Bayesian Hierarchical Model for Pitch Efficiency in Major League Baseball

Jacob Andros

December 7, 2021

Background

Model

Level 1:

$$y_{ij}|\theta_i \sim \text{Poisson}(\theta_i) \tag{1}$$

Level 2:

$$\theta_i|a_i, b_i \sim \text{Gamma}(a_i, b_i) \tag{2}$$

Level 3:

$$a_i \sim \text{N}(16, 3) \quad b_i \sim \text{N}(1, \frac{1}{16}) \tag{3}$$

$$i \in \{1, 2, 3, 4, 5, 6\}, \;\; j \in \{1, \ldots, n_i\}$$

Computation

$$
\begin{aligned}
\mathrm{P}(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b} | \boldsymbol{y}) &\propto \mathrm{P}(\boldsymbol{y}|\boldsymbol{\theta})\mathrm{P}(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{b})\mathrm{P}(\boldsymbol{a}, \boldsymbol{b}) \\
&\propto \prod_{i=1}^{6} \Big[\prod_{j=1}^{n_i} \Big(\frac{\theta_i^{y_{ij}} e^{-\theta_i}}{y_{ij}!} \Big) \Big(\frac{b_i^{a_i}}{\Gamma(a_i)} \theta_i^{a_i - 1} e^{-\theta_i b_i} \Big) \exp\{ -\frac{1}{2}(a_i - 16)^2 - \frac{1}{2}(b_i - 1)^2 \} \Big] \\
&\propto \prod_{i=1}^{6} \Big[\frac{1}{\Gamma(a_i)} \theta_i^{a_i - 1 + \sum y_{ij}} e^{-n_i \theta_i} b_i^{a_i} \exp\{ -\theta_i b_i - \frac{1}{2}(a_i - 16)^2 - \frac{1}{2}(b_i - 1)^2 \} \Big] \\
\mathrm{P}(\boldsymbol{\theta} | \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{y}) &\propto \prod_{i=1}^{6} \Big[\theta_i^{a_i - 1 + \sum y_{ij}} e^{-n_i \theta_i - b_i \theta_i} \Big] \propto \prod_{i=1}^{6} \Big[\theta_i^{a_i - 1 + \sum y_{ij}} e^{-n_i \theta_i - b_i \theta_i} \Big] \\
&\implies \theta_i | \cdot \sim \mathrm{Gamma}(a_i + \sum_{j=1}^{n_i} y_{ij}, n_i + b_i)
\end{aligned}
$$

