age, they are significantly more likely to be flagged for abnormal performances.



(a) $E[Z_t|y_{\cdot,1:T}]$ (b) $E[\zeta_{i,t}|y_{\cdot,1:T}]$

Figure 14: Summaries for proportion of population whose performance is abnormal. Left: Distribution of $E[Z_t|y_{\cdot,1:T}]$, the expectation of the AP indicator across players. Right: Distribution of $E[\zeta_{i,t}|y_{\cdot,1:T}]$ for full population.

Figure 14b presents the posterior distribution of $E[\zeta_{i,t}|y_{\cdot,1:T}]$ for the full population at each age. Each point in the distribution is the expected value of $\zeta_{i,t}$ for a single player. Figure 14b shows that most players are not associated with performance inflation. It is interesting to observe the continuum of the distribution. Mass in the posterior is concentrated near zero with outliers being distributed along the continuum of probabilities from zero to one. It is important that the method be able to express its uncertainty over player level AP indicators with expected values in the middle of the unit interval.

6.3. Prediction

In addition to validating the model by examining specific cases of known PED use, we validate the model by examining its ability to predict a player's future home run total. To assess the predictive capability, we conducted a second analysis where the data was constructed from a sample beginning in 1990 and ending in 2005. The prediction exercise is to forecast home run performance out of sample in 2006. We choose these years to coincide with a predictive analysis conducted in Jensen et al. (2009). Just as in Jensen et al. (2009), we consider the predictive performance for the full sample and a sample of 118 elite players.

Table 3 presents the predictive performance of the sDGLM compared against the method of Jensen et al. (2009), Pecota, a commercially available and hand curated forecasting system, Marcel, a widely used open source forecasting method, and a naive forecast, which is the home run total from the previous season. The sDGLM is competitive in its predictive performance for the full sample with Jensen et al. (2009). For the sample of 118 top home run hitters, the sDGLM outperforms Marcel but falls slightly behind both Pecota and Jensen et al. (2009).

Table 3: Predictive performance of sDGLM compared to Jensen et al. (2009), Pecota, Marcel, and a naive forecast where the forecast for 2006 is the home run total from 2005.

	Top 118	Full Sample
Pecota	7.11	-
Jensen et al. (2009)	7.33	5.30
sDGLM	7.73	5.37
Marcel	7.82	-
Naive	11.11	8.54

Figure 15 presents the RMSE of our method plotted against age. Despite the relatively large number of players in the sample between ages 25 and 30, the RMSE is not significantly different for those ages than for ages with relatively few players. In fact, the RMSE has a negative trend with age. Since our model learns sequentially with age, this makes sense. Prediction accuracy that improves with age has the added benefit of delivering increasingly reliable predictions for players in the years of their careers when they earn the most money.

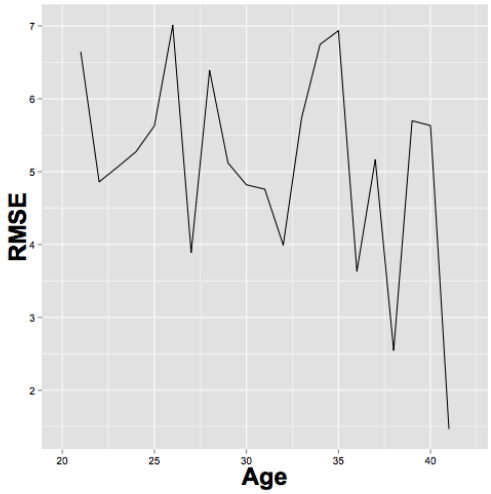


Figure 15: Predictive RMSE of the sDGLM for the full sample by age.

