Bayesian Analysis (2008) **3**, Number 2, pp. 317–338

Semi-parametric Bayesian Inference for Multi-Season Baseball Data

Fernando A. Quintana[*], Peter Müller[†], Gary L. Rosner[‡] and Mark Munsell[§]

Abstract. We analyze complete sequences of successes (hits, walks, and sacrifices) for a group of players from the American and National Leagues, collected over 4 seasons. The goal is to describe how players' performances vary from season to season. In particular, we wish to assess and compare the effect of available occasion-specific covariates over seasons. The data are binary sequences for each player and each season. We model dependence in the binary sequence by an autoregressive logistic model. The model includes lagged terms up to a fixed order. For each player and season we introduce a different set of autologistic regression coefficients, i.e., the regression coefficients are random effects that are specific to each season and player. We use a nonparametric approach to define a random effects distribution. The nonparametric model is defined as a mixture with a Dirichlet process prior for the mixing measure. The described model is justified by a representation theorem for order-k exchangeable sequences. Besides the repeated measurements for each season and player, multiple seasons within a given player define an additional level of repeated measurements. We introduce dependence at this level of repeated measurements by relating the season-specific random effects vectors in an autoregressive fashion. We ultimately conclude that while some covariates like the ERA of the opposing pitcher are always relevant, others like an indicator for the game being into the seventh inning may be significant only for certain seasons, and some others, like the score of the game, can safely be ignored.

Keywords: Dirichlet Process, Partial Exchangeability, Semiparametric Random Effects

1 Introduction

Albright (1993) discusses a data set of the entire sequences of successes for baseball players from the American and National Leagues, over the 4 seasons spanning the years from 1987 to 1990. Albright defines a success in terms of either getting on base or moving players along the bases. Thus, hit, walk, or sacrifice constitute a success, and we follow this definition. The data set contains for each player, season, and at-bat

[*] Departamento de Estadística, Pontificia Universidad Católica de Chile, Chile, mailto:quintana@mat.puc.cl

[†] Department of Biostatistics, The University of Texas, M. D. Anderson Cancer Center, mailto:pmueller@mdanderson.org

[‡] Department of Biostatistics, The University of Texas, M. D. Anderson Cancer Center, mailto:glrosner@mdanderson.org

[§] Department of Biostatistics, The University of Texas, M. D. Anderson Cancer Center, mailto:mfmunsell@mdanderson.org

DOI:10.1214/08-BA312

occasion, a binary outcome, defined as $y_{ijk} = 1$ if a success occurred and $y_{ijk} = 0$ otherwise. Here, k indexes at-bat occasions, j indexes seasons, and i denotes players, where $k = 1, \ldots, n_{ij}$; $j = 1, \ldots, n_i$; and $i = 1, \ldots, n$. The data set also includes 11 occasion-specific covariates for each binary outcome's at-bat appearance that possibly affect the success probability. For each at-bat appearance, the covariate *I7* equals 1 if the occasion occurs in the 7th inning or later and 0 otherwise; *O2* is 1 if there are 2 outs and 0 otherwise; *Score* equals the number of runs that separate the batter's team and the opposing team (positive if the batter's team is ahead and negative if behind); *R123* is 1 if any runners are on base when the player is at bat and 0 otherwise; *R23* is 1 if any runners are on 2nd or 3rd base and 0 otherwise; *Game*, corresponding to a chronological index of the game number; *DN* equals 1 for a night game and 0 for a day game; *HA* is 1 if the game is played at home and 0 if it is an away game; *T* is defined as 1 if the opposing pitcher is right-handed and 0 if left-handed; *ERA* equals the opposing pitcher's earned run average for that season; and *Turf* equals 1 if the field is natural grass and 0 if artificial turf.

Common sense suggests that the *ERA*, a measure of the opposing pitcher's ability to hold down the opposing team's batters, should have substantial impact on the hitting probability. Since ERA is likely to be an important covariate, we are careful about modeling its effect. We use three dummy variables *ERA1* through *ERA3* to allow for non-linear effects. Letting $q_1 < q_2 < q_3$ denote the quartiles of the observed *ERA* values in a given season, we define $ERA1 = I\{ERA \le q_1\}$, $ERA2 = I\{q_1 < ERA \le q_2\}$ and $ERA3 = I\{q_2 < ERA \le q_3\}$. We denote the 13-dimensional vector of occasion-specific covariates as $\boldsymbol{x}_{ijk}$.

One of the key issues that motivated the analysis in Albright (1993) was the question whether players exhibit streakiness in their hitting patterns. A related question is whether this streakiness was permanent or just limited to a given season. The assessment of streakiness for this dataset was previously discussed in a number of articles. In particular, we build on Quintana and Müller (2004) who concluded, based on data from the 1990 season only, that a first-order Markovian dependence was appropriate to model streakiness.

In this paper, we focus on inference about the evolution of players over seasons. In particular, we want to study how streakiness and covariate effects change over seasons. This includes questions such as "Does the home field advantage effect on batting success probabilities change over seasons?" Addressing such questions requires a model that takes into account not only the longitudinal nature of the binary responses within a season but also the dependence across seasons that arises from observing the same player across multiple seasons.

Because we are specifically interested in serial dependence across seasons, we focus only on players with available data for consecutive seasons. We find $n = 76$ players with data across consecutive seasons. Table 1 lists the specific players. Of these 76 players, 27 have data recorded for the first two seasons ($n_i = 2$), 14 have data for seasons 1 through 3 ($n_i = 3$), and for 35 players data are reported for all seasons ($n_i = 4$). Finally, the available data only includes "regular players", i.e., those who were at least 500 times at-bat during a given year, so that $n_{ij} \geq 500$.

The main features of the proposed model are an autologistic model for the binary outcomes across at-bat occasions, a non-parametric prior for the random effects distribution of season-specific success probabilities, and an autoregressive dependence structure across seasons. We use the term "nonparametric" to refer to models that cannot be defined in terms of finite-dimensional parameters. Nonparametric random effects models have been successfully applied in various contexts. Models with similar nonparametric priors have been used, among others, in Bush and MacEachern (1996), Müller and Rosner (1997), Mukhopadhyay and Gelfand (1997) and Kleinman and Ibrahim (1998). Bush and MacEachern (1996) was the first paper to introduce the now commonly used semiparametric modeling approach with parametric priors for fixed effects and nonparametric model components for random effects. For the particular case of binary data, related models appear in Liu (1996) and Basu and Mukhopadhyay (2000), among many others. We chose a nonparametric random effects distribution because we were concerned that a parametric model, such as a traditional multivariate normal random effects model, would assume too much homogeneity for the population of players. Technically, the proposed nonparametric model can be described as a mixture of normal models. The model preserves many of the computational advantages of a parametric model. But by introducing a mixture the model allows us to learn about heterogeneity in the population. Another important motivation for the nonparametric model choice is a representation theorem for partially exchangeable sequences. If we believe that the probability model for a binary sequence should depend only on the number of order-ℓ transitions, then it can be argued that the model should include a nonparametric random effects distribution for season-specific success probabilities. In this sense the nonparametric model allows us to report inference without the limitation of a specific assumed sampling model, beyond the general notion of order-ℓ exchangeability. We provide further details in Section 2.

In our application we base the nonparametric specification on the popular Dirichlet process (DP) (Ferguson 1973). The choice of the DP prior is mainly driven by the simplicity of the resulting posterior simulation schemes. Recent reviews of semi- and non-parametric Bayesian models can be found in Walker et al. (1999) and in Müller and Quintana (2004). Another key feature of the proposed model is the use of an autoregression to define the dependence of random effects corresponding to the same player across different seasons. An implication of this modelling choice is that the marginal models for random effects within a season remain nonparametric, while dependence across seasons is modeled parsimoniously by the autoregression. Another feature is that the model implies increased prior uncertainty for random effects for later seasons corresponding to a given player. This simply reflects the fact that many changes usually take place from one season to other: players change teams, some retire and some minor-leaguers enter the major leagues, some get injured, etc.

The rest of this article is organized as follows. In section 2, we describe the main features of the proposed model, emphasizing the two levels of dependence. A Markov chain Monte Carlo (MCMC) simulation approach for posterior inference is briefly described in Section 3. Section 4 summarizes the findings when fitting the model to the baseball dataset described earlier. In particular, we answer the motivating and other questions. Finally, Section 5 presents a discussion and extensions of our analysis.