Convergence

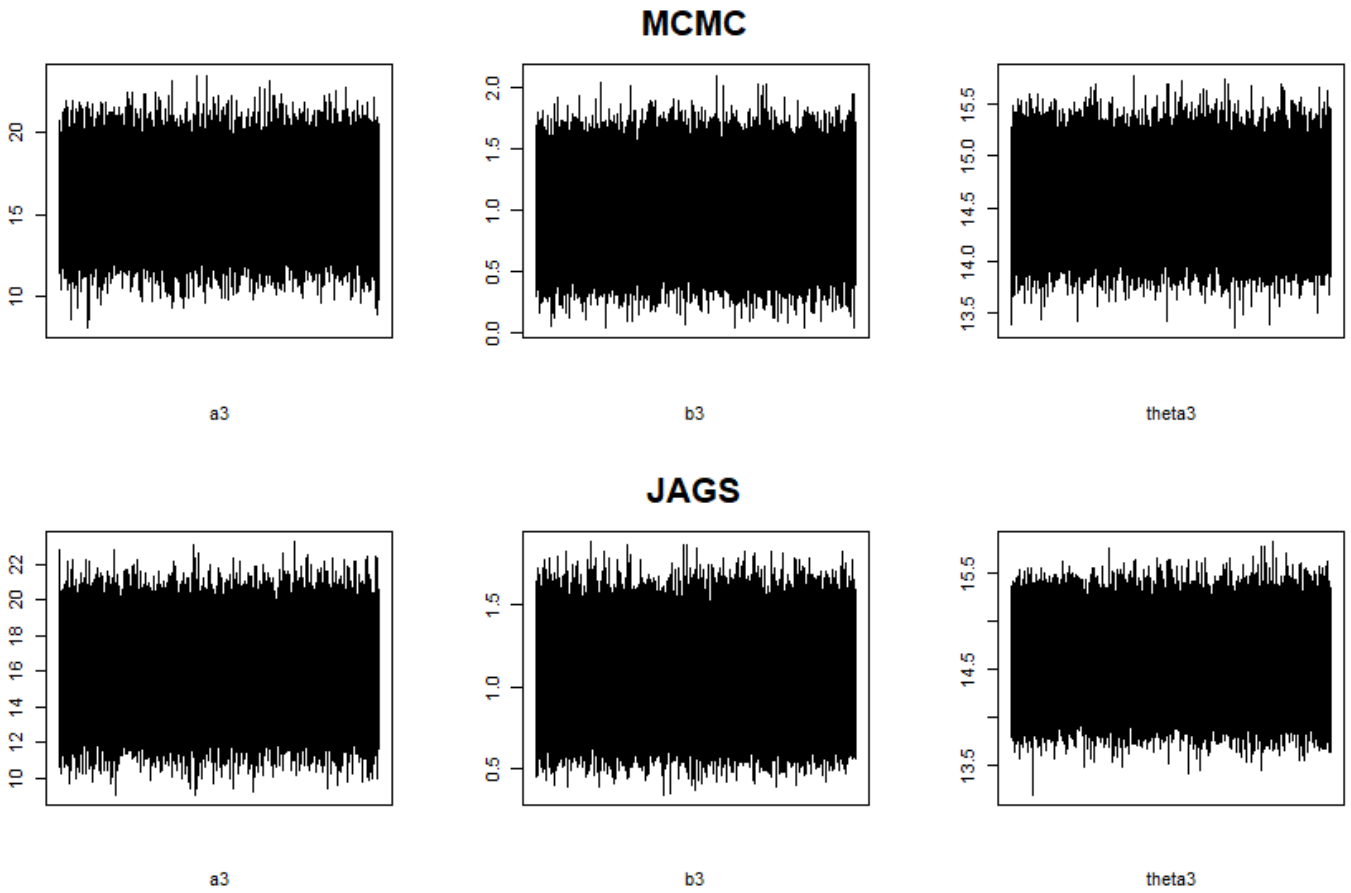

Figure 1: Trace plots for Zack Wheeler (a_3, b_3, θ_3) under each sampling method. The trace plots for the other 5 pitchers are comparable.

Convergence

Better overall convergence and computational efficiency with JAGS, but both methods work.

	Parameters	Chains	Acceptance	ESS	Geweke	$\hat{R}$
MCMC	$\boldsymbol{a}$	1	0.55	32070	.26	NA
	$\boldsymbol{b}$	1	0.28	50022	.09	NA
	$\boldsymbol{\theta}$	1	NA	99673	.09	NA
JAGS	$\boldsymbol{a}$	8	NA	158585	.10	1
	$\boldsymbol{b}$	8	NA	158989	0.37	1
	$\boldsymbol{\theta}$	8	NA	159534	.09	1

