

Babe Ruth and Bayesian Shrinkage

Shane T. Jensen and Jonathan Weinberg

Department of Statistics, The Wharton School,

University of Pennsylvania

stjensen@wharton.upenn.edu

Abstract

Hierarchical linear models are used in many applications where the observed data is considered to be a linear function of one or more latent variables that are themselves generated from one or more underlying population models. In the Bayesian framework, the population model for each type of latent variable is represented by a shared prior distribution. A well-known effect of this framework is the phenomenon of **shrinkage**: the estimated values for each latent variable is pulled towards their common prior (or population) mean. When the correct prior distribution is used, this shrinkage has well-documented benefits, but if the underlying prior distribution is misspecified, Bayesian shrinkage estimators can have a dramatic effect on the tails of the latent variable distribution. As an example, using a normal prior distribution will lead to dramatic shrinkage of extreme observations. We examine the shrinkage effects of different Bayesian estimators in an application to the historical modeling of hitting performance in baseball and discuss several strategies for the objective use of prior distributions in hierarchical modeling.

Motivation: Modeling Home Run Hitting

Berry et al. (1999) developed a hierarchical model with the intended goal of linking different eras of baseball players based upon two measures of performance: home run and batting average. We focus on their model for home-run hitting, where we have home-run totals from around 7000 players across 96 seasons and 78 ballparks through the 1996 season.

Home-run totals h_{ij} for a particular player i in year j of their career is modeled as a binomial outcome,

$$h_{ij} \sim \text{Binomial}(m_{ij}, \pi_{ij})$$

with probability π_{ij} satisfying the following equation,

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \theta_i + \delta_y + \epsilon_p + f_i(a_{ij}).$$

θ_i represents the intrinsic ability of player i

δ_y is an effect of the year y

ϵ_p is an effect of home ballpark p

$f_i(a_{ij})$ is a player-specific function of the player's age a_{ij}

There is an additional level of common prior distributions for these parameters,

$$\delta_y \sim \text{Normal}(0, 1)$$

$$\epsilon_p \sim \text{Normal}(0, 1)$$

$$\theta_i \sim \text{Normal}(\mu_d, \sigma_d^2)$$

μ_d and σ_d^2 are decade-specific ability parameters that have additional common prior distributions with fixed hyperparameters:

$$\mu_d \sim \text{Normal}(-3.5, 1)$$

$$\sigma_d^2 \sim \text{Inv} - \text{Gamma}(3, 3)$$

Results of Berry Model

The following table gives the players with the top 10 estimated ability parameters θ_i from the Berry model (along with their unadjusted home run totals and proportions).

Rank	Player	Berry Estimates		Home Run Statistics	
		Mean(θ_i)	SD(θ_i)	HR	HR / AB
1	M.McGwire	0.104	0.006	329	0.0806
2	J. Gonzalez	0.098	0.008	214	0.0684
3	B. Ruth	0.094	0.004	714	0.0850
4	D. Kingman	0.093	0.004	442	0.0662
5	M. Schmidt	0.092	0.005	548	0.0656
6	H. Killebrew	0.090	0.005	573	0.0703
7	F. Thomas	0.089	0.007	222	0.0675
8	J. Canseco	0.088	0.004	328	0.0647
9	R. Kittle	0.086	0.006	176	0.0650
10	W. Stargell	0.084	0.003	475	0.0599

It is strange to see that two of the top four career leaders in home runs, H. Aaron (755 HR) and W. Mays (660 HR), do not appear in the top 10. However, this result is not as surprising when you consider that the proportion of home runs for both H. Aaron (0.0610) and W. Mays (0.0607) are not as high as any member of the top 10 list.

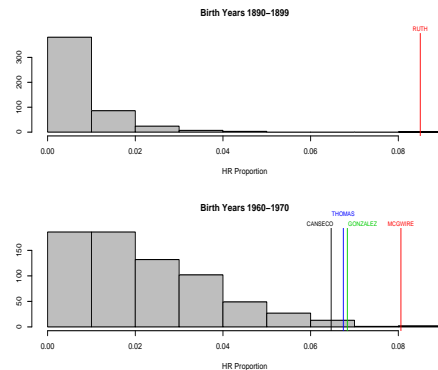
The more surprising result is that **Babe Ruth has a clearly superior HR total and HR proportion** relative to the rest of the top 10, yet the Berry model estimates that his ability parameter θ_i to be only third highest, behind M. McGwire and J. Gonzalez.

Shrinkage to the Decade-Specific Mean

The Berry model assumption that all players in a particular decade d share a common normal distribution for their ability parameters θ_i has the effect of shrinking observations in towards the common estimated mean μ_d . This shrinkage property of hierarchical models can be both a blessing and a curse, depending on the inferential goal.

