7-2010

A Point-Mass Mixture Random Effects Model for Pitching Metrics

James M. Piette III
University of Pennsylvania

Alexander Braunstein

Blakeley B. McShane
University of Pennsylvania

Shane T. Jensen
University of Pennsylvania

Recommended Citation

Piette, J. M., Braunstein, A., McShane, B. B., & Jensen, S. T. (2010). A Point-Mass Mixture Random Effects Model for Pitching Metrics. *Journal of Quantitative Analysis in Sports, 6* (3), http://dx.doi.org/10.2202/1559-0410.1237

A Point-Mass Mixture Random Effects Model for Pitching Metrics

Abstract

A plethora of statistics have been proposed to measure the effectiveness of pitchers in Major League Baseball. While many of these are quite traditional (e.g., ERA, wins), some have gained currency only recently (e.g., WHIP, K/BB). Some of these metrics may have predictive power, but it is unclear which are the most reliable or consistent. We address this question by constructing a Bayesian random effects model that incorporates a point mass mixture and fitting it to data on twenty metrics spanning approximately 2,500 players and 35 years. Our model identifies FIP, HR/9, ERA, and BB/9 as the highest signal metrics for starters and GB%, FB%, and K/9 as the highest signal metrics for relievers. In general, the metrics identified by our model are independent of team defense. Our procedure also provides a relative ranking of metrics separately by starters and relievers and shows that these rankings differ quite substantially between them. Our methodology is compared to a Lasso-based procedure and is internally validated by detailed case studies.

Keywords

baseball, Bayesian models, entropy, mixture models, random effects

Disciplines

Applied Statistics | Statistics and Probability

Journal of Quantitative Analysis in Sports

Volume 6, *Issue* 3 2010 *Article* 8

A Point-Mass Mixture Random Effects Model for Pitching Metrics

James Piette, *University of Pennsylvania*
Alexander Braunstein, *Google, Inc.*
Blakeley B. McShane, *Kellogg School of Management, Northwestern University*
Shane T. Jensen, *University of Pennsylvania*

Recommended Citation:
Piette, James; Braunstein, Alexander; McShane, Blakeley B.; and Jensen, Shane T. (2010) "A Point-Mass Mixture Random Effects Model for Pitching Metrics," *Journal of Quantitative Analysis in Sports*: Vol. 6: Iss. 3, Article 8.

DOI: 10.2202/1559-0410.1237

A Point-Mass Mixture Random Effects Model for Pitching Metrics

James Piette, Alexander Braunstein, Blakeley B. McShane, and Shane T. Jensen

Abstract

A plethora of statistics have been proposed to measure the effectiveness of pitchers in Major League Baseball. While many of these are quite traditional (e.g., ERA, wins), some have gained currency only recently (e.g., WHIP, K/BB). Some of these metrics may have predictive power, but it is unclear which are the most reliable or consistent. We address this question by constructing a Bayesian random effects model that incorporates a point mass mixture and fitting it to data on twenty metrics spanning approximately 2,500 players and 35 years. Our model identifies FIP, HR/9, ERA, and BB/9 as the highest signal metrics for starters and GB%, FB%, and K/9 as the highest signal metrics for relievers. In general, the metrics identified by our model are independent of team defense. Our procedure also provides a relative ranking of metrics separately by starters and relievers and shows that these rankings differ quite substantially between them. Our methodology is compared to a Lasso-based procedure and is internally validated by detailed case studies.

KEYWORDS: baseball, Bayesian models, entropy, mixture models, random effects

1 Introduction

Nobody likes to hear it, because it's dull, but the reason you win or lose is darn near always the same - pitching.

Earl Weaver

There is an avid interest in assessing the degree of signal in various metrics of pitching performance. A key element in this discussion is the differentiation between events a pitcher can control and events that he cannot. Events within the control of the pitcher should be better indicators of his true ability and thus more predictive of future performance. In contrast, events beyond the control of the pitcher can be attributed to chance variation and are not predictive of future performance.

The baseball literature contains a litany of articles devoted to finding pitching metrics that are within a pitcher's control. McCracken (2001) provides an approach based on defensive independent pitching statistics (DIPS), which aims to remove outside influences (e.g., team defense) from common measures such as ERA. This strategy is supported by an examination of batting average on balls in play (BABIP). Specifically, McCracken (2001) observed that: (i) a player's BABIP was only weakly correlated from year to year, (ii) team BABIP was a better predictor of a given pitcher's BABIP next year than his own BABIP this year, and (iii) the range of BABIP exhibited over the course of a career was "about the same as the range you would expect from random chance."

In contrast, Tippett (2003) detected some degree of stability in BABIP from year to year and found that players who faced more batters over the course of their careers (i.e., more successful pitchers) had a lower BABIP than those who faced fewer batters. However, the year to year correlation of BABIP was still observed to be much weaker than defense independent measures such as strikeout rate.