Table 1: Convergence comparison between MCMC and JAGS

Comparison of JAGS with my MCMC

MCMC	Pitcher	Posterior Median	95% Credible Interval	95% MC Interval
θ_1	Walker Buehler	14.718	(14.16, 15.29)	(14.71977, 14.71978)
θ_2	Nathan Eovaldi	15.543	(14.94, 16.17)	(15.54786, 15.54787)
θ_3	Zack Wheeler	14.585	(14.05, 15.13)	(14.58555, 14.58556)
θ_4	Taijuan Walker	14.682	(14.05, 15.33)	(14.68592, 14.68593)
θ_5	Shane McClanahan	15.663	(14.90, 16.44)	(15.66531, 15.66532)
θ_6	Max Fried	15.130	(14.49, 15.78)	(15.13280, 15.13281)
JAGS	Pitcher	Posterior Median	95% Credible Interval	95% MC Interval
θ_1	Walker Buehler	14.713	(14.16, 15.29)	(14.71576, 14.71576)
θ_2	Nathan Eovaldi	15.543	(14.94, 16.16)	(15.54521, 15.54522)
θ_3	Zack Wheeler	14.580	(14.05, 15.13)	(14.58233, 14.58234)
θ_4	Taijuan Walker	14.677	(14.04, 15.33)	(14.68056, 14.68057)
θ_5	Shane McClanahan	15.660	(14.91, 16.44)	(15.66372, 15.66373)
θ_6	Max Fried	15.128	(14.50, 15.78)	(15.13112, 15.13112)

Table 2: Posterior summarization under each method. The medians and credible intervals are virtually identical.

Comparison of JAGS with my MCMC

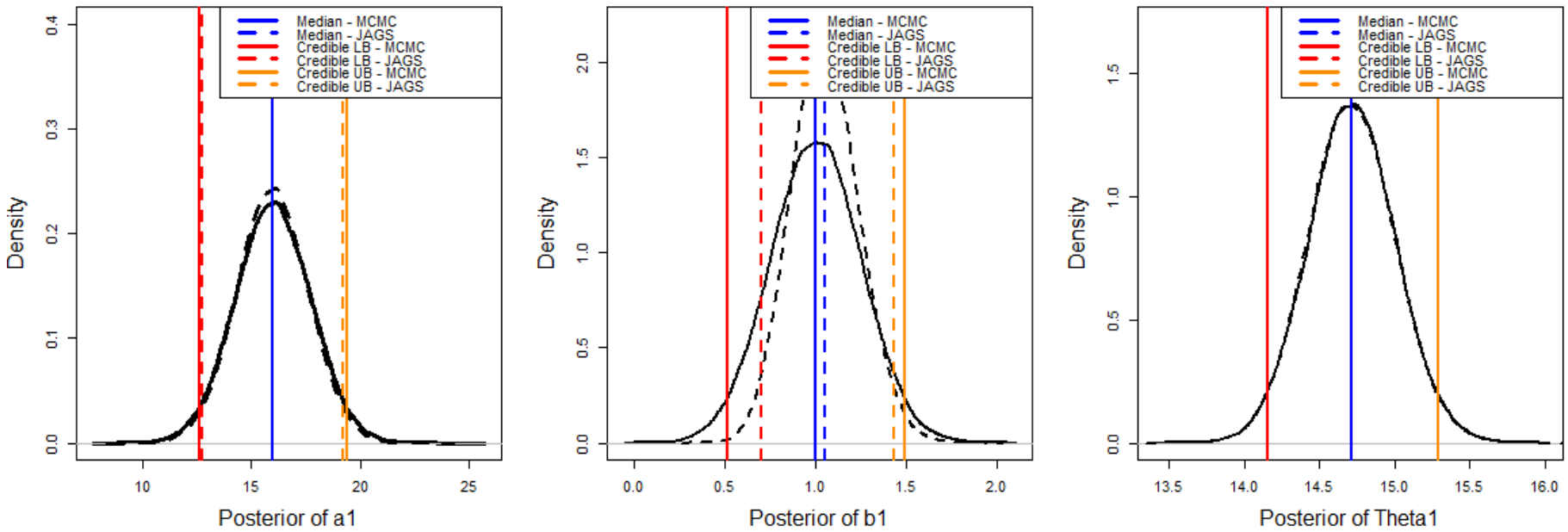

Figure 2: Posterior densities of a_1, b_1, and θ_1 which corresponds to Walker Buehler. Monte Carlo error bounds are not plotted because they are visually no different from the point estimates.