If one is interested in estimating the average performance of players in a decade, then this type of model makes quite a bit of sense. However, if we are interested in estimating the "best" players of an era, then we are dealing with observations from the tail of the distribution which may be shrunk too far in towards the common mean.

The ability parameter θ_i for Babe Ruth is dramatically affected by this shrinkage to a common mean because he played in an era where home runs were much less common. This can be seen by comparing the distribution of HR proportions between players born in the same decade (1890-1899) as Babe Ruth and players born in the same decade as Mark McGwire and Juan Gonzalez.



The home run hitting of Babe Ruth is much more extreme relative to his contemporaries than the home-run hitting of modern-day players. However, the model attributes much of his success to luck due to the fact that he is such an aberrant observation. This shrinkage effect is counter-intuitive, Babe Ruth is penalized for his dominance over his contemporaries instead of being rewarded.

Shrinkage and the Prior Distribution

This shrinkage effect is especially dramatic when a **Normal prior distribution** is used, since the thin tails make extreme observations very unlikely. Babe Ruth's HR proportion is **9.1 standard deviations** above the mean of players in his decade, and assuming a Normal population distribution, we would expect **1 in every 2.21×10^{19} players** to be this extreme! If we instead assume player abilities have a t -distribution with 4 degrees of freedom, we would expect **1 in every 2473 players** to be this extreme.

References

- S.M. Berry, C.S Reese and P.D. Larkey (1999). Bridging different eras in sports. *JASA* 94:661-685.
- B. Efron and C. Morris (1971). Limiting the risk of Bayes and Empirical Bayes estimators - Part I: The Bayes case. *JASA* 66:807-815
- J.O. Berger (1990). Robust Bayesian analysis: sensitivity to the prior. *JSPI* 25:303-328

Alternative Model Suggestions

This phenomenon is well-established in the statistical literature and several alternatives have been suggested:

MLE and Limited Translation Rules:

Efron & Morris (1971) examine a Normal-Normal model and show that the Bayes estimate can be higher risk than the maximum likelihood estimate when the underlying population (prior) distribution contains a sub-population of extreme observations and suggest constraining the Bayes shrinkage to be a maximum distance away from the maximum likelihood estimate, and thereby restricting the risk for observations which are extreme relative to the prior distribution.

t distribution as a Robust Alternative:

Berger (1990) and many others have developed many strategies for robust Bayesian analysis in the presence of extreme outliers, such as the use of a t distribution with low degrees of freedom (say, 4). This prior assumption would be equivalent to the Berry use of the normal distribution if the variance σ^2 was allowed to vary between observations (with an inverse χ^2 prior distribution).

Comparison in Application

Using MCMC methods, we implemented the simplified version of the Berry model under two alternative prior distributions for the ability parameters θ : flat and t_4 . Park and year effects were also included, but not the aging function $f(\cdot)$. The table below shows the top 10 players (according to ability θ) for the original Berry model and the simplified Berry model with flat and t prior distributions.

Rank	Original Berry	Flat Prior	t Prior
1	M.McGwire	B. Ruth	B. Ruth
2	J. Gonzalez	M.McGwire	J. Foxc
3	B. Ruth	R. Kiner	M.McGwire
4	D. Kingman	J. Foxc	R. Kiner
5	M. Schmidt	T. Williams	H. Greenburg
6	H. Killebrew	H. Greenburg	H. Killebrew
7	F. Thomas	H. Killebrew	T. Williams
8	J. Canseco	P. Seerey	P. Seerey
9	R. Kittle	D. Kingman	D. Kingman
10	W. Stargell	R. Kittle	R. Kittle

Compared to the original Berry model, these two alternative priors have resulted in a greater number of players from early baseball being included on the top 10 list. **Only two of the Berry top ten players were retired by 1980, compared with seven of the top ten players from the models with flat and t priors.** This result was expected, since these players played in eras with generally less home-run hitting, and thus are penalized by the Berry model more dramatically than either alternative prior. Correspondingly, several modern-day players are considered less impressive by the models with flat or t priors. As an example, J. Gonzalez was ranked 2nd by the Berry model but only 14th and 13th by the flat and t model respectively. Finally, we see that Babe Ruth is considered the top home run hitter by the flat and t models, though it should be noted that our implementation of a simplified Berry model (no aging function) with a normal prior also has Babe Ruth ranked first, which suggests that the presence/absence of the aging function also plays a role in these results.

Another Option: Mixture Normal Prior

If the underlying population (prior) distribution might contain a sub-population of extreme observations, an alternative prior specification would be to directly model this sub-population with a mixture of normal (or t) distributions which would allow for additional inference about the sub-population in addition to reducing the risk due to shrinkage. This work is part of a continuing collaboration with ESPN to evaluate the past and predicted future performance of major league baseball players. A key component of this project will be the modeling of individuals within a hierarchical Bayesian framework, with particular attention given to the extremes of player performance.