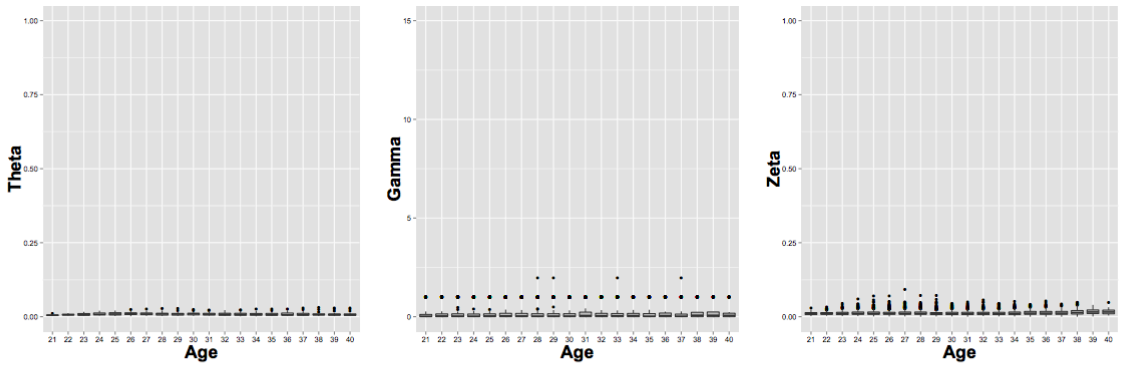
6.4. Reproducibility

One of the major concerns with any analysis utilizing MCMC is Markov chain convergence. If the chain has not converged to its stationary distribution, samples from different chains will lead to different inferences. To address the issue of reproducibility, 8 parallel MCMC simulations were run. The chains were initialized by sampling from the prior distributions for all parameters and latent variables with increased variance. For initializing $\theta_{k,t}$, the variance utilized was twice the prior variance. We believe this leads to sufficiently disperse initializations.

Figure 16a presents the maximum difference in the inference for each of the $\theta_{k,t}$. More formally, the points in the boxplot represent the set of points

$$\{\max_j\{|\hat{\theta}^j_{1,t} - \hat{\theta}^1_{1,t}|\}, \ldots, \max_j\{|\hat{\theta}^j_{k,t} - \hat{\theta}^1_{k,t}|\}, \ldots, \max_j\{|\hat{\theta}^j_{K,t} - \hat{\theta}^1_{K,t}|\}\}$$

, where $\hat{\theta}^j_{k,t}$ is the posterior mean for $\theta_{k,t}$ in the j^{th} parallel simulation. The j index takes values on $\{2, \ldots, 8\}$. All differences are computed with respect to the first initialization. Figure 16a provides convincing evidence that our MCMC based estimates of $\theta_{k,t}$ are reproducible.



(a) Θ_t (b) γ_t (c) ζ_t

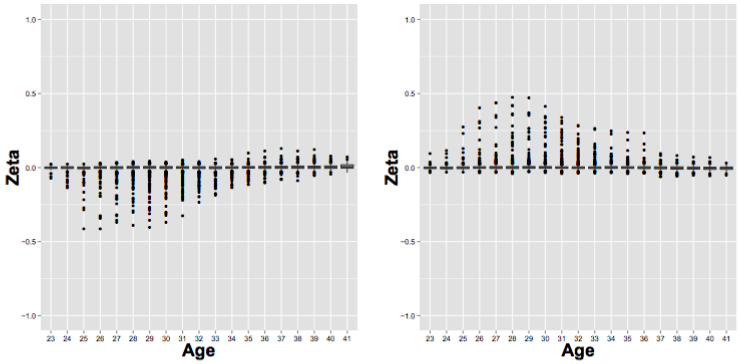
Figure 16: Reproducibility of inference from 8 parallel MCMC simulations. Left: maximum differences across 8 chains for each $\theta_{k,t}$. Middle: maximum differences across 8 chains for each player's estimate of $E[\gamma_{i,t}|y_{\cdot,1:T}]$. Right: maximum difference across 8 chains for each players estimate of $E[\zeta_{i,t}|y_{\cdot,1:T}]$.

Figures 16b and 16c consider differences across the MCMC chains at the individual player level. In these figures, each point in the distribution is the maximum difference for a single player.