Posterior Densities

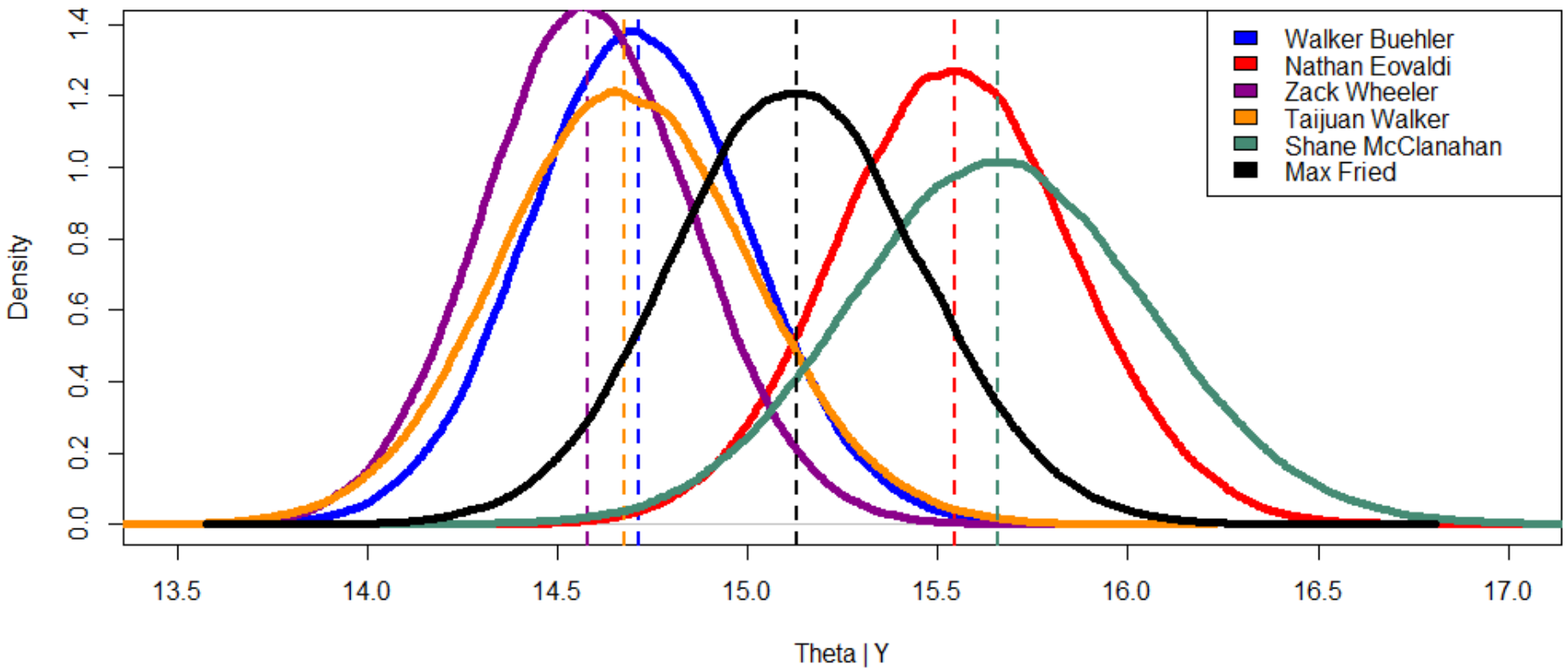

Figure 3: Posterior densities and medians of $(\theta_1, \ldots, \theta_6 | \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{y})$

Sensitivity Analysis

Alternative Priors:

1) $\theta_i \sim \text{Gamma}(a_i, b_i)$; $a_i \sim \text{N}(16, 1)$; $b_i \sim \text{N}(1, 0.01)$

