Predicting Pitch Type: A Case Study in Bayesian Hierarchical Multinomial Regression

Christian Stratton
Dr. Andrew Hoegh, Dr. Jennifer Green
Department of Mathematical Sciences
Montana State University

Setting the scene

- Series tied 2-2
- Bottom of the 9^{th}, Dodgers up 4-3
- One out, runners on 1^{st} and 2^{nd}
- Kershaw comes in
- First batter pops out to second, second batter, Wilmer Difo, at a 1-1 count

2016 NLDS - Game 5 - LAD vs WSN

Motivation

Why build a pitch prediction model?

- Sign stealing controversy
  - Red Sox steal signs from Yankees using Apple Watch
- iPads in the dugout
  - Recent decision by MLB to allow iPads in the dugout
- $27.6 million to World Series Champion in 2017

Outline

- Data Collection/EDA
- Model Specification
- Results and Model Evaluation
- Conclusions

Where to find data to build a pitch prediction model?

- PITCHf/x database
  - Contains pitch-by-pitch information on every game since 2008
- *pitchRx* package
  - 20000 Clayton Kershaw pitches between 2008 and 2015
  - 15000 training observations, 5000 validation observations
- Classifier - neural network
- Covariates
  - Count
  - Previous Pitch
  - Hierarchy by catcher

EDA

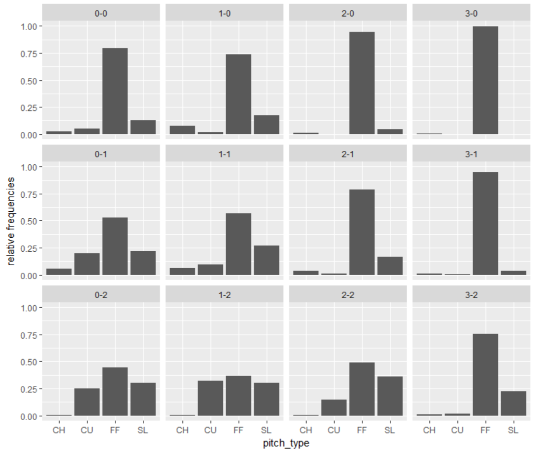


Figure: Pitch Type by Count

EDA Cont.

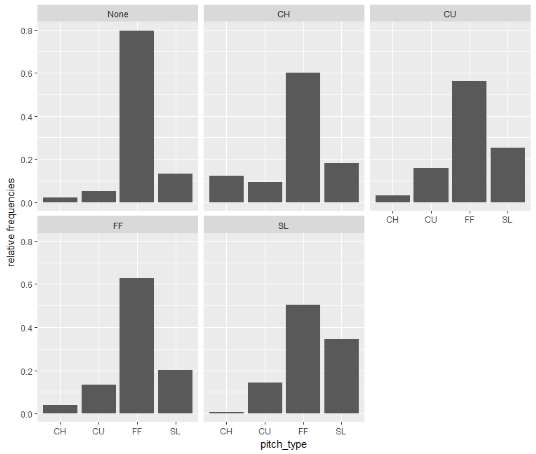


Figure: Pitch Type by Previous Pitch

Model Specification

Model 1

$$\boldsymbol{y}_1, \ldots \boldsymbol{y}_n | \boldsymbol{\pi}_i \sim \text{multinomial}(1, \boldsymbol{\pi}_i), \qquad \pi_{ij} = \frac{e^{\boldsymbol{x}_i \boldsymbol{\beta}_j}}{\sum_{k=1}^{J} e^{\boldsymbol{x}_i \boldsymbol{\beta}_k}}$$

$$\beta \sim \text{mvnormal}(\mathbf{0}, 9\boldsymbol{I})$$

Model 2

$$\boldsymbol{y}_1, \ldots \boldsymbol{y}_{n_m} | \boldsymbol{\pi}_{im} \sim \text{multinomial}(1, \boldsymbol{\pi}_{im}), \qquad \pi_{ijm} = \frac{e^{\boldsymbol{x}_i \boldsymbol{\beta}_{jm}}}{\sum_{k=1}^{J} e^{\boldsymbol{x}_i \boldsymbol{\beta}_{km}}}$$