In Figure 16b, the quantity of interest is $max_j|\hat{\gamma}^j_{i,t} - \hat{\gamma}^1_{i,t}|$, the maximum absolute difference in the posterior mean of a player's ability class. Each point in the boxplot corresponds to the maximum difference for a single player. In 16c, the quantity of interest is $max_j|\hat{\zeta}^j_{i,t} - \hat{\zeta}^1_{i,t}|$. Again, both figures provide convincing evidence that the eight chains generate the same set of inferences for Γ_t, and Z_t.

6.5. Sensitivity analysis

One of the limitations of our model is the necessity for the analyst to choose the number of ability classes K. For this reason, we conduct a sensitivity analysis on two different choices of K. Because the prior distribution $P(\gamma_{i,0} = k)$ and the transition kernels Q^{γ}_t are functions of K, the analysis also allows for fluctuations in the prior distributions. We compare the difference between the probability of a player's AP indicator being unity with K = 15 (baseline) against K = 13 and K = 17 classes, respectively. Figure 17a presents the boxplot of $E[\zeta_{i,t}|y_{\cdot,1:T}, K = 15] - E[\zeta_{i,t}|y_{\cdot,1:T}, K = 13]$ for player level AP status when comparing a model with K = 15 and K = 13 ability classes. While inference for a large majority of players is unchanged, there are a few outliers in which AP inference is moderately changed. Positive values correspond to cases where the probability of being flagged by the AP indicator in the K = 15 model is higher than in the model with K = 13. When outliers do occur, it is most typical that $E[\zeta_{i,t}|y_{\cdot,1:T}, K = 13] > E[\zeta_{i,t}|y_{\cdot,1:T}, K = 15]$.



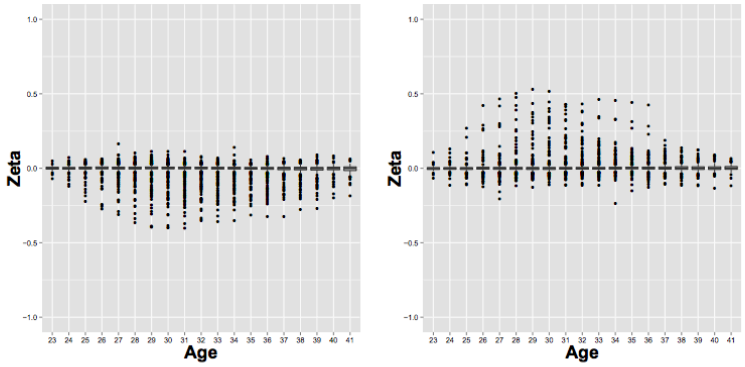
(a) $K = 13$ (b) $K = 17$

Figure 17: Sensitivity of $E[\zeta_{i,t}|y_{\cdot,1:T},]$ to choice of K. Left: $E[\zeta_{i,t}|y_{\cdot,1:T}, K = 15] - E[\zeta_{i,t}|y_{\cdot,1:T}, K = 13]$. Right: $E[\zeta_{i,t}|y_{\cdot,1:T}, K = 15] - E[\zeta_{i,t}|y_{\cdot,1:T}, K = 17]$

The opposite phenomenon is observed in the comparison with K = 17 ability classes. In Figure 17b, most estimates of the distribution for AP status are unchanged; however, it is shown that when outliers do occur, it is more common that $E[\zeta_{i,t}|y_{\cdot,1:T}, K = 17] < E[\zeta_{i,t}|y_{\cdot,1:T}, K = 15]$. As the number of ability classes increases, fewer and fewer players reach those higher clusters. As a result, the high home run hitting rates of the elite ability classes are preserved. By contrast, when there are fewer ability classes, more players are assigned to the highest ability classes, which diminishes the elite status assigned to those classes in the prior distribution. When the home run hitting rates of the highest ability classes are low, it is more likely that players require the increase in performance associated with the AP indicator to model the observed data.