2 A Model for Batting Performance

Recall that y_{ijk} denotes the binary outcome (hit, walk or sacrifice) recorded at the kth time that player i was at bat during the j-th season, and that $\boldsymbol{x}_{ijk}$ corresponds to the 13-dimensional vector of covariates, where $k = 1, \ldots, n_{ij}$, $j = 1, \ldots, n_i$, and $i = 1, \ldots, n = 76$. Denote the entire binary sequence for player i during season j as $\boldsymbol{y}_{ij} = (y_{ijk},\ 1 \leq k \leq n_{ij})$, and the complete collection of responses as $\boldsymbol{y} = (\boldsymbol{y}_{ij},\ 1 \leq j \leq n_i,\ 1 \leq i \leq n)$. We model the longitudinal sequence $\boldsymbol{y}_{ij}$ by a Markov model. The Markov chain has a binary state space $y_{ijk} \in \{0, 1\}$. The Markov chain is defined by specifying the transition probabilities from y_{ijk} to $y_{ij,k+1}$. We use an autologistic regression model to do this. The regression includes lagged responses and covariates $\boldsymbol{x}_{ijk}$, following the approach in Quintana and Müller (2004). They proposed to model the $\boldsymbol{y}_{ij}$ sequences as mixtures of Markov chains of a certain order ℓ. The mixture is defined with respect to the transition probabilities and includes a nonparametric prior on the mixing measure. The model is inspired by the notion of partial exchangeability (Quintana and Newton 1998). A probability model for a binary sequence is partially exchangeable of order ℓ if it is invariant under any permutation that leaves the initial portion of the sequence and the transition counts up to order ℓ unaltered. It has been shown (Freedman 1962a,b; Quintana and Newton 1998) that such invariance plus some technical conditions imply that the joint distribution $p(\boldsymbol{y}_{ij})$ can be written as a mixture of Markov chains. For order-ℓ models, the Markov chain is characterized by a transition matrix of dimension $2^\ell \times 2^\ell$, which can be parameterized by only 2^ℓ transition probabilities.

We propose a more parsimonious and restricted version of the general representation of a partially exchangeable model. We assume

$$\text{logit}\left[P(y_{ijk} = 1 \mid \ell, \boldsymbol{\theta}_{ij}, y_{ij,k-\ell}, \ldots, y_{ij,k-1})\right] = \theta_{ij0} + \theta_{ij1} y_{ij,k-1} + \cdots + \theta_{ij\ell} y_{ij,k-\ell} + \boldsymbol{x}'_{ijk} \boldsymbol{\beta}_j, \tag{1}$$

with $\boldsymbol{\theta}_{ij} = (\theta_{ij0}, \ldots, \theta_{ij\ell})$, for a fixed value of ℓ. The $\boldsymbol{\theta}_{ij}$ parameters are the autologistic regression coefficients, with θ_{ij0} being the intercept and $\theta_{ij1}, \ldots, \theta_{ij\ell}$ being the coefficients of the lagged responses $y_{ij,k-1}, \ldots, y_{ij,k-\ell}$. This $(\ell + 1)$-dimensional Markovian representation is equivalent to the full model with 2^ℓ parameters for the cases $\ell = 0$ and $\ell = 1$. Note also that the covariates enter model (1) linearly in the logit scale. The 13-dimensional coefficient vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_4$ are season specific and represent the global effect of occasion-specific covariates on the success probability, expressed on the logit scale. Posterior inference on $\boldsymbol{\beta}_j$ will allow us to formally address questions about how the effect of covariates, such as home field advantage, changes over seasons. Inference about the change of streakiness across seasons is achieved by comparing $\boldsymbol{\theta}_{ij}$ across seasons j.

A model similar to (1), with fully parametric priors and without reference to order-ℓ partial exchangeability was used in Erkanli et al. (2001). To define ℓ we recall the analysis that Quintana and Müller (2004) carried out using data from the 1990 season only. They concluded that an order-1 model was appropriate. It is not clear that this conclusion can be extrapolated to data over four seasons. Instead we make a conservative choice and use $\ell = 5$.

The next stage in our model building is the definition of a probability model for the random effects $\boldsymbol{\theta}_{ij}$. We proceed by defining a model for the first season and then extend this to subsequent seasons. A conventional and technically convenient choice for a random effects distribution in a model like (1) would be a multivariate normal random effects distribution for $\boldsymbol{\theta}_{i1}$. However, as part of our substantial prior information we believe that the player population is not homogeneous. We wish to explicitly allow for different sub-populations being characterized by different levels and different kinds of streakiness. Such heterogeneity is well represented by mixture models. Also recall the earlier mentioned representation of partially exchangeable random binary sequences as mixtures of Markov chains. These two considerations lead us to use a nonparametric random effects distribution that takes the form of a mixture of normal models. Starting with the 1987 season, we define a nonparametric random effects model as a mixture of normal distributions. The mixture is with respect to the normal location parameter. The model becomes nonparametric by assuming a nonparametric prior on the mixing measure.

$$\boldsymbol{\theta}_{i1} \overset{\text{ind}}{\sim} \int N(\boldsymbol{\theta}_{i1};\ \boldsymbol{\mu}_{i1}, \boldsymbol{S})\, \mathrm{d}F(\boldsymbol{\mu}_{i1}) \qquad \text{and} \qquad F \sim \mathcal{D}(M, F_0), \tag{2}$$

where $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{S})$ denotes the multivariate normal distribution on $\boldsymbol{x}$, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{S}$, and $\mathcal{D}(M, F_0)$ denotes the Dirichlet process (DP) with baseline distribution F_0 and total mass parameter M (Ferguson 1973). The DP defines a random probability measure (RPM), that is, a distribution on the space of distributions.

We briefly summarize some key features of the DP model that are helpful for understanding the nature of the proposed model. Actual implementation, as described in Section 3, will not make use of the nonparametric model itself. Technically, the implementation of posterior inference will be based on model (2) with the random probability measure F integrated out, i.e., we will only have to manipulate the resulting finite dimensional probability model on $(\boldsymbol{\theta}_{i1},\ i = 1, \ldots, n)$. Sethuraman (1994) shows that a random probability F generated from a DP prior can be written as follows. Consider any event B. Then $F(B) = \sum_{h=1}^{\infty} w_h \delta_{U_h}(B)$, where $U_1, U_2, \ldots$ are a random sample from F_0 and the weights w_h are defined as $w_1 = V_1$ and $w_h = \prod_{j=1}^{h-1}(1 - V_j)V_h$ for $h \geq 2$, with $V_1, V_2, \ldots$ a random sample from the $\text{Beta}(1, M)$ distribution. In summary, the representation is in terms of an infinite mixture of point-masses with locations generated from F_0 and the weights determined by M. An immediate consequence of this representation is that the DP is almost surely discrete.

Model (2) defines a mixture of a normal kernels with respect to the normal location parameter. The mixing measure has a DP prior. The resulting continuous distribution is known as a DP mixture (Antoniak 1974). The introduction of latent variables $\boldsymbol{\mu}_i$ allows us to write (2) as a hierarchical model. Also, we introduce a modification to the semi-parametric mixture model (2) to facilitate the inclusion of player-specific prior information.

The logistic intercept θ_{ij0} is special. It represents the marginal probability of a success for player i in season j. We characterized the prior for each player's marginal "ability" for the 1987 season with a normal distribution based on the player's own 1986 season averages. We inflate the prior variances by a factor 10 to allow for additional prior uncertainty. The averages included walks and sacrifices as "successes" to agree with the data we analyzed. One player, Ellis Burks, began his major league career in 1987. For Burks's normal prior, we computed a predictive distribution for a rookie in the 1986 season and used the first two moments from the predictive distribution. We determined the predictive distribution based on a hierarchical model, analyzing the data for all rookies whose major league careers started in 1986. Since several rookies in 1986 had no successes, we used empirical logits for each rookies' data (that is, we added 0.5 to the number of successes and failures before taking the logarithm), with the variance suggested by Gart (1966). All 1986 data came from version 5.4 of Sean Lahman's baseball database (www.baseball1.com/statistics/). The resulting player-specific prior means and variances are shown in Table 1.