Bradbury (2005) showed that the previous season DIPS is a better predictor of current ERA than previous season ERA, and concluded that "while pitchers may have some ability to prevent hits on balls in play, the effect is small. And any effect a pitcher does have is reflected within DIPS metrics". Gassko (2007) showed that previous year DIPS is predictive of current year BABIP even when controlling for previous year BABIP, league, park, and defense effects. In other studies, Keri (2007) developed a measure called fielding independent pitching (FIP),Caruth (2007) examined home run to fly ball ratio (HR/FB), and Studeman (2006) examined the percentage of balls in play that are line drives (LD%).

Albert (2006) presents a beta-binomial model to tease out the variation in pitcher performance due to variation in pitcher talent versus random chance. This

model is applied to seven metrics: walk rate, strikeout rate, home run rate, non-home run rate, run rate, and earned run rate. Albert (2006) finds the strikeout rate to be the highest signal metric and explores the changes in these metrics between eras.

Generally, the previous literature in this area lacks systematic evidence of signal across (i) many metrics for (ii) many seasons with (iii) many players. Most previous arguments for consistency of a metric are based on year-to-year correlations which do not provide a complete story. We address this issue with a Bayesian random effects model that assesses the degree of signal in twenty different pitching metrics using data from approximately 2,500 unique pitchers (starters and relievers) over thirty-five seasons.

2 Methodology

2.1 Data and Model

Our data is taken from the *Fangraphs* database (www.fangraphs.com) and spans thirty-five seasons (1974-2008). We examine the twenty pitching metrics outlined in our appendix separately for starting pitchers and relief pitchers. We exclude player-seasons for starters who pitch less than 100 innings and relievers who pitcher less than 40 innings. For pitchers who move from one role to another in a season, we consider their starter innings and reliever innings separately.

Our interest is determining which metrics have signal and which are dominated by noise. If a metric were pure noise, pitchers would be completely inconsistent from year to year, and the best prediction one could make for a given pitcher in future years would be the league average. This theory suggests a *minimum threshold* for a metric: pitchers must perform consistently with respect to that metric over the course of their careers.

We assess the consistency of a metric with a Bayesian point-mass mixture random effects model. Our focus is on the novel results of our analysis, so we provide a brief overview of the model here. A more detailed description of the model and estimation procedure can be found in McShane et al. (2009).

DOI: 10.2202/1559-0410.1237

$$
\begin{aligned}
y_{ij}|\alpha_i,\sigma^2,\mu &\sim N(\mu+\alpha_i,\sigma_{ij}^2) && \text{Likelihood}\\
\alpha_i|\tau^2,\gamma_i &\sim \begin{cases}\approx 0 & \text{if } \gamma_i=0\\ N(0,\tau^2) & \text{if } \gamma_i=1\end{cases} && \text{Player Difference}\\
\gamma_i &\sim Bernoulli(p_1) && \text{Player Indicator}\\
\mu &\sim N(0,100^2) && \text{Prior for League Mean}\\
p_1 &\sim Beta(1,1) && \text{Prior for Mixing Proportion}\\
\sigma^2 &\sim IG(.01,.01) && \text{Prior for Season Variance}\\
\tau^2 &\propto 1/\tau && \text{Prior for Player Variance}
\end{aligned}
$$

In the model, y_{ij} is the observation for pitcher i in season j of a particular metric. μ is the overall major league baseball mean whereas α_i is the difference between pitcher i's mean and this overall league mean. We note that each player-season has a specific variance $\sigma_{ij}^2 = \sigma^2 \cdot \bar{n}/n_{ij}$ where n_{ij} is the number of innings pitched by pitcher i in seasons j and $\bar{n}$ is the average number of innings pitched. This adjustment accounts for the fact that seasons for which a pitcher pitched more contain more information than those for which he pitched less. The overall season-to-season variance is captured by the global parameter σ^2.

The key modeling assumption is that the random effects α_i come from a mixture of two components: either a point-mass at zero or a non-zero random effect with variance τ^2. For each pitcher i, the variable γ_i indicates whether his individual mean is equal to the league mean ($\gamma_i = 0$) or different ($\gamma_i = 1$). A key global parameter for each metric will be p_1, the proportion of pitchers that have non-zero random effects α_i.

In plain terms, our model assumes there are two kinds of pitchers: those who have an "intrinsic talent" equal to that of the league and those who do not. Approximately p_1 percent of pitchers are in the latter category and they deviate from the league mean by about $\pm\tau$. Finally, for a given pitcher with fixed talent level, his season to season variation is on the order of $\pm\sigma$. Our model generalizes a standard random effects model, which estimates non-zero α_i for all pitchers. In a high signal setting where the standard random effects model holds, we should estimate p_1 to be close to one.