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M | \psi \sim \text{mvnormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad \psi = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$$

$$\boldsymbol{\mu} \sim \text{mvnorm}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0), \qquad \boldsymbol{\mu}_0 = \mathbf{0}, \boldsymbol{\Lambda}_0 = 100\boldsymbol{I}$$

$$\boldsymbol{\Sigma} \sim \text{inverse-Wishart}(\eta_0, \boldsymbol{S}_0), \qquad \eta_0 = 64, \boldsymbol{S}_0 = \boldsymbol{I}$$

where i indexes the observation, j the level of the response and m the member of the hierarchy

Conditional Bayes Factor

$$BF_{21} = \frac{\int f_2(\boldsymbol{Y}_2|\boldsymbol{\theta}_2)g_2(\boldsymbol{\theta}_2|\boldsymbol{Y}_1)d\boldsymbol{\theta}_2}{\int f_1(\boldsymbol{Y}_2|\boldsymbol{\theta}_1)g_1(\boldsymbol{\theta}_1|\boldsymbol{Y}_1)d\boldsymbol{\theta}_1}$$

Strength of evidence (Kass and Raftery 1995)

2log(BF)	strength of evidence
0 to 2	not worth more than a bare mention
2 to 6	positive
6 to 10	strong
$>$ 10	very strong

For these data,

$$2log(BF_{21}) = 140.33$$

Conclusions

Things we liked:

- Overall, hierarchical component worked well
- Sampler worked, good experience
- Good foundation for more complicated models

Things we would like to consider:

- Add dynamic component
- Add spatial prediction component

What about Difo?

2016 NLDS - Game 5 - LAD vs WSN Revisited

2016 NLDS - Game 5 - LAD vs WSN Revisited

The End

References

[1] Brooks, S. P., and Gelman, A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7*, 4 (1998), 434–455.

[2] Chen, Z., and Kuo, L. A note on the estimation of the multinomial logit model with random effects. *The American Statistician 55*, 2 (May 2001), 8995.

[3] Efron, B., and Morris, C. Stein's paradox in statistics. *Scientific American 236*, 5 (1977), 119–127.

[4] Gelfand, A. E. Gibbs sampling. *Journal of the American Statistical Association 95*, 452 (2000), 1300–1304.

[5] Gelman, A., and Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science 7*, 4 (1992), 457–472.

[6] Gelman, Andrew, e. a. *Bayesian Data Analysis*, 3 ed. CRC Press, 2014.

[7] Grün, B., and Leisch, F. Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification 25*, 2 (Nov 2008), 225–247.

[8] Hoff, P. D. *A First Course in Bayesian Statistical Methods*. Springer, 2010.

[9] Madigan, D., Genkin, A., Lewis, D. D., and Fradkin, D. Bayesian multinomial logistic regression for author identification. *AIP Conference Proceedings 803*, 1 (2005), 509–516.

[10] Novy-Williams, E. Mlb teams allowed to use ipads in dugout all season. *The Chicago Tribune* (Mar 2016).

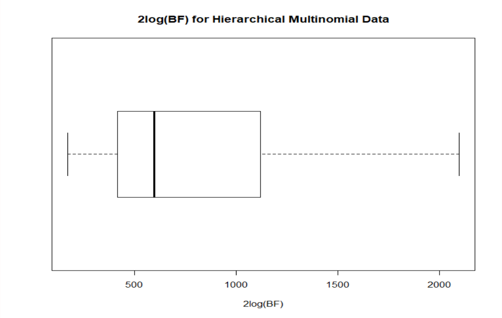
[11] Schmidt, M. S. Boston red sox used apple watch to steal steal signs against yankees. *The New York Times* (Sep 2017).

[12] Sievert, C. Taming pitchf/x data with xml2r and pitchrx. *The R Journal 6*, 1 (2014), 5 – 19.

Advantages of Hierarchical Modeling

Why introduce the complexity of hierarchical modeling?

- Anticipate differences across catchers
- Lower expected MSPE with shrinkage estimators
- Verified with simulation



Convergence Diagnostics - Model 1