Let $\mu_{i10} \sim N(m_{i0}, V_{i0})$ denote the historical prior on the ability of player i based on the 1986 season. Also, let $\boldsymbol{\mu}_{i1}^{-} = (\boldsymbol{\mu}_{i11}, \ldots, \boldsymbol{\mu}_{i1q})$ denote $\boldsymbol{\mu}_{i1}$ with the intercept $\boldsymbol{\mu}_{i10}$ removed (recall that $q = 5$ in our case). Writing the modified mixture model (2) as a hierarchical model we get:

$$\boldsymbol{\theta}_{i1}|\boldsymbol{\mu}_{i1} \overset{\text{ind}}{\sim} N(\boldsymbol{\mu}_{i1}, \boldsymbol{S}), \quad \boldsymbol{\mu}_{i1}^{-}|F, \boldsymbol{\phi} \overset{\text{ind}}{\sim} F, \quad \boldsymbol{\mu}_{i10} \sim N(m_{i0}, V_{i0}), \quad F|\boldsymbol{\phi} \sim \mathcal{D}(M, F_0(\boldsymbol{\phi})), \tag{3}$$

where $\boldsymbol{\phi}$ is an optional vector of hyperparameters for the baseline distribution F_0.

Next we extend the model to seasons $j = 2, \ldots, n_i$. The sampling model is already defined in (1). We still need to define a random effects distribution for $\boldsymbol{\theta}_{ij}$. As in (3) we introduce latent parameters $\boldsymbol{\mu}_{ij}$ and define a normal prior, $\boldsymbol{\theta}_{ij} \sim N(\boldsymbol{\theta}_{ij};\ \boldsymbol{\mu}_{ij}, \boldsymbol{S})$. Completing the prior model for $\boldsymbol{\mu}_{ij}$ we have three goals in mind. (i) We want to explicitly relate seasons for a given player. This is important to address the desired inference about changes in streakiness across seasons. (ii) We want a flexible marginal model for random effects from each season, similar to (2) or (3). And, (iii) we want to reflect the increased prior uncertainty in the random effects for a given player as seasons progress (recall the earlier discussion). We achieve these goals by considering an autoregressive relationship among latent parameters:

$$\boldsymbol{\mu}_{ij} = \boldsymbol{\alpha}_0 + D(\boldsymbol{\alpha}_1)\boldsymbol{\mu}_{i,j-1} + \boldsymbol{\epsilon}_{ij}, \qquad j > 1, \tag{4}$$

where $\{\boldsymbol{\epsilon}_{ij}\}$ are independent normal residuals with covariance matrix $\boldsymbol{R}$, $D(\boldsymbol{a})$ is a diagonal matrix with the elements of $\boldsymbol{a}$ as diagonal entries, and F is a DP random measure, as before. In words, the model specifies that the player and season-specific parameters that characterize success probability and a hot hand are related across seasons by an autoregressive structure. The autoregressive model is initialized in season $j = 1$ with a nonparametric prior, i.e., without strict parametric assumptions. The autoregressive coefficients α_{1k} quantify the strength of the correlation across seasons.

Random effects $\boldsymbol{\theta}_{ij}$ for $j = 1, \ldots, n_i$ are thus related because the corresponding latent $\boldsymbol{\mu}_{ij}$ parameters are stochastically dependent. Also, the nonparametric nature of the model for the latent $\boldsymbol{\mu}_{i1}$'s and the autoregressive specification (4) imply a nonparametric model for the marginal distribution of $\boldsymbol{\mu}_{ij}$, $j \geq 2$. As a consequence, random effects $\boldsymbol{\theta}_{ij}$ have marginally an induced mixture model with a DP prior, as in (2). Only the normal kernel $N(\boldsymbol{\mu}_{i1}, \boldsymbol{S})$ in (2) is replaced by the normal distribution implied by the convolution of the $N(\boldsymbol{\mu}_{i1}, \boldsymbol{S})$ distribution with the normal autoregressive conditional distributions $p(\boldsymbol{\mu}_{ij} \mid \boldsymbol{\mu}_{i,j-1})$.

The final step in the model construction is a prior probability model for all the remaining hyperparameters, including $\boldsymbol{\phi}$. For convenience we choose the baseline distribution F_0 in (2) as $F_0(x; \boldsymbol{\phi}) = N(x; \boldsymbol{\phi})$, where $\boldsymbol{\phi} = (\boldsymbol{m}, \boldsymbol{V})$. Denote by $\boldsymbol{a} = (\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$ a stacked version of the autoregression coefficients. We assume a normal prior $p(\boldsymbol{a} \mid \boldsymbol{a}_0, \boldsymbol{A}) = N(\boldsymbol{a}; \boldsymbol{a}_0, \boldsymbol{A})$. For $\boldsymbol{\phi}$ we choose another normal-inverse Wishart prior $p(\boldsymbol{m}, \boldsymbol{V}) = N(\boldsymbol{m}; \boldsymbol{m}_0, \boldsymbol{V}) \times IW(\boldsymbol{V}; \boldsymbol{V}_0, \nu_V)$. Finally, for $\boldsymbol{S}$ we assume $p(\boldsymbol{S}) = IW(\boldsymbol{S}; \boldsymbol{S}_0, \nu_S)$. See Section 4 for specific choices of M, $\boldsymbol{a}_0$, $\boldsymbol{A}$, ν_V, $\boldsymbol{m}_0$, $\boldsymbol{V}_0$, ν_S and $\boldsymbol{S}_0$. For later reference, we write the whole vector of hyperparameters as $\boldsymbol{\eta} = (\boldsymbol{\phi}, \boldsymbol{a}, \boldsymbol{S})$.

The model is completed with a prior probability model for $\boldsymbol{\beta}_j$, the covariate effects in season j. We use a similar AR prior as in (4):

$$\boldsymbol{\beta}_1 \sim N(\mathbf{0}, \boldsymbol{B}) \qquad \text{and} \qquad p(\boldsymbol{\beta}_j \mid \boldsymbol{\beta}_{j-1}) \sim N(\boldsymbol{\beta}_{j-1}, c^2 \boldsymbol{B}), \tag{5}$$

where c is a scalar factor that controls the amount of smoothing across seasons and $\boldsymbol{B}$ is a hyperparameter that specifies the a priori correlation of the covariate effects.

3 Posterior Simulation Algorithm

The proposed model includes conditional independence at several levels. As a consequence many complete conditional distributions are easily recognized and allow efficient random variate generation. This allows us to define an efficient Gibbs sampling algorithm for posterior simulation.

Posterior simulation proceeds after analytically marginalizing model (3) with respect to the random probability measure F. The resulting posterior distribution for $\boldsymbol{\theta}_{ij}$ allows an efficient Gibbs sampling implementation. We describe the main steps, and refer the interested reader to the cited references for a detailed description. The discrete nature of the random probability measure F in (3) implies a positive probability for ties among the $\boldsymbol{\mu}_{i1}$ parameters. Denote by $k \equiv k(n)$ the number of distinct values among the components of $\boldsymbol{\mu}_1 = (\boldsymbol{\mu}_{11}, \ldots, \boldsymbol{\mu}_{n1})$. Let the unique values (or *locations*) be denoted as $\boldsymbol{\mu}_1^* = (\boldsymbol{\mu}_{11}^*, \ldots, \boldsymbol{\mu}_{k1}^*)$. Define *membership* indicators $\boldsymbol{s} = (s_1, \ldots, s_n)$ as $s_i = j$ if $\boldsymbol{\mu}_{s_i 1} = \boldsymbol{\mu}_{j1}^*$. The sets of players sharing a common location can be interpreted as *clusters*, describing groups of players with similar behavior. It is then convenient to represent $\boldsymbol{\mu}_1$ as $(\boldsymbol{\mu}_1^*, \boldsymbol{s})$. Updating $\boldsymbol{\mu}_1$ is carried out by updating $\boldsymbol{\mu}_1^*$ and $\boldsymbol{s}$, conditional on all the other quantities. This defines the perhaps simplest Gibbs sampling scheme for DP mixture models (Bush and MacEachern 1996; MacEachern and Müller 1998). Our implementation is based on these algorithms. Alternative approaches based on Metropolis-Hastings moves are discussed in Neal (2000), Dahl (2003) and Jain and Neal (2004). The most time-consuming step in the posterior simulation is the updating of membership indicators $\boldsymbol{s}$. Luckily, the normal distribution assumption $\boldsymbol{\theta}_{i1}|\boldsymbol{\mu}_{i1} \overset{\text{ind}}{\sim} N(\boldsymbol{\mu}_{i1}, \boldsymbol{S})$ together with the conjugate normal specification of F_0 allows us to analytically integrate out $\boldsymbol{\mu}^*_{j1}$. This allows for an efficient implementation, as described in MacEachern and Müller (1998, 2000).