2.2 Evaluating Metrics

As mentioned above, a minimum threshold for a high signal metric would be that pitchers perform consistently with respect to it over time, so that predictions for

a pitcher's future would be based on their personal history rather than the league average. This suggests two important criteria for a metric: (i) player-specific information trumps league information for a large fraction of players and (ii) one has high confidence about the specific pitchers who truly differ from league average. By fulfilling both criteria, a useful metric contains both global evidence (large fraction of players with signal) and local evidence (high confidence for individual players) of high signal.

The first criterion is addressed by our model parameter p_1, which identifies exactly the fraction of players for which individual information trumps league information. We will use the posterior mean $\widehat{p_1}$ of this parameter to assess whether or not a metric fulfills this first criterion.

Our model can also be used to address the second criterion. For each player i, we can calculate the posterior mean $\widehat{\gamma_i}$ of their indicator variable γ_i. When $\widehat{\gamma_i}$ is close to one, we are very confident that player i has a non-zero random effect (i.e., a different personal mean compared to the league mean). Conversely, when $\widehat{\gamma_i}$ is close to zero, we are very confident that player i is not different from the league mean.

Hence, metrics with $\widehat{\gamma_i}$ near zero or one for most pitchers are metrics which give high confidence about the quality of individual pitchers. We can formalize this concept into a single value, the average $-$Entropy (Jaynes, 1957):

$$-\text{Entropy} = \frac{1}{m}\sum_{i=1}^{m}[\widehat{\gamma_i}\log(\widehat{\gamma_i}) + (1-\widehat{\gamma_i})\log(1-\widehat{\gamma_i})] \tag{1}$$

Metrics with large values of both $\widehat{p_1}$ and $-$Entropy are high signal metrics: they are metrics which have (i) individual means different from the league mean for a large fraction of players and for which (ii) it is clear which players are different from the league mean and which are not. In the next section, we will show how each of the twenty pitching metrics perform on these two measures separately for starters and relievers.

3 Results

Our discussion of results is divided into three subsections. We first examine the overall conclusions of our model before delving into a more detailed discussion of several high signal metrics. Our model inference is then compared to the results of a simpler analysis using the Lasso (Tibshirani, 1996) approach to sparse regression.

DOI: 10.2202/1559-0410.1237

3.1 Overall Evaluation of Metrics

An ideal metric demonstrates strong signal in our model by having both $\widehat{p_1}$ near one and $-$Entropy near zero. In Figure 1, we plot $\widehat{p_1}$ versus $-$Entropy for each of the twenty metrics separately by starters and relievers. Starters are given in black and relievers in red[1]. We also provide the data from Figure 1 in Table 1.

There appear to be three clusters of metrics visible in the upper-left plot of Figure 1. First, there are metrics demonstrating strong signal in a consistent manner, with $\widehat{p_1}$ near one and $-$Entropy near zero. These high signal metrics are indicated by the dotted rectangle and are plotted in more detail in bottom panel of the figure. We also see a continuum of metrics which have moderate signal, with $-$Entropy less than -0.25 but $\widehat{p_1}$ greater than 0.5. These are indicated by the dashed rectangle and are plotted in more detail in the top right panel of the figure. Finally, we see some isolated metrics with low signal ($\widehat{p_1} < 0.5$).

The metrics with highest signal are GB%, FB%, and K/9 for relievers and FIP, HR/9, Pitches, ERA, and BB/9 for starters, and these findings have some support in the previous literature. K/9, BB/9, and HR/9 fit into the context of DIPS/FIP paradigm of McCracken (2001)and Keri (2007). Our findings about GB% and FB% are supported by Tippett (2003), who noted that certain pitchers tend to induce more ground ball or fly ball outs than others. Our results are also relevant to previous discussions (Treder, 2004) of bullpen specialization: ground ball specialists such as Chad Qualls and lefty/sidearm strikeout specialists such as Mike Myers, Will Ohman, and Chad Bradford are excellent examples of the importance of GB%, FB%, and K/9 for relievers.

In contrast, K/BB and GB/FB stand out as particularly noisy measures. We have observed that the ratios of two metrics will be quite noisy even when both the numerator and denominator contain signal. Thus, K/9, BB/9, GB%, and FB% all demonstrate moderate to high amounts of signal while their ratios do not. LD% is another metric which shows little to no signal, a claim which has found both support (Gassko, 2006) and criticism (Studeman, 2004).

Our model provides an interesting examination of the common ERA metric. ERA has long been considered to be noisy and inferior to defense independent measures such as DIPS or FIPS (McCracken, 2001). Our model confirms the consensus that FIP is a better measure, with a larger $\widehat{p_1}$ and $-$Entropy than ERA. Our analysis also suggests, however, that ERA does contain substantial signal, in most part due to the large component of FIPS contained within ERA (Keri, 2007). Interestingly, our results suggest that ERA dominates WHIP even though WHIP is often thought of as a superior measure.