2) $\theta_i \sim \text{Gamma}(a_i, b_i)$; $a_i \sim \text{N}(32, 3)$; $b_i \sim \text{N}(2, \frac{1}{16})$

3) $\theta_i \sim \text{N}(a_i, b_i)$; $a_i \sim \text{Unif}(14, 17)$; $b_i \sim \text{Unif}(0.1, 1)$

Sensitivity Analysis

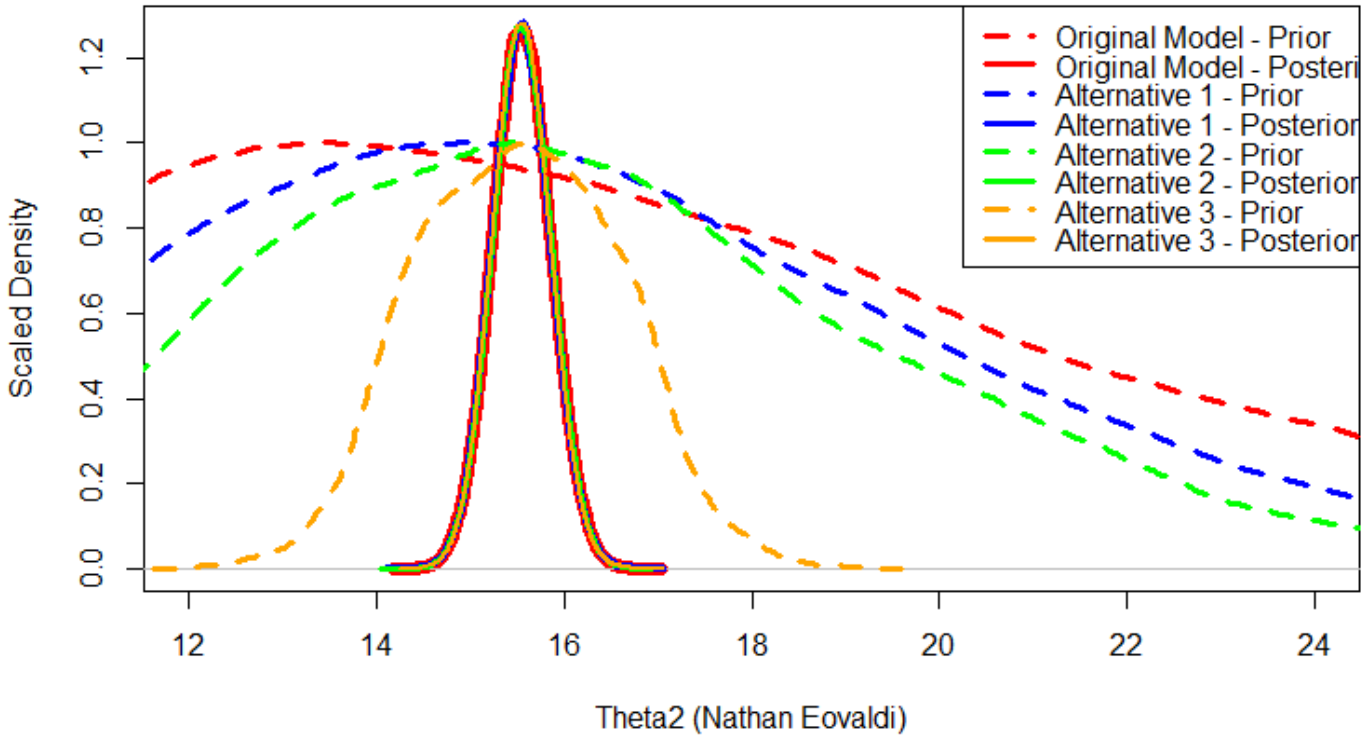

Figure 4: Four different prior distributions on θ_2 and their resulting posteriors. θ_2 corresponds to Nathan Eovaldi.