Once a new value for $\boldsymbol{s}$ is imputed, the $\boldsymbol{\mu}^\star_{i1}$ are updated from the corresponding full conditional given the new $\boldsymbol{s}$ and all other parameters. This reduces, for each imputed cluster, to a simple parametric model with prior F_0 and restricted to players sharing that cluster. The remaining $\boldsymbol{\mu}_{ij}$ vectors for $2 \leq j \leq n_i$ can be updated one by one from their respective full conditional posterior distributions, all of normal type. By the AR(1) assumption (4), these conditional posterior distributions depend on the latent parameters $(\boldsymbol{\mu}_{i,j-1}, \boldsymbol{\mu}_{i,j+1})$ for the previous and next seasons, respectively (except for the last season $j = n_i$). The $\boldsymbol{\theta}_{ij}$ parameters are updated one at a time by drawing from the corresponding logistic-normal distribution, as described in, e.g. Carlin and Louis (1996).

Updating the autoregressive coefficients $\boldsymbol{\alpha}$ is easily implemented by noting that (4) is linear in $\boldsymbol{\alpha}$, so that the normal prior assumption implies a normal conditional posterior distribution. The remaining parameters, namely $\boldsymbol{m}$, $\boldsymbol{V}$ and $\boldsymbol{S}$ can be updated straightforwardly, given the conjugate-style prior assumptions.

4 Results

Recall the assumption of an order of dependence $\ell = 5$. Thus, each of the $\boldsymbol{\theta}_{ij}$ and $\boldsymbol{\mu}_{ij}$ vectors is of dimension 6. The θ_{ij0} coefficients define the probability of success (in the logit scale) when all covariates are 0 and the player suffered a streak of misses of length (at least) 5. In turn, θ_{ijm} for $m = 1, \ldots, 5$ are the regression coefficients for the binary lagged responses. The regression coefficients $\boldsymbol{\beta}_j$ are season-specific and are the primary focus of interest in our analysis.

For the fixed hyperparameters, we use $\boldsymbol{m}_0 = \boldsymbol{0}$, $\boldsymbol{a}_0 = \boldsymbol{0}$, $\nu_S = 8$, $\nu_V = 8$, $\boldsymbol{A} = 0.25\boldsymbol{I}_{12}$, and $\boldsymbol{V}_0$ and $\boldsymbol{S}_0$ such that the resulting prior means for $\boldsymbol{V}$ and $\boldsymbol{S}$ are both equal to $0.25\boldsymbol{I}_6$, where $\boldsymbol{I}_m$ is the identity matrix of dimension $m \times m$. For the prior on the covariate effects we choose $\boldsymbol{B} = \boldsymbol{I}$ and $c = 0.1$. We also choose $M = 1$, which implies a priori for the expected number of clusters, $E(k(n)) \approx M \log((M+n)/M) = 4.34$, and for the corresponding variance, $\text{Var}(k(n)) \approx M\{\log((M+n)/M) - 1\} = 3.34$ (Liu 1996).

Figure 1 summarizes the marginal posterior distributions for the autoregressive coefficients $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ in (4). Most of the posterior mass for α_{0m} is concentrated away from 0 for $m = 0, \ldots, 3$. In each season a player is subject to the same overall probability of streakiness, characterized by the positive values for $(\alpha_{01}, \alpha_{02}, \alpha_{03})$. The first three lagged at-bats appear to be globally important.

The posterior distribution of α_{10} is bounded away from zero, unlike the posteriors of

α_{1m}, $m \geq 1$. Thus, only the marginal probability of getting on base (i.e., a success) is significantly correlated across seasons for a player. The strength of the autocorrelation (i.e., the propensity to streaky behavior), however, is not. There is no evidence even in this extensive data set that streakiness of a player persists across seasons. In other words, based on the near-zero values of $(\alpha_{11}, \alpha_{12}, \alpha_{13})$ there is no evidence that a player's streakiness is correlated across seasons. A possible explanation is simply the fact that approximately 6 months separate one season from the next one.

Looking at individual players, we find substantial variability. This variability is illustrated in Figures 2 through 4, which summarize the entire collection of marginal posterior distributions for the $\boldsymbol{\theta}_{ij}$ parameters. We observe some substantial variability in the evolution of those coefficients over seasons. Take, for instance, players 7 (Barry Bonds) and 76 (Robin Yount). Comparing the evolution of the respective intercept coefficients θ_{ij0}, we see quite the opposite behavior for these two players. For Bonds, we see mostly increasing values over the 1987 to 1990 period, while for Yount there is a decreasing pattern. It is interesting to note that the history of official batting averages (defined as the ratio of hits to official at bats) for 1987-1990 was .312, .306, .318, and .247 for Yount, and .261, .283, .248 and .301 for Bonds (available from www.espn.com). The posterior means of the θ_{ij0} coefficients mostly follow the same trends, although batting averages are not a one-to-one function of θ_{ij0}, because we use a broader definition of success and include more appearances (walks and sacrifices) in the denominator than official "at bats." Considering Barry Bonds' record, we note that he had increasingly more intentional walks over seasons. This helps to explain the increasing pattern. For Robin Yount, the 1990 season was already his 17th season; He ended his career in 1993. We find increasingly fewer at-bats across the four seasons (635 through 587). Other, more zigzagging patterns are also found, e.g., for player 5, George Bell. Also, the posterior variances for all coefficients tend to increase over seasons, in agreement with the notion of increased prior uncertainty for random effects over seasons. Another cause for the increasing posterior variances is the fact that not all players have data on all seasons, as noted earlier. The number of effective data points for season 4 is less than half of what is available for season 1.

Figures 5, 6 and 7 show marginal posterior summaries of the regression coefficients for all thirteen covariates. The variation in regression coefficients β_j across seasons is natural. Players change teams, retire, have injuries, etc. All seasons are different. The importance of these covariates is uneven. The results suggest that knowing the score of the game (*Score*) when the player gets up to bat seems to make little difference on the success probability. The handedness of the pitcher (T) did not appear to matter. This lack of an effect may reflect the need to know the handedness of the pitcher relative to how the batter is batting. Conventional wisdom is that right-handed batters hit better against left-handed pitchers and vice versa. The data set does not contain this information, though.

In contrast, knowing whether there are already two outs (*O2*), runners on base (*R23* and *R123*), whether the game is played in the batter's home field (home-field advantage *HA*), and the magnitude of the earned run average (*ERAj*, $j = 1, 2, 3$) of the opposing pitcher are always relevant. *O2* is a stress covariate in the sense that the hitter is in a limit situation; thus the posterior mass for the corresponding coefficient is mostly concentrated on negative numbers. Also, with two outs, a hitter can not try for a sacrifice or may hit into a fielder's choice situation, which may contribute to the lower success probability when batting with two outs.

Having a runner already on base has a positive effect on the probability of success, as shown by the positive effects for *R23* and *R123*. The larger positive effect of *R23* may have a twofold explanation: first, it reflects the possibility of intentional walks, which count as successes; and second, getting a hit when there is a player on second or third base almost always gets a runner home. Having players on base may also lead to sacrifices on the part of the batter in the attempt to move the player already on base closer to home. Additionally, a pitcher may be distracted trying to keep a runner from stealing and allow a batter to hit or, perhaps, walk.

The home-field advantage (*HA*) is also an important psychological stimulus for the player, who might be trying to please the usually eager fans. Also, a player plays half their games at home. Therefore, they are much more familiar with their home field than any opposing field. Opposing fielders are less familiar with the field as well. These 2 factors may also help explain HA. The *ERA* is a measure of the pitcher's quality, with higher values indicating a less skilled player. Our parameterization has the highest quartile pitchers as the baseline group, so better pitchers (i.e., those with lower ERAs) should have a negative effect on the batter's log-odds of success. Indeed, most of the posterior mass for the coefficients of all three ERA variables is concentrated on negative numbers, which is natural since $ERA1 = ERA2 = ERA3 = 0$ represents the weakest pitchers. The reported posterior means for the dummy variables ERA1, ERA2 and ERA3 show a trend (success probabilities associated with ERA1 being lower than those when facing pitchers in the second quartile, which are lower than associated with ERA3). Additionally, the posterior means suggest the presence of a nonlinear effect of ERA on the hitting probabilities, with the greatest between-ERA-quartile effect on the batter's success appearing when facing a pitcher in the third quartile, relative to the fourth.

The other covariates do not appear to be important, as the corresponding posterior means are at most one standard deviation away from zero. In other words, in all these cases there is at least 15% to 20% probability to each side of zero.