[1] We fit our model to IFFB% but do not report our results for this metric since it does not fulfill the normality assumption of our model. Thus, any conclusions about IFFB% are held very tentatively.

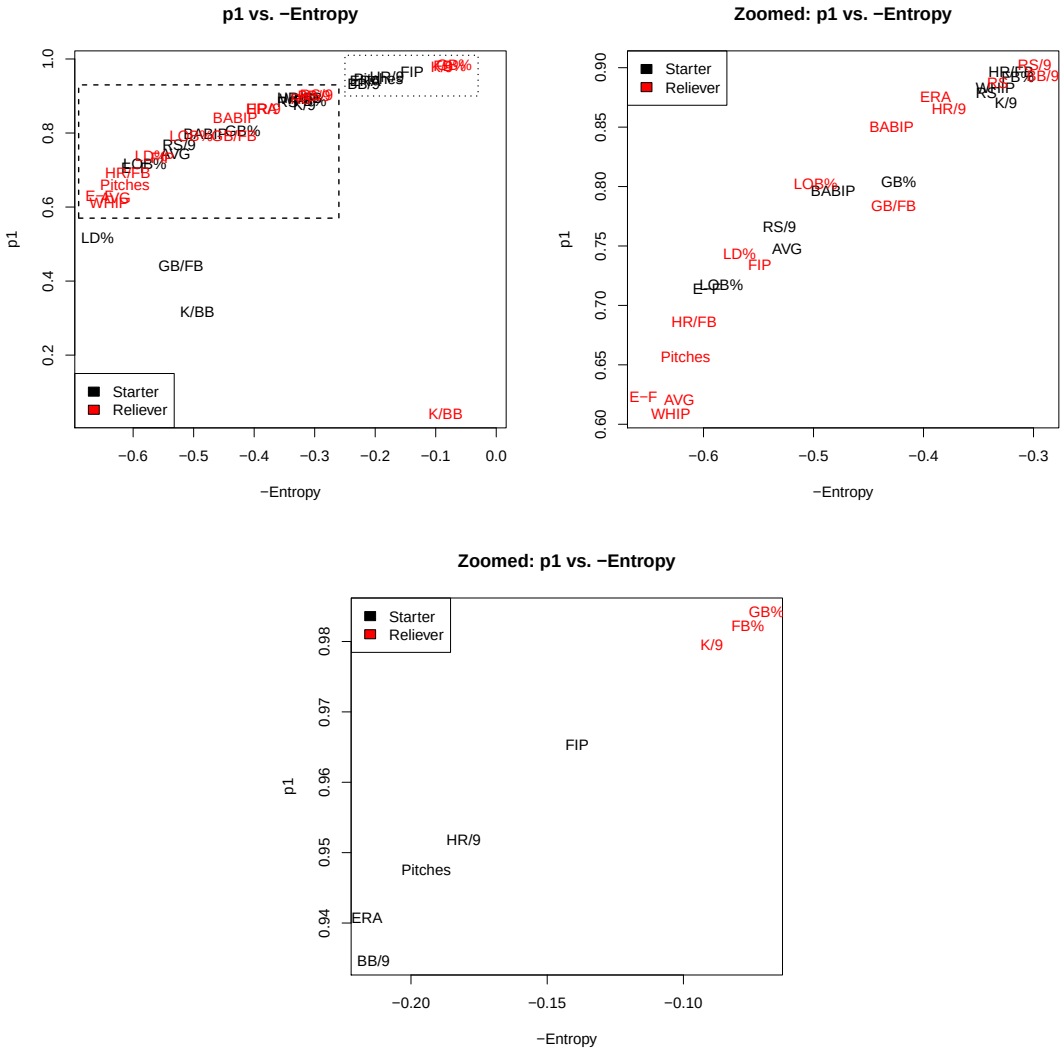

Figure 1: On the top left we plot $\widehat{p_1}$ for each pitching metric versus $-$Entropy. On the top right, we zoom in on the metrics indicated by the dashed rectangle in the first plot. On the bottom, we zoom in on the high signal metrics indicated by the dotted rectangle in the first plot. IFFB% is excluded as it does not fulfill the model assumption of normality.