The modeler's choice of the parameter α could also impact inference. The α parameter is the effective stickiness of ability class transitions. The higher α, the stickier class transitions are. To assess the sensitivity of our AP inference to α, we again compute the difference between probability of the AP flag being unity for the baseline case (α = 5) and two alternative cases with α = 4 and α = 6. The case with α = 4 corresponds to more volatility in class switching. The case with α = 6 corresponds to more autocorrelated class membership across age.

In Figures 18a and 18b, each point in the distributions across age is the difference in probability of the AP flag being unity for α = 5 and the alternative case. Note the similarity in Figures 18a and 17a. Decreasing α has a similar effect to reducing the number of ability classes. More common players reach the highest ability classes. As a result, the elite status of the highest ability classes is reduced and players are more frequently flagged for abnormal performance. A similar parallel exists between Figure 18b and 17b. Increasing α, which implies stickier class transitions, has the same effect as adding more latent classes. Fewer players reach the highest ability classes, and the elite status of those classes is preserved. As a consequence, fewer players are flagged for abnormal performance.



(a) $\alpha = 4$ (b) $\alpha = 6$

Figure 18: Sensitivity of $E[\zeta_{i,t}|y_{\cdot,1:T}]$ to choice in α. Left: $E[\zeta_{i,t}|y_{\cdot,1:T}, \alpha = 5] - E[\zeta_{i,t}|y_{\cdot,1:T}, \alpha = 4]$. Right: $E[\zeta_{i,t}|y_{\cdot,1:T}, \alpha = 5] - E[\zeta_{i,t}|y_{\cdot,1:T}, \alpha = 6]$

7. Discussion

In this paper, we have developed a model for the ability of Major League Baseball players to hit home runs at different ages. Our method borrows information both locally in age and across career trajectories of players with similar natural ability. In modeling age trajectories, we have allowed for contributions to ability which change incrementally with age and also exhibit large jumps which correspond to an abnormal increase in performance. This AP status variable allows us to identify players who have deviated sharply from the age trajectory followed by their peers of similar natural ability. While the AP status that we include in our model is simply an unexplained increase in performance, our results show that the unexplained increase aptly models the data of known PED users.

To partially disentangle a player's natural ability level and AP status, we elicited prior distributions that induce a marginal prior distribution for home run hitting ability which is consistent with physiological aging patterns. We find that our method flags performances by Mark McGwire, Alex Rodriguez, and Barry Bonds as being abnormally inflated. We demonstrate that our prior distributions have partially resolved the identifiability issue by comparing the inferences for Alex Rodriguez and Barry Bonds.

We validate this model by examining its predictive capability and find that it is competitive with methods of Jensen et al. (2009), Pecota, and Marcel. We also find that the accuracy of our predictions increases with age. Accurate predictions for players entering the prime years of their careers are important as team executives try to fairly compensate players for future and not past performance. Further, we conduct a sensitivity analysis and demonstrate that, for the vast majority of players, our inferences are robust to difference choices for the number of latent classes and the stickiness parameter governing class transitions. We find that increasing the number of ability classes has a similar effect as increasing the stickiness of the class transitions.

No statistical method is capable of completely resolving the identifiability challenge. The contribution of this work is a modeling framework that is capable of incorporating expert opinion and external sources of data to construct player specific prior distributions. Outside information could include a player's physical condition and injury information across age. It could also include results from previous drug tests or disciplinary findings. The dynamic model we develop is

amenable to intervention and change as new information arises. With the addition of outside information, estimation of the AP binary variable will be more robust. Integrating outside information into our existing framework is an important area of future work.

This method is not intended to replace drug testing programs or form the basis of disciplinary actions. It is a statistical analysis of performance data where some career trajectories are best modeled by an unexplained increase in performance. We believe this method is best suited for directing drug testing and investigative resources toward players whose performances are flagged as abnormal.