The list of covariates that are reported in the data from Albright (1993) excludes some potentially important explanatory variables. These include handedness of the hitter, the ball park, batting average against the pitcher rather than the pitcher's ERA, characterization of the ball park as a hitter's park versus pitcher's park, team affiliation, and career stage. In principle an interested reader may attempt the difficult task of collecting additional occasion-specific information. For simplicity in this article, we restricted inference to the data set reported in Albright (1993). Besides the choice of covariates, the data set covers a relatively short time span of the seasons 1987–1990. However, we note that four years is reasonable, considering the length of time a player stays with a team. Rather than changing the data set, which we fear could compromise the pre-processing and selection chosen by Albright, we opted to incorporate some additional information on each player by means of informative priors.

Finally, we also explored sensitivity of our results to some modeling and hyperparameter choices. Specifically, we considered increasing hyperparameter M in the DP prior to something very large, practically ∞. In addition, given that the lagged terms of order 4 and 5 did not seem to have much impact on the inferences, we considered the same model but decreasing ℓ from 5 to 3. Figure 1 contains the posterior means of the $\boldsymbol{\alpha}$ coefficients for the proposed and the two alternative models. Similar results were obtained for the regression coefficients (not shown). No significant changes are found in these plots, which suggests robustness of the results to our particular choices.

5 Conclusion

We have analyzed the performance of batters over the course of 4 seasons, using a semiparametric Bayesian random effects model. The distinctive feature of the model is the incorporation of two levels of dependence, reflecting the nature of the data as nested repeated measurements. While one level of repetition (at bats within a season) was modelled using a Markovian-style autologistic regression, for the other (seasons) we used an AR(1) model linking latent parameters specific to each season and player.

From the analysis we ultimately concluded that the model captured well some of the main aspects of the problem. Also, by looking at the entire collection of posterior distributions $p(\theta_{ijk} \mid \boldsymbol{y})$ (not shown) we can conclude that there is substantial variability among players in terms of streakiness and also for players across seasons. For instance, for player 17 (Alvin Davis) a number of lags from various seasons appear to be important (i.e. the corresponding posterior has significant portion of probability mass to one side of zero), while for Bobby Bonilla (player 8), only the second lag from season 2 stands out. This suggests that streakiness may also be related to seasons.

A somewhat similar conclusion is reached regarding the available occasion-specific covariates. Some are important in all seasons, some are important only for specific seasons, and some are never relevant so that they may be safely discarded from the analysis. Among the latter, it is somewhat surprising to observe that the handedness of the pitcher does not give significant information, a fact that may appear to contradict conventional wisdom. This was partly explained by the fact that the data set did not include information on the batter's handedness relative to the pitcher.

References

Albright, S. C. (1993). "A statistical analysis of hitting streaks in baseball (with Discussion and a reply from the author)." *Journal of the American Statistical Association*, 88: 1175–1196. 317, 318

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The Annals of Statistics*, 2: 1152–1174. 321

Basu, S. and Mukhopadhyay, S. (2000). "Bayesian analysis of binary regression using symmetric and asymmetric links." *Sankhyā*, 62: 372–387. 319

Bush, C. A. and MacEachern, S. N. (1996). "A semiparametric Bayesian model for randomised block designs." *Biometrika*, 83(2): 275–285. 319, 323

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall. 324

Dahl, D. B. (2003). "An improved merge-split sampler for conjugate Dirichlet Process mixture models." Technical Report 1086, Department of Statistics, University of Wisconsin. 324

Erkanli, A., Soyer, R., and Angold, A. (2001). "Bayesian Analyses of Longitudinal Binary Data Using Markov Regression Models of Unknown Order." *Statistics in Medicine*, 20(5): 755–770. 320

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1: 209–230. 319, 321

Freedman, D. A. (1962a). "Invariants under mixing which generalize de Finetti's theorem." *Annals of Mathematical Statistics*, 33: 916–923. 320

— (1962b). "Mixtures of Markov processes." *Annals of Mathematical Statistics*, 33: 114–118. 320

Gart, J. J. (1966). "Alternative analyses of contingency tables." *Journal of the Royal Statistical Society. Series B. Methodological*, 28: 164–179. 322

Jain, S. and Neal, R. M. (2004). "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model." *Journal of Computational and Graphical Statistics*, 13(1): 158–182. 324

Kleinman, K. and Ibrahim, J. (1998). "A Semi-parametric Bayesian Approach to the Random Effects Model." *Biometrics*, 54: 921–938. 319

Liu, J. S. (1996). "Nonparametric Hierarchical Bayes via Sequential Imputations." *The Annals of Statistics*, 24(3): 911–930. 319, 324

MacEachern, S. N. and Müller, P. (1998). "Estimating Mixture of Dirichlet Process Models." *Journal of Computational and Graphical Statistics*, 7(2): 223–338. 323, 324

— (2000). "Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichlet Process Mixture Models." In Ruggeri, F. and Ríos-Insua, D. (eds.), *Robust Bayesian Analysis*, 295–316. New York: Springer-Verlag. 324

Mukhopadhyay, S. and Gelfand, A. E. (1997). "Dirichlet Process Mixed Generalized Linear Models." *Journal of the American Statistical Association*, 92: 633–639. 319

Müller, P. and Quintana, F. (2004). "Nonparametric Bayesian Data Analysis." *Statistical Science*, 19: 95–110. 319

Müller, P. and Rosner, G. (1997). "A Bayesian population model with hierarchical mixture priors applied to blood count data." *Journal of the American Statistical Association*, 92: 1279–1292. 319

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9: 249–265. 324

Quintana, F. and Müller, P. (2004). "Nonparametric Bayesian Assessment of the Order of Dependence for Binary Sequences." *Journal of Computational and Graphical Statistics*, 13: 213–231. 318, 320, 321

Quintana, F. A. and Newton, M. A. (1998). "Assessing the Order of Dependence for Partially Exchangeable Binary Data." *Journal of the American Statistical Association*, 93: 194–202. 320

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4(2): 639–650. 321

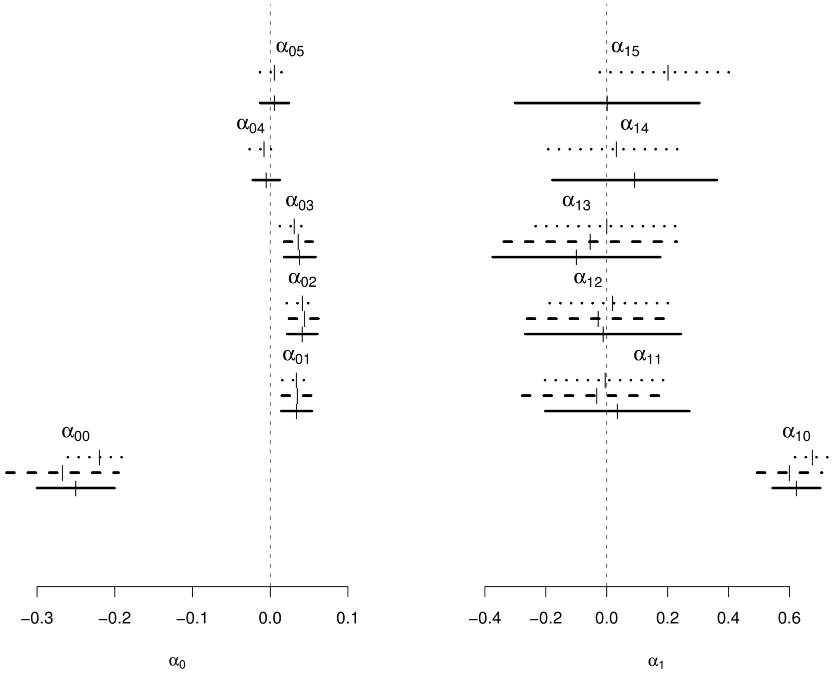
Walker, S. G., Damien, P., Laud, P. W., and Smith, A. F. M. (1999). "Bayesian Nonparametric Inference for Random Distributions and Related Functions (with discussion and a reply from the authors)." *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61: 485–527. 319