DOI: 10.2202/1559-0410.1237

Starters		
Metric	$\widehat{\mathbf{p_1}}$	**-Entropy**
FIP	0.965	-0.139
HR/9	0.952	-0.181
Pitches	0.948	-0.194
ERA	0.941	-0.216
BB/9	0.935	-0.219
HR/FB	0.896	-0.325
WHIP	0.892	-0.325
FB%	0.888	-0.307
RS	0.886	-0.345
K/9	0.877	-0.317
GB%	0.807	-0.419
BABIP	0.799	-0.480
RS/9	0.769	-0.524
AVG	0.745	-0.530
LOB%	0.719	-0.581
E-F	0.705	-0.597
LD%	0.516	-0.659
GB/FB	0.442	-0.521
K/BB	0.317	-0.494

Relievers		
Metric	$\widehat{\mathbf{p_1}}$	**-Entropy**
GB%	0.984	-0.069
FB%	0.982	-0.076
K/9	0.979	-0.090
RS/9	0.904	-0.297
BB/9	0.901	-0.301
RS	0.894	-0.323
ERA	0.867	-0.387
HR/9	0.865	-0.385
BABIP	0.843	-0.431
LOB%	0.794	-0.504
GB/FB	0.793	-0.433
LD%	0.739	-0.570
FIP	0.732	-0.551
HR/FB	0.694	-0.609
Pitches	0.662	-0.613
E-F	0.632	-0.655
AVG	0.627	-0.629
WHIP	0.613	-0.639
K/BB	0.041	-0.084

Table 1: Table of Metrics by Starter and Reliever: each table gives estimated $\widehat{p_1}$ and $-$Entropy and is sorted by the former. IFFB% is excluded as it does not fulfill the model assumption of normality.

Another controversial pitching measure, batting average on balls in play, is shown to have moderate signal by our analysis. BABIP has a reasonably large $\widehat{p_1}$ for both starters and relievers, supporting the findings from Tippett (2003). However, BABIP does not have an impressive $-$Entropy, which leads some credence to McCracken (2001).

We see substantial discrepancy between the high-signal metrics for relievers versus starters. Part of this divergence is due to the different situations faced by each type of pitcher. For example, it is important for starters to minimize the number of pitches they throw in order to maximize the number of batters they face. So, effective starters tend to favor pitching to contact, alleviating the risk of a walk. In contrast, relievers do not have the same pitch count worries and have less concern about the pitch count costs associated with walks. This allows them to place more

Starter's FIP		
Player	**Mean** $(\mu+\alpha_i)$ **Estimate**	**SD**
Best Five Players		
Nolan Ryan	3.02	0.12
Pedro Martinez	3.03	0.14
J. R. Richard	3.06	0.18
Roger Clemens	3.13	0.10
Jon Matlack	3.16	0.18
Worst Five Players		
Scott Elarton	5.28	0.24
Jamey Wright	5.02	0.20
Ricky Bones	5.01	0.23
Rob Bell	4.98	0.29
Ramon Ortiz	4.93	0.21
Population Mean $\hat{\mu} = 4.16$		

Starter's ERA		
Player	**Mean** $(\mu+\alpha_i)$ **Estimate**	**SD**
Best Five Players		
Jim Palmer	3.16	0.20
Pedro Martinez	3.17	0.19
Roger Clemens	3.22	0.15
Jose Rijo	3.25	0.24
Greg Maddux	3.27	0.14
Worst Five Players		
Jamey Wright	4.88	0.26
Darren Oliver	4.80	0.26
LaTroy Hawkins	4.89	0.33
Eric Milton	4.76	0.23
Scott Elarton	4.86	0.29
Population Mean $\hat{\mu} = 4.11$		

Table 2: Best and Worst Five Players (by $\widehat{\alpha}_i$) for Starter's FIP and ERA

focus on accruing the most effective outs, strikeouts. We discuss the distinction between starting and relief pitching in more detail in Sections 3.2 and 3.3 below.

3.2 Examining Starters on High Signal Metrics

We focus our examination of individual players on the metrics FIP, ERA, HR/9, and BB/9, which were found in Section 3.1 to be high signal for starting pitchers (we omit discussion of Pitches because it is not particularly illuminating). In Table 2, we present the best and worst five starters in terms of their estimated random effects ($\widehat{\alpha}_i$) for the FIP metric. As discussed previously, FIP is considered the "signal" component from the more common measure ERA, so we also include the best and worst starters for ERA and note there is overlap between the two on the best and worst five.

Table 2 features several hall-of-fame caliber pitchers among the top five pitchers in terms of FIP, as well as several great players with careers cut short by injury (J. R. Richard and Jose Rijo). Scott Elarton is found to be the worst pitcher (by a substantial margin) on this measure. However, we suspect that his FIP was

DOI: 10.2202/1559-0410.1237

Starter's HR/9		
Player	**Mean $(\mu+\alpha_i)$**	
	Estimate	**SD**
Best Five Players		
J. R. Richard	0.51	0.09
Bruce Berenyi	0.53	0.11
Steve Rogers	0.54	0.07
Tommy John	0.56	0.07
Nolan Ryan	0.56	0.05
Worst Five Players		
Scott Elarton	1.53	0.11
Jose Lima	1.48	0.09
Eric Milton	1.43	0.08
Brian Anderson	1.39	0.09
Rick Helling	1.38	0.09
Population Mean $\hat{\mu} = 0.94$		