Comparison to Frequentist Methods

Casella and Berger (2001) used the sufficient statistic $W = \sum y_i$ to obtain the formula for a $(1 - \alpha)\%$ confidence interval on θ in the Poisson distribution.

$$\left\{\theta : \frac{1}{2n}\chi^2_{W,1-\alpha/2} \leq \theta \leq \frac{1}{2n}\chi^2_{2(W+1),\alpha/2}\right\}$$

Comparison to Frequentist Methods

Pitcher	Parameter	Sample Mean	95% Confidence Interval
Walker Buehler	θ_1	14.713	(14.16, 15.29)
Nathan Eovaldi	θ_2	15.544	(14.94, 16.17)
Zack Wheeler	θ_3	14.579	(14.04, 15.13)
Taijuan Walker	θ_4	14.676	(14.04, 15.33)
Shane McClanahan	θ_5	15.663	(14.90, 16.45)
Max Fried	θ_6	15.128	(14.49, 15.78)

Table 3: Point and interval estimates for $\theta_1, \ldots, \theta_6$ using maximum likelihood and the Poisson 95% confidence interval.

Applications in Prediction

1) Predict the number of pitches thrown in a new inning for one of the pitchers from the study.

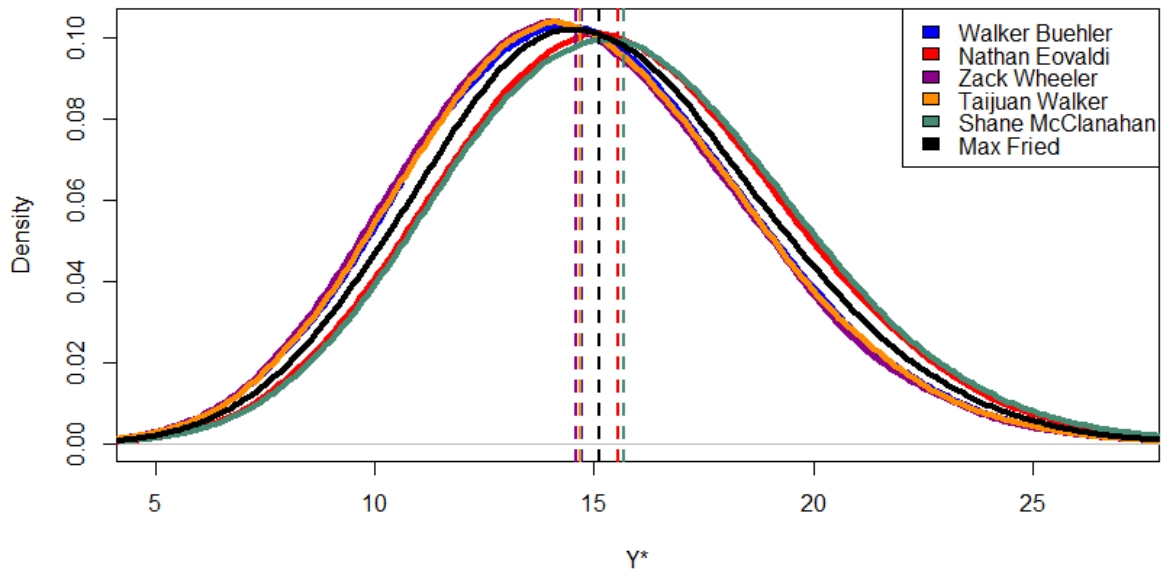

Figure 5: Posterior predictive density for each pitcher. This predictive density is perhaps the most useful application of the analysis to baseball teams and managers. The dotted lines now represent the means rather than the medians, since the median would be exactly 15 for most of the pitchers and would not convey much information visually.

Applications in Prediction: Example

The manager of the Tampa Bay Rays has Shane McClanahan at 90 pitches after 5 innings, and is trying to decide whether to send him out for the sixth. His predictive distribution could be used to estimate the probability of completing the next inning in 10 pitches or less (so as to keep him at the 100 pitch limit). In his case, this probability is estimated at 9.05%. On the other hand, the Phillies could be a little more optimistic about Zack Wheeler, for whom this probability is 14.18%.