Name	# Seasons	Mean	Variance
			(×10)
1 Harold Baines	3	−0.614	0.0710
2 Jesse Barfield	4	−0.508	0.0635
3 Marty Barrett	2	−0.513	0.0599
4 Kevin Bass	2	−0.558	0.0675
5 George Bell	4	−0.584	0.0631
6 Wade Boggs	4	−0.153	0.0581
7 Barry Bonds	4	−0.678	0.0925
8 Bobby Bonilla	4	−0.572	0.0874
9 Sid Bream	2	−0.603	0.0740
10 Ellis Burks	4	−0.799	4.4320
11 George Brett	2	−0.371	0.0782
12 Jose Canseco	2	−0.702	0.0662
13 Gary Carter	2	−0.562	0.0755
14 Joe Carter	3	−0.632	0.0622
15 Will Clark	4	−0.555	0.0942
16 Vince Coleman	4	−0.791	0.0695
17 Alvin Davis	4	−0.494	0.0756
18 Eric Davis	4	−0.473	0.0868
19 Glenn Davis	3	−0.598	0.0668
20 Andre Dawson	2	−0.620	0.0805
21 Rob Deer	4	−0.644	0.0811
22 Bill Doran	3	−0.489	0.0661
23 Brian Downing	3	−0.388	0.0658
24 Darrell Evans	2	−0.578	0.0723
25 Dwight Evans	4	−0.458	0.0658
26 Tony Fernandez	4	−0.628	0.0606
27 Scott Fletcher	4	−0.509	0.0718
28 Julio Franco	4	−0.637	0.0695
29 Gary Gaetti	4	−0.588	0.0659
30 Andres Galarraga	4	−0.651	0.1247
31 Ozzie Guillen	4	−0.898	0.0843
32 Tony Gwynn	4	−0.468	0.0602
33 Mel Hall	2	−0.610	0.0913
34 Billy Hatcher	2	−0.784	0.1026
35 Jack Howell	3	−0.526	0.2448
36 Kent Hrbek	2	−0.557	0.0681
37 Brook Jacoby	4	−0.608	0.0683
38 Howard Johnson	4	−0.646	0.1752
39 Wally Joyner	3	−0.509	0.0633
40 Carney Lansford	4	−0.667	0.0697
41 Chet Lemon	2	−0.667	0.0976
42 Don Mattingly	3	−0.371	0.0558
43 Willie McGee	2	−0.784	0.0862
44 Mark McGwire	4	−1.053	0.8992
45 Kevin McReynolds	4	−0.500	0.0664
46 Paul Molitor	3	−0.625	0.0914
47 Keith Moreland	2	−0.641	0.0679
48 Lloyd Moseby	3	−0.657	0.0666
49 Dale Murphy	3	−0.627	0.0637
50 Eddie Murray	2	−0.385	0.0718
51 Pete O'Brien	3	−0.447	0.0656
52 Gerald Perry	2	−0.565	0.5409
53 Jim Presley	2	−0.783	0.0704
54 Kirby Puckett	4	−0.541	0.0595
55 Harold Reynolds	4	−0.905	0.1003
56 Cal Ripken Jr.	4	−0.560	0.0611
57 Juan Samuel	2	−0.780	0.0733
58 Ryne Sandberg	4	−0.656	0.0652
59 Benito Santiago	2	−0.740	0.7035
60 Steve Sax	2	−0.409	0.0592
61 Dick Schofield	2	−0.620	0.0831
62 Kevin Seitzer	4	−0.243	0.3499
63 John Shelby	2	−0.989	0.1183
64 Ruben Sierra	4	−0.771	0.1125
65 Ozzie Smith	4	−0.437	0.0689
66 Cory Snyder	3	−0.846	0.1099
67 Darryl Strawberry	4	−0.517	0.0760
68 Danny Tartabull	3	−0.599	0.0756
69 Alan Trammell	2	−0.556	0.0661
70 Willie Upshaw	2	−0.615	0.0664
71 Andy Van Slyke	4	−0.614	0.0934
72 Tim Wallach	4	−0.766	0.0857
73 Frank White	2	−0.683	0.0723
74 Willie Wilson	2	−0.767	0.0684
75 Dave Winfield	2	−0.574	0.0665
76 Robin Yount	4	−0.419	0.0702

Table 1: Player numbers, names, number of seasons available for the analysis and means and variances used in the construction of informative prior distributions for logistic intercepts θ_{ij0} as described in Section 2.

Acknowledgments

This research was supported, in part, by grants FONDECYT 1060729, Laboratorio de Análisis Estocástico PBCT ACT-13, and CA075981 from the U.S. National Cancer Institute. The authors wish to thank the reviewers for the helpful suggestions and comments to improve this manuscript.



(a) $p(\alpha_{0m} \mid Y)$

(b) $p(\alpha_{1m} \mid Y)$

Figure 1: *Marginal posterior means and standard deviations for* $\boldsymbol{\alpha}_0$ *and* $\boldsymbol{\alpha}_1$ *coefficients for the proposed model (solid lines), for the model that results when changing* ℓ *to* 3 *(dashed lines) and when letting* $M \to \infty$ *in the original model (dotted lines). The horizontal bars in panel (a) show the marginal posterior mean* $E(\alpha_{0m} \mid Y)$ *(marked by "*$\mid$*") plus/minus one posterior standard deviation* $SD(\alpha_{0m} \mid Y)$*. Panel (b) shows the same for* α_{1m}*.*

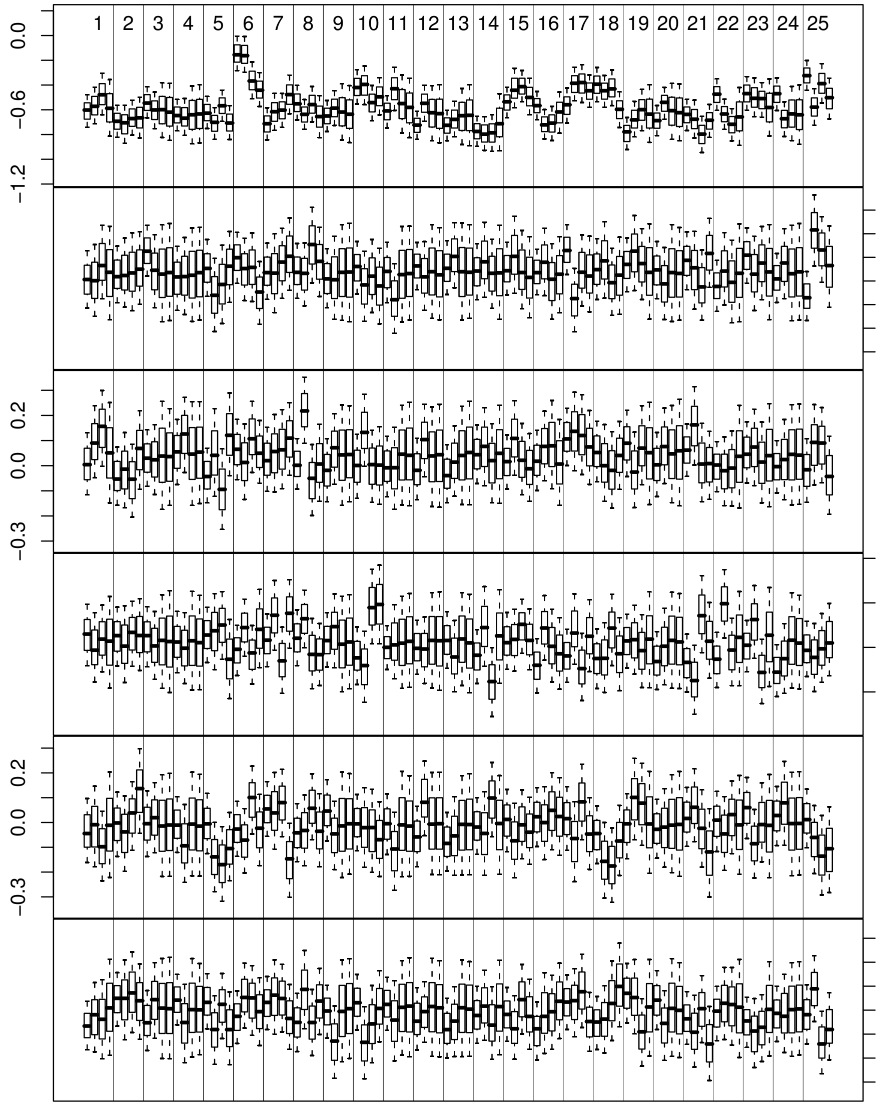


Figure 2: *Posterior marginal boxplots of posterior draws for coordinates of* $\boldsymbol{\theta}_{ij}$ *for players* 1 *through* 25*. Row panels are ordered by component proceeding from the top (*$k = 0$*) down to the bottom (*$k = 5$*). Sectors in each plot correspond to player with index i indicated in the top panel, and within a sector, coordinates are sorted by season, from left to right.*

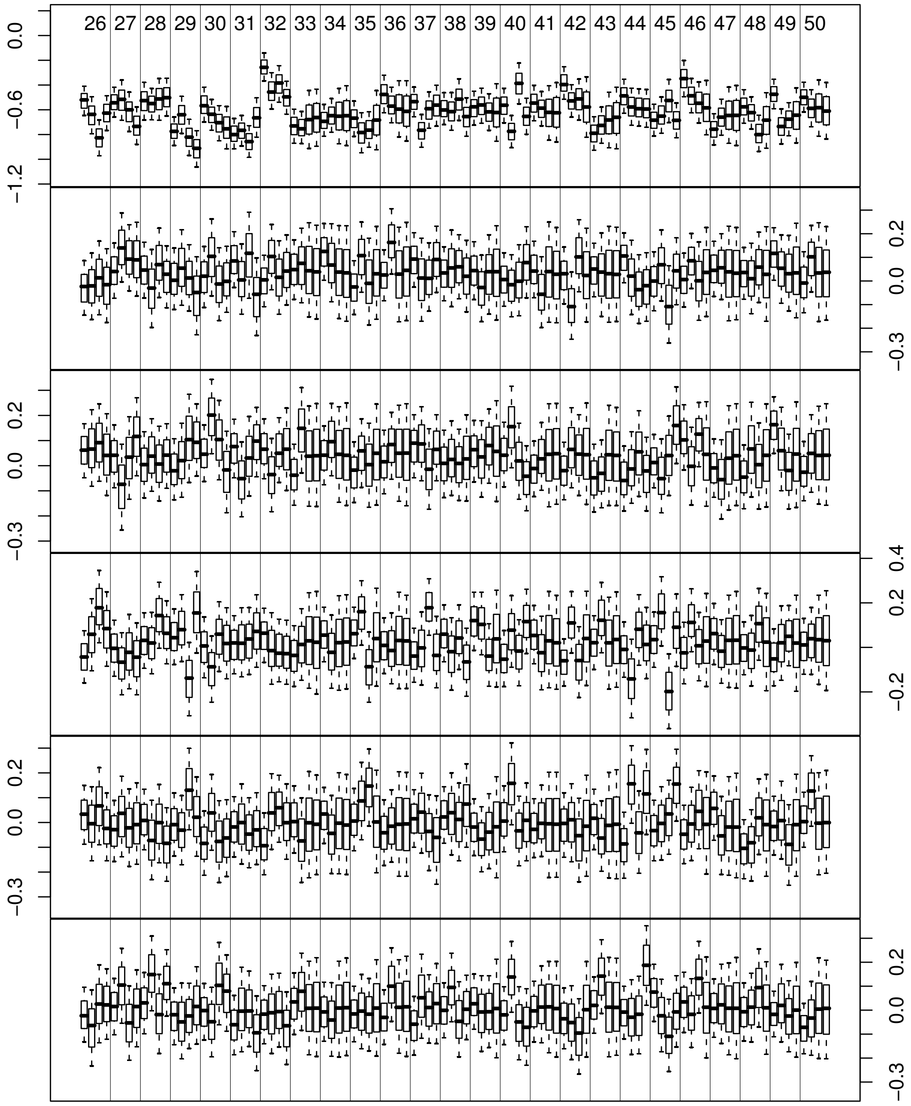


Figure 3: *Posterior marginal boxplots of posterior draws for coordinates of* $\boldsymbol{\theta}_{ij}$ *for players* 26 *through* 50*. Row panels are ordered by component proceeding from the top (*$k = 0$*) down to the bottom (*$k = 5$*). Sectors in each plot correspond to player with index i indicated in the top panel, and within a sector, coordinates are sorted by season, from left to right.*

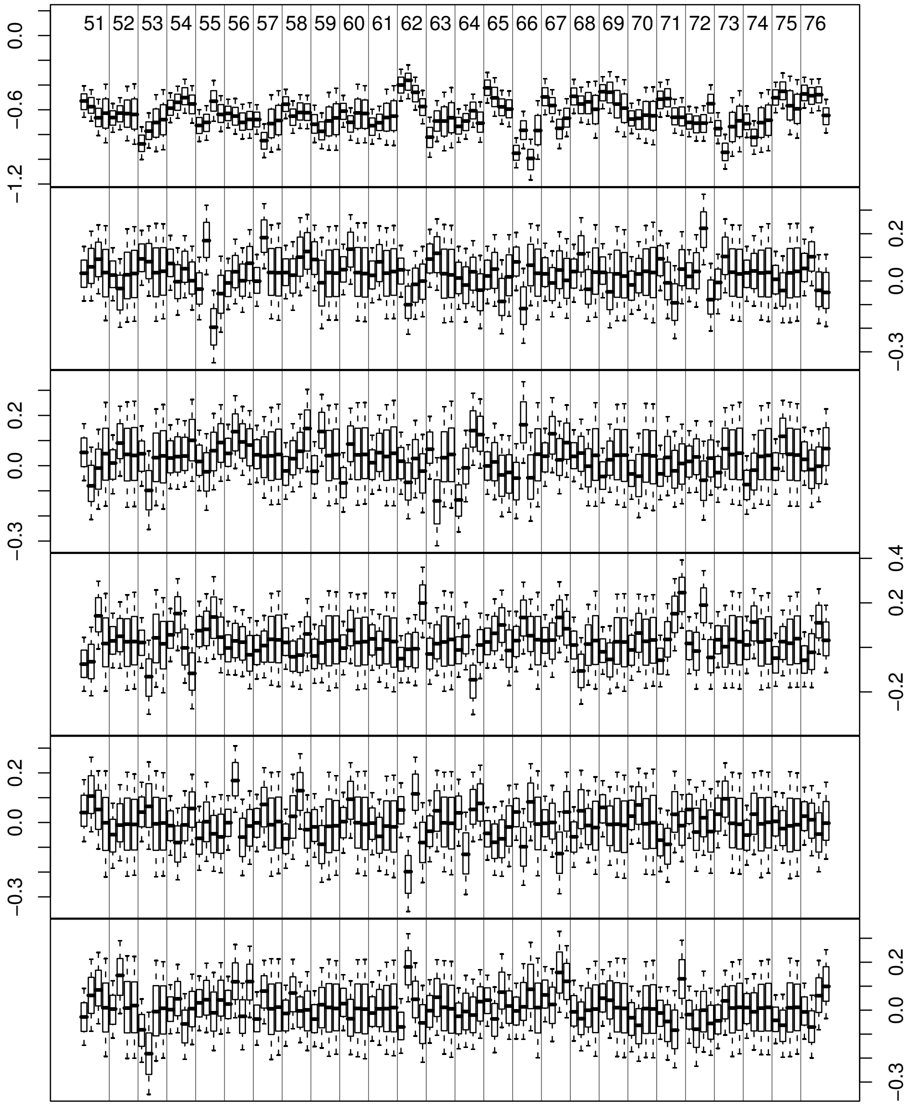


Figure 4: *Posterior marginal boxplots of posterior draws for coordinates of* $\boldsymbol{\theta}_{ij}$ *for players* 51 *through* 76*. Row panels are ordered by component proceeding from the top* ($k = 0$) *down to the bottom* ($k = 5$)*. Sectors in each plot correspond to player with index i indicated in the top panel, and within a sector, coordinates are sorted by season, from left to right.*