Starter's BB/9		
Player	**Mean $(\mu+\alpha_i)$**	
	Estimate	**SD**
Best Five Players		
Bob Tewksbury	1.52	0.20
Carlos Silva	1.63	0.26
Gary Nolan	1.67	0.35
Bret Saberhagen	1.70	0.17
Brad Radke	1.73	0.17
Worst Five Players		
Kazuhisa Ishii	5.10	0.35
Bobby Witt	4.85	0.17
Jose DeJesus	4.84	0.38
Daniel Cabrera	4.78	0.27
Jason Bere	4.76	0.26
Population Mean $\hat{\mu} = 3.14$		

Table 3: Best and Worst Five Players (sorted by $\widehat{\alpha}_i$) for Starter's HR/9 and BB/9.

inflated by the fact that he played in Mile High Stadium (pre-humidor) for three seasons, a park known for harming a pitcher's numbers.

In Table 3, we examine the best and worst starting pitchers in terms of the HR/9 and BB/9 metrics. For both metrics, we see that there is a huge difference between these pitchers and the league average, which implies a large τ^2 parameter in both cases. For the HR/9 measure, one contributing factor is ballpark: both Scott Elarton and Eric Milton spent large parts of their career in home-run-friendly ballparks. In fact, Milton has been a league average pitcher excluding his years spent in Citizens Bank Park and the Great American Ballpark.

For the BB/9 measure, we see several pitchers among the best five that are famous for their control. In the cases of Bob Tewksbury and Gary Nolan, their ability to restrict walks allowed them to remain successful late into their careers. In contrast, the worst five list is dominated by "power pitchers" such as Kazuhisa Ishii and Daniel Cabrera who displayed impressive velocity in their careers but struggled with control. Ishii, the worst pitcher, struck out nearly one batter per inning in his first season, a very good K/9 for any pitcher. However, he walked 6.19 hitters per nine innings that season, nearly twice the league average rate. The Dodgers employed him another two seasons, but his control barely improved.

Reliever's GB%		
Player	**Mean** $(\mu+\alpha_i)$ **Estimate**	**SD**
Best Five Players		
Cla Meredith	0.662	0.026
Bill Swift	0.634	0.026
Chad Bradford	0.629	0.020
Roger McDowell	0.625	0.014
Roy Corcoran	0.623	0.039
Worst Five Players		
Troy Percival	0.287	0.016
Jeff Reardon	0.291	0.019
Gabe White	0.295	0.021
Al Reyes	0.301	0.023
Ugueth Urbina	0.308	0.017
Population Mean $\hat{\mu} = 0.449$		

Reliever's FB%		
Player	**Mean** $(\mu+\alpha_i)$ **Estimate**	**SD**
Best Five Players		
Troy Percival	0.544	0.016
Carlos Marmol	0.511	0.028
Al Reyes	0.499	0.023
Scott Proctor	0.497	0.026
Julio Mateo	0.496	0.022
Worst Five Players		
Cla Meredith	0.178	0.026
Bill Swift	0.188	0.025
Roger McDowell	0.194	0.014
Derek Lowe	0.196	0.026
Chad Bradford	0.206	0.018
Population Mean $\hat{\mu} = 0.351$		

Table 4: Best and Worst Five Players (sorted by $\widehat{\alpha}_i$) for Reliever's GB% and FB% Compared to $\hat{\mu}$.

3.3 Examining Relievers on High Signal Metrics

Unlike starters, relievers are typically employed in a more situationally dependent manner. Rather than measurements of game-long performance (as for starters), the high signal metrics for relievers more closely relate to individual at bats. Table 4 focuses on two of these high signal metrics: GB% and FB%.

We expect a negative correlation between GB% and FB%, which is not surprising considering grounders and flys are both alternative outcomes for a ball in play. This negative correlation is quite dramatic in Table 4, where several of the best pitchers in terms of GB% correspond to the worst pitchers in terms of FB%. Cla Meredith, Bill Swift, and Roger McDowell are routinely recognized as elite ground ball pitchers, appearing with $\widehat{\alpha}_i$ very positive for GB% and very negative for FB%. We also see Chad Bradford, known for his quirky side arm delivery releasing the ball inches from the ground, among the top ground-ball pitchers.

Table 5 shows the best and worst relief pitchers by the high signal metric K/9. The worst pitchers on this measure are not particularly notable, but the best five pitchers include several elite closers of the last few years. Eric Gagne has the longest streak of converted saves (84), and Brad Lidge had a perfect 2008 season

Reliever's K/9		
Player	**Mean** $(\mu + \alpha_i)$	
	Estimate	**SD**
Best Five Players		
Brad Lidge	12.01	0.46
Rob Dibble	11.89	0.48
Billy Wagner	11.57	0.35
Octavio Dotel	11.42	0.45
Eric Gagne	11.25	0.52
Worst Five Players		
Dan Quisenberry	3.40	0.31
Mike Proly	3.54	0.49
Steve Comer	3.55	0.68
Jim Todd	3.56	0.48
Doug Sisk	3.56	0.44
Population Mean $\hat{\mu} = 6.45$		

Table 5: Best and Worst Five Players (sorted by $\widehat{\alpha_i}$) for Reliever's K/9 Compared to $\hat{\mu}$.

