Bayesian Analysis (2009) **4**, Number 4, pp. 665–668

Comment on Article by Jensen et al.

Fernando A. Quintana* and Peter Müller †

1 Introduction

We congratulate Shane T. Jensen, Blake McShane and Abraham J. Wyner (henceforth JMW) for a very well written and interesting modeling and analysis of hitting performance for Major League Baseball players. JMW proposed a hierarchical model for data extracted from the Lahman Baseball Database. They model the player/year-specific home run rate using covariate information such as the player's age, home ballpark, and position. The proposed approach successfully strikes a balance of parsimonious assumptions where detail does not matter versus structure where it is important for the underlying decision problem. An interesting feature of the model is the time-dependence that is induced by assuming the existence of a hidden Markov chain that drives the transition of players between "elite" and "non-elite" conditions. In the former case, JMW postulate that the home run rate is increased by a certain position-dependent quantity. The model is used to predict home run totals for the 2006 season, and the results compared to some external methods (MARCEL and PECOTA). The comparison gives some mixed results, with the proposed method rating generally well, compared to their competitors.

2 Some general comments

Inference for the Lahmann baseball data raises a number of practical challenges. The data include records on over 2,000 players, but for many of them there is information for only a couple of years. In many cases there are several years with missing information. As usual in sports data, there is tremendous heterogeneity and unbalance among the experimental units (players). We suspect this is partly the reason why the focus is on predictions for a subset of players. However, this opens the question of whether the model actually provides a good fit for *all* the players. We believe an interesting challenge is to extend the modeling approach to larger subsets, and maybe all players. For such extended inference the model needs to be extended to properly reflect the increased heterogeneity across all players. We propose a possible approach below. Also, the inference focus would shift from prediction to more emphasis on an explanatory model.

Model (2) and the proposed variations, have the interesting feature of incorporating in the home run rates θ_{ij} an explicit dependence on player position k, home ballpark

*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, Houston, TX, mailto:pmueller@mdanderson.org

 DOI:10.1214/09-BA424C

b and a smooth position-specific age trajectory, expressed as an hypothesized linear combination in the logit scale. The smooth function of age seems to capture interesting nonlinear features of the home run rates evolution on time, as seen in Figures 3 and 5. One may even venture the existence of an "optimal" age for hitting, and a natural decay in abilities with progressing age. In fact, such conclusions have been reached elsewhere, and even if not the target of this work, it is a nice feature of the analysis that the same kind of findings are uncovered.

The hidden Markov model for "elite" status is the model component that is responsible for introducing dependence across seasons for a given player. The extended model allows for player-specific transition parameters, i.e., individual trajectories for the binary elite indicator variables. Concretely, JMW assume the parameters (ν^i_{00}, ν^i_{11}) controlling these transitions to be a priori independent and Beta-distributed, with conditional independence across players sharing a same position k. These assumptions imply flexibility in the evolution of the $\{E_{ij}\}$ elite indicators, which are well defined regardless of missing data patterns along the sequences of home runs. Looking at the results of the analysis, it is quite remarkable that a large number of players achieve elite status after only one or two major league seasons, as seen in Figure 2. Intuitively one would have expected a peak more likely around 3-5 years. JMW seem to be equally surprised at such findings, when they comment that the sum over years 2 through 11 still represents 75% of the cases considered.

Another consequence of the elite/non-elite model is that the effect on home run rates θ_{ij} is only through a position-specific added term $\alpha_k = \alpha_{k0}(1 - E_{ij}) + \alpha_{k1}E_{ij}$ on the logit scale. While this has the advantage of borrowing strength across players with the same position, it may be not flexible enough to capture highly heterogeneous home run profiles.

3 Extending the proposed approach

The latent elite indicator E_{ij} defines a mixture model for the observed home run totals. The use of E_{ij} is an elegant way to formalize inference about top players. The model balances parsimony with sufficient structure to achieve the desired inference. The authors correctly point out some of the remaining limitations. Perhaps the most important limitation is that the model reduces the heterogeneity of the population of all players to a mixture of only two homogeneous subpopulations. This is particularly of concern in the light of the underlying decision problem. The resulting inference only informs us about the probability of a player being in the elite group. Some evidence for more heterogeneity beyond the mixture of only two subpopulations is seen in Figure 4. The wide separation of the credible intervals suggests scope for intermediate performance groups in the model. The population of players is highly heterogeneous, but not in such a sharply bimodal fashion. It is also interesting to note in the same figure the almost preserved ordering across positions between elite and non-elite groups.

A minor extension of the model could generalize the mixture to a random partition into H subpopulations, which could help closing the gap just pointed out. Each cluster

could have a cluster-specific set of intercepts α_{kh}, $h = 0, \ldots, H-1$ for the logistic regression prior (2) of player-season home run rates θ_{ij}. Like in JMW's model, the intercepts remain ordered $\alpha_{kh} \leq \alpha_{k,h+1}$, $k = 1, \ldots, 9$. This allows us to interpret the clusters labels $h = 0, \ldots, H-1$ as latent player performance.

Formally the model extension would replace (2) by

$$\text{logit}(\theta_{ij}) = \alpha_{ih} + \beta_b + f_k(A_{ij}), \tag{1}$$

where β_b and $f_k(A_{ij})$ are as earlier, and $h = E_{ij}$ is the imputed cluster membership for player i in season j. The prior for $\boldsymbol{\alpha_k} = (\alpha_{kh},\ h = 0, \ldots, H-1)$ is similar to (9), now for the $H-$dimensional vector $\boldsymbol{\alpha_k}$. The prior for the latent cluster membership E_{ij} remains as in (3), extended to transitions between H states. The number of transition parameters ν_{rs} remains unchanged with prior probability ν_{01} for upgrades in elite level, prior probability ν_{10} for downgrades and ν_{00} for the probability of remaining in state $E_{ij} = 0$ and ν_{11} for the probability of remaining in a performance state $E > 0$. Like in (7) the transition probabilities are position-specific.

The number of states H would itself be treated as unknown, with a geometric prior $p(H) = (1-p)^{H-1}p$ and a hyperparameter p. The only additional step in the MCMC implementation is a transition probability to change H. We consider two transitions, "birth" of an additional performance level by splitting an existing level h into two new levels and the reverse "death" move. This could be implemented as a reversible jump move.

The generalized model defines a random partition of the player-years (ij) into performance clusters $h = 0, \ldots, H-1$. The unique features of this random partition model would be the ordering of the clusters and the dependence across j. Both features are naturally accommodated by the outlined model-based clustering. We see it as an interesting and challenging application of model-based clustering. In contrast to much of the of clustering models in the recent Bayesian literature, the use of standard clustering models such as the ubiquitous Polya urn would be inappropriate. The Polya urn model does not naturally allow the desired ordering of cluster-specific parameters and time-dependence of cluster membership indicators.

4 Final words

We realize the above proposal can be extended/modified in many different ways, the main point being the possibility of improving on the analysis and model proposed by JMW. Our aim here was not to criticize the model but to help improve it. We indeed think the hidden Markov component is a very nice feature, which combined with a flexible extension, could motivate further analysis of the data under a more general framework.

Acknowledgments

Fernando Quintana was partially funded by Fondecyt grant 1060729.