With our model, it is possible to estimate a player's career natural home run total. With the sDGLM, we can adjust home run totals for performance enhancement and compare those adjusted totals to existing historical records. Given the historical importance of career milestones in baseball, we believe this to be an important method for putting records set during the PED era in proper historical context. We leave this is future work.

References

Albert, J. (2002). Smoothing career trajectories of baseball hitters. Technical report, Department of Mathematics and Statistics, Bowling Green State University.

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

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. Biometrika, 81(3):541–553.

Fainaru-Wada, M. and Williams, L. (2007). Game of Shadows: Barry Bonds, BALCO, and the Steroids Scandal that Rocked Professional Sports. Gotham, New York, NY.

Fair, R. (2008). Estimated age effects in baseball. Journal of Quantitative Analysis in Sports, 4(1):1–41.

Ferreira, M. A. R., Holan, S. H., and Bertolde, A. I. (2011). Dynamic multiscale spatiotemporal models for gaussian areal data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(5):663–688.

Fox, E. (2009). Bayesian Nonparametric Learning of Complex Dynamical Phenomena. Ph.D. thesis, MIT, Cambridge, MA.

Fruhwirth-Schnatter, S. (2001). Fully bayesian analysis of switching gaussian state space models. Annals of the Institute of Statistical Mathematics, 53(1):31–49.

Frhwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. Journal of Time Series Analysis, 15(2):183–202.

Gaines, C. (2014). Biogenesis investigation will reveal more MLB players who used steroids. Business Insider. [Online; accessed 01/09/2016].

Gamerman, D. (1998). Markov chain monte carlo for dynamic generalised linear models. Biometrika, 85(1):215–227.

Ghahramani, Z. and Hinton, G. E. (2000). Variational learning for switching state-space models. Neural Computation, 12(4):831–864.

Jensen, S. T., McShane, B. B., and Wyner, A. J. (2009). Hierarchical Bayesian modeling of hitting performance in baseball. Bayesian Analysis, 4(4):631–652.

Jung, D. (2005). Congressional hearings on steroids in baseball. Washington Post. [Online; accessed 01/09/2016].

Kepner, T. (2010). Mcgwire admits that he used steroids. New York Times. [Online; accessed 04/15/2016].

Kim, C.-J. (1994). Dynamic linear models with markov-switching. Journal of Econometrics, 60(12):1 – 22.

Lahman, S. (2014). Lahman's baseball database. Downloaded from: http://www.seanlahman.com/baseball-archive/statistics/.

Mitchell, G. J. (2007). Report to the commissioner of baseball of an independent investigation into the illegal use of steroids and other performance enhancing substances by players in major league baseball. Technical report, DLA Piper U.S. LLP.

Nieswiadomy, M. L., Strazicich, M., and Clayton, S. (2012). Was there a structural break in Barry Bonds's bat? Journal of Quantitative Analysis in Sports, 8(3):1–19.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using polya-gamma latent variables. Journal of the American Statistical Association, 108(504):1339– 1349.

Schmotzer, B. J., Switchenko, J., and Kilgo, P. D. (2008). Did steroid use enhance the performance of the Mitchell batters? The effect of alleged performance enhancing drug use

on offensive performance from 1995 to 2007. Journal of Quantitative Analysis in Sports, 4(3):1–17.

Shumway, R. H. and Stoffer, D. S. (1991). Dynamic linear models with switching. Journal of the American Statistical Association, 86(415):763–769.

Washington Post (2011). Lawyer: Bonds didn't know he used steroids. Washington Post. [Online; accessed 04/15/2016].

Weaver, J. (2014). Alex Rodriguez's DEA confession: Yes, i used steroids from fake Miami doctor. Miami Herald. [Online; accessed 04/15/2016].

West, M. and Harrison, J. (1997). Bayesian Forecasting and Dynamic Modeling. Springer-Verlag, New York, NY, second edition.

Whiteley, N., Andrieu, C., and Doucet, A. (2010). Efficient Bayesian inference for switching state-space models using discrete particle Markov chain Monte Carlo methods. ArXiv e-prints.