(converting each of his 41 regular and 7 post season save opportunities). Although our approach focuses on career performance rather than seasonal performance, our results for strikeout rates are similar to those of Albert (2006).

3.4 Comparison to a Lasso-based Approach

The Lasso is a penalized least squares regression model which constrains some parameter estimates to be zero (Tibshirani, 1996), and so it is commonly used for variable selection. Thus, an alternative way to examine pitching metrics would be to classify a particular metric as high signal if it had a large "Lasso %": the percentage of players estimated to have non-zero means by the Lasso. This parameter is the Lasso analogue of $\widehat{p_1}$ from our model.

Figure 2 provides pairwise plots of Lasso % against $\widehat{p_1}$ (left) and $-$Entropy (right). Our model and the Lasso show general agreement in terms of both, though there are three notable exceptions for $\widehat{p_1}$: Starters' GB/FB, Starters' K/BB, and Relievers' K/BB. All three of these are ratio metrics, which are typically noisier than their component numerator and denominator. Our methodology captures this noisiness by estimating $\widehat{p_1}$ near one-half and low $-$Entropy values whereas the

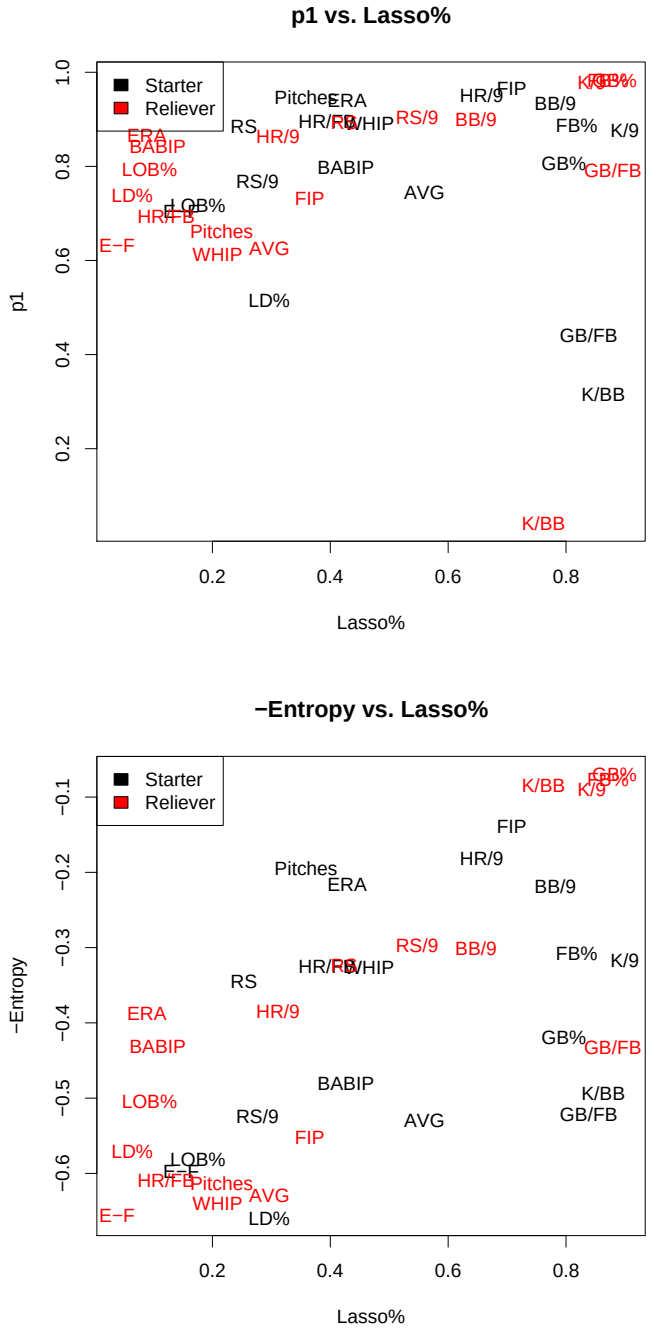

Figure 2: Left: Plot of $\widehat{p}_1$ (y-axis) against the percentage of players with non-zero means selected by the Lasso. Right: Plot of $-$Entropy (y-axis) versus percentage of players with non-zero means selected by the Lasso. IFFB% is excluded as it does not fulfill the model assumption of normality.

DOI: 10.2202/1559-0410.1237

Lasso estimates an individual mean for about 80% of players for these three metrics. For these three measures, we prefer the conservative approach of our model.