Applications in Prediction

2) Predict the rate parameter θ_j for a new pitcher or the number of pitches they throw in a new inning.

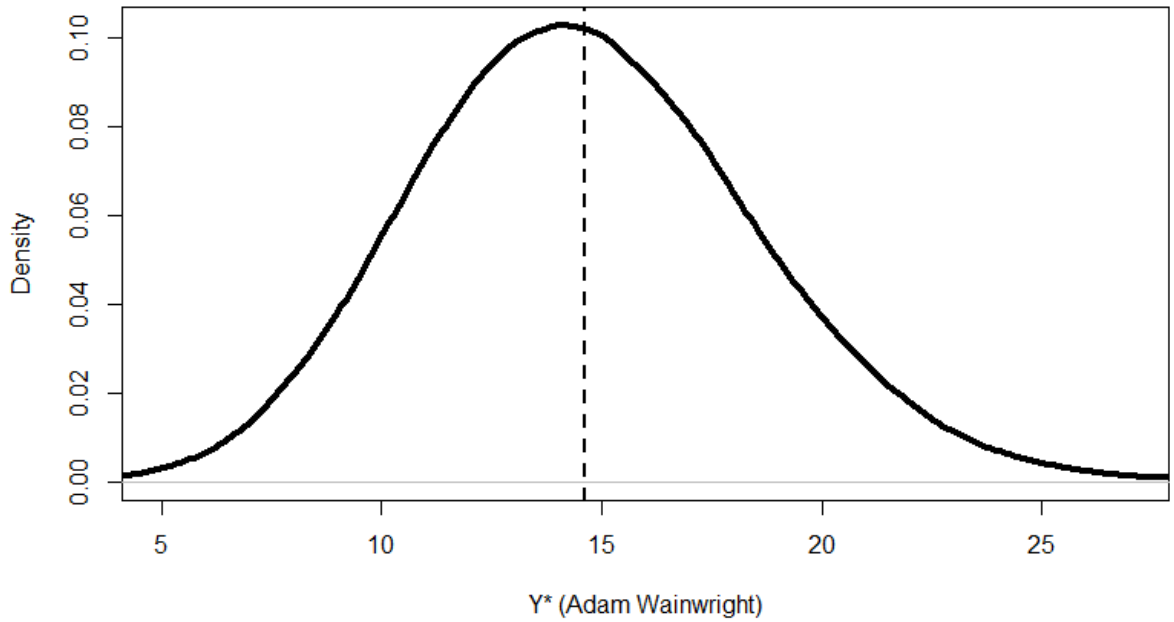

Figure 6: The posterior predictive density for Adam Wainwright of the St. Louis Cardinals, another Tommy John comeback pitcher. His average projected P/IP falls in between that of Zack Wheeler and Taijuan Walker, so he appears to be quite efficient with his pitches.

Applications of the Hyperparameters

Can be used to understand the nature of the distribution of P/IP for any pitcher.

Example:
$\hat{a}_5 = 16.06, \hat{b}_5 = 1.02$ (Shane McClanahan)
$\hat{a}_3 = 15.96, \hat{b}_3 = 1.05$ (Zack Wheeler)

$\Longrightarrow$ More uncertainty in prediction/inference on the pitch count for McClanahan than for Wheeler.

Conclusion

- Samples can be drawn from a posterior predictive distribution for any given pitcher in Major League Baseball.
  - Establishes a direct application to baseball teams trying to determine the longevity and efficiency of their starters.
- Every pitcher is different due to a variety of factors (as is every pitcher's individual start).
  - Age, climate, team, opposing team, health history, etc.
  - Future models could incorporate data on some of these covariates to assess mathematical relationships (instead of being predictive only).