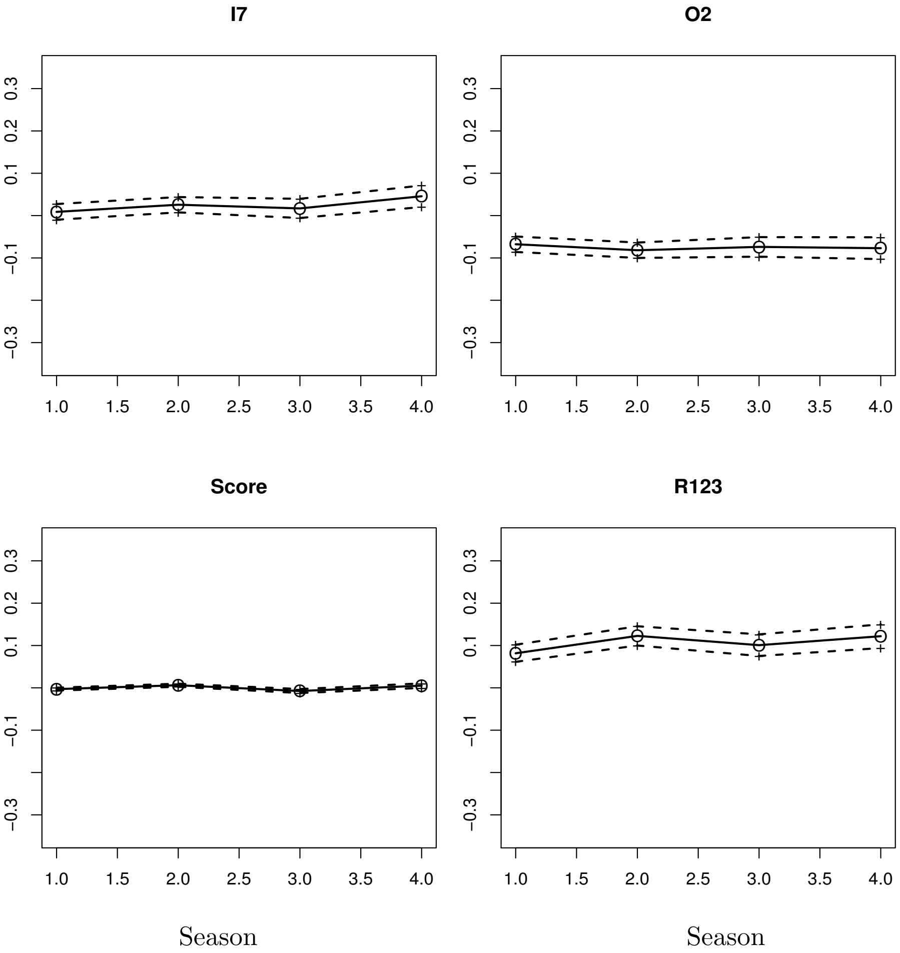


Figure 5: *Posterior marginal means plus and minus one standard deviation for the first four covariate coefficients across seasons. See the text for an explanation.*

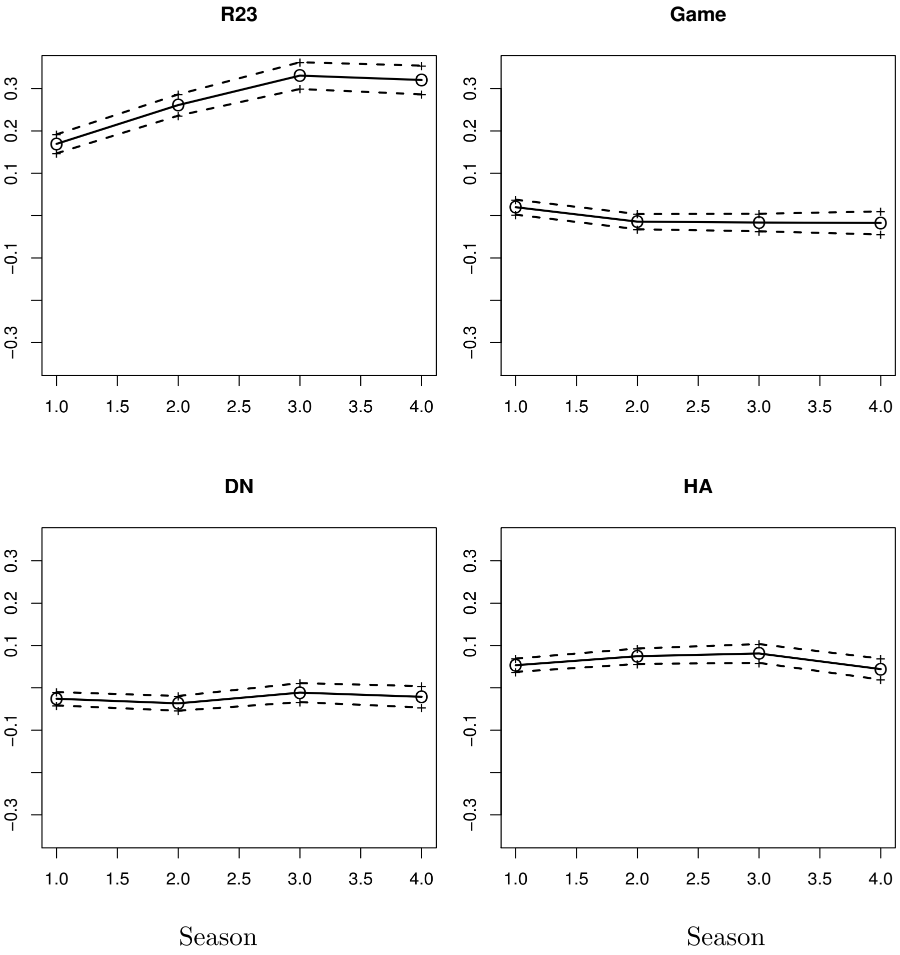


Figure 6: *Posterior marginal means plus and minus one standard deviation for the following four covariate coefficients across seasons. See the text for an explanation.*

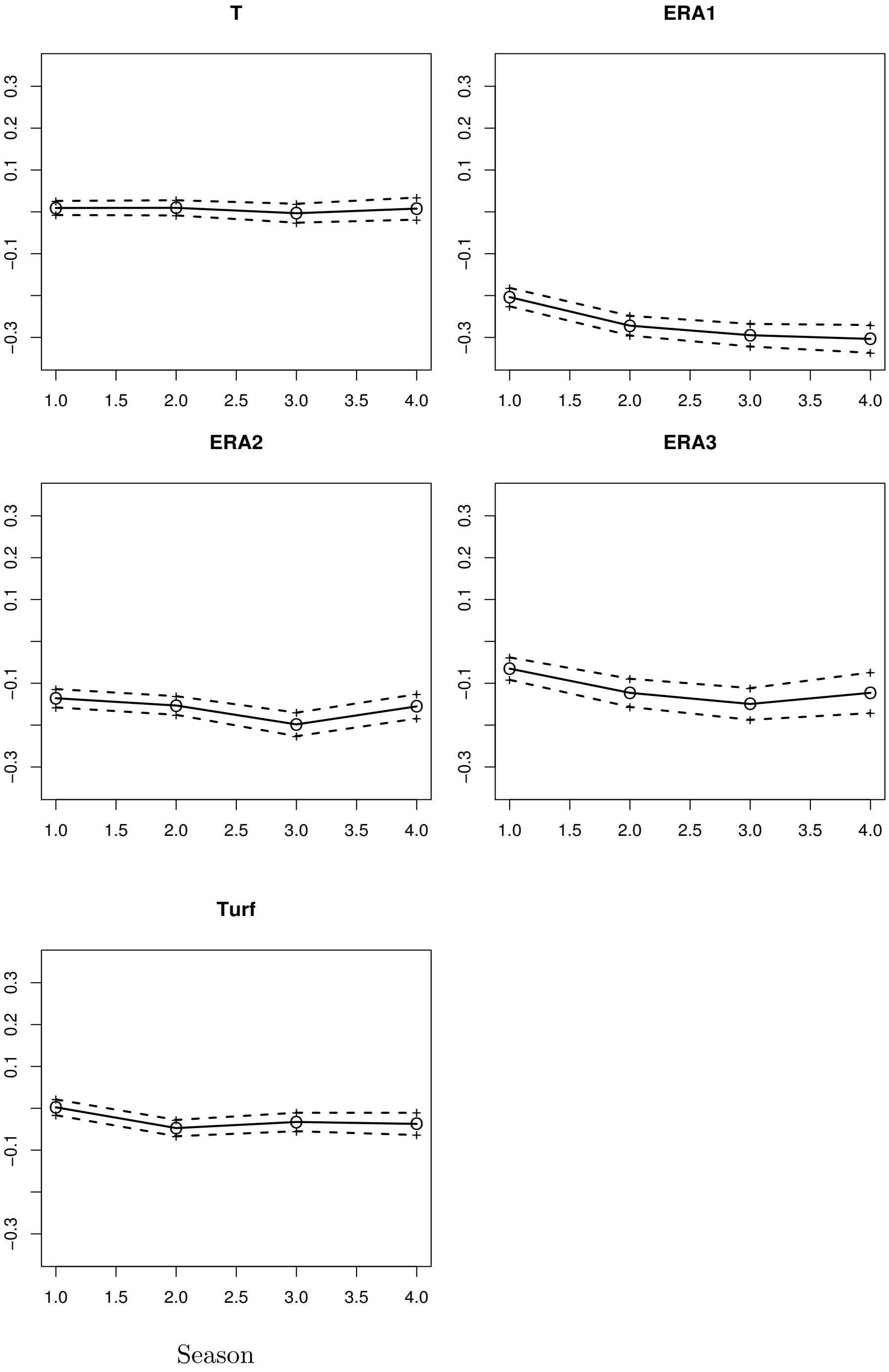


Figure 7: *Posterior marginal means plus and minus one standard deviation for the remaining five covariate coefficients across seasons. See the text for an explanation.*