4 Discussion

We have developed a Bayesian random effects model which incorporates a point mass mixture to distinguish players that are different from league average. We argue that high signal metrics (i) have a large fraction of players which are different from the league average and (ii) give high confidence about which players are not league average. These criteria provide a framework for finding high signal metrics. Our analysis considers starters and relievers separately, and we find a large disparity in high signal metrics between these two types of pitchers.

Our model identifies FIP, HR/9, ERA, and BB/9 as the highest signal metrics for starters and GB%, FB%, and K/9 as the highest signal metrics for relievers. Anecdotally, these measures are independent of team defense and thus support previous work (McCracken, 2001; Keri, 2007). BABIP, which has been a controversial measure in past studies (McCracken, 2001; Tippett, 2003), is found to be middle of the pack in terms of the amount of signal it contains. ERA is identified as a high signal metric, though this is explained by the fact that FIP (an even higher signal metric) is a significant component of ERA.

There are several directions for future work. We plan to fit our model to other metrics that attempt to remove park, team, or league effects. Our approach is also based on the assumption that true performance on each measure is constant across a pitcher's career. We plan to extend our model by building in non-constant career trajectories or perhaps correlated error structures.

A Pitching Metrics

Our 20 pitching measures are enumerated in the table below. Several term referenced in the table below are IBB (intentional walk), IP (innings pitched), H (hits), HR (home run), AB (at-bats/batters faced), FB (fly balls), GB (ground balls), LD (line drives), and BIP (balls in play). Note that for the batted ball measures (FB%, LD%, GB%, IFFB%, GB/FB, HR/FB, RS/9, RS, and Pitches), there are only 19 seasons worth of data. The rest of the measures are based on 35 seasons (1974-2008).

Metric y_{ij}	**Description**
AVG	batting average (H/AB)
BABIP	batting average for balls in play (H/BIP)
FB%	fly ball percentage (FB/BIP)
GB%	ground ball percentage (GB/BIP)
LD%	line drive percentage (LD/BIP)
IFFB%	infield fly ball percentage (IFFB/FB)
K/9	strikeouts per 9 innings
BB/9	walks per 9 innings
HR/9	home runs per 9 innings
K/BB	ratio of K to BB
GB/FB	ratio of GB to FB
HR/FB	ratio of HR to FB
LOB%	one minus the ratio of runners scoring to all runners reaching base
ERA	earned runs per 9 innings
FIP	Fielding Indep. Pitching = (HR*13 + (BB + HBP - IBB)*3 - K*2) / IP
E-F	ERA minus FIP
WHIP	(BB + H)/IP
RS	runs scored
RS/9	RS per 9 innings
Pitches	pitches thrown per season

References

Albert, J. (2006): "Pitching statistics, talent and luck and the best strikeout seasons of all-time," *Journal of Quantitative Analysis in Sports*, 2, 2.

Bradbury, J. C. (2005): "Another look at dips," *http://www.hardballtimes.com/main/article/another-look-at-dips1*, May 24, 2005.

Caruth, M. (2007): "Groundballs and homerun rates," *http://www.hardballtimes.com/main/article/groundballs-and-homerun-rates*, May 11, 2007.

Gassko, D. (2006): "The truth about the grounder," *http://www.hardballtimes.com/main/printarticle/the-truth-about-the-grounder*, May 12, 2006.

Gassko, D. (2007): "Uncovering dips," *http://www.hardballtimes.com/main/article/uncovering-dips*, January 4, 2007.

Jaynes, E. T. (1957): "Information theory and statistical mechancics," *Physical Review*, 106, 620–630.

Keri, J., ed. (2007): *Baseball Between the Numbers: Why Everything You Know about the Game Is Wrong*, Basic Books.

McCracken, V. (2001): "Pitching and defense: How much control do hurlers have?" *http://www.baseballprospectus.com/article.php?articleid=878*, January 23, 2001.

McShane, B. B., A. Braunstein, J. Piette, and S. T. Jensen (2009): "A Bayesian variable selection approach to major league baseball hitting metrics," Technical report, arXiv:0911.4503.

Studeman, D. (2004): "Groundballs, flyballs, and lin drives," *http://www.hardballtimes.com/main/article/groundballs-flyballs-and-line-drives*, May 9, 2004.

Studeman, D. (2006): "Inside der," *http://www.hardballtimes.com/main/article/inside-der*, January 26, 2006.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *J. R. Statist. Soc. B*, 58, 267–288.

Tippett, T. (2003): "Can pitchers prevent hits on balls in play?" *http://www.diamond-mind.com/articles/ipavg2.htm*, July 21, 2003.

Treder, S. (2004): "The closer and the damage done," *http://www.hardballtimes.com/main/article/the-closer-and-the-damage-done*, August 17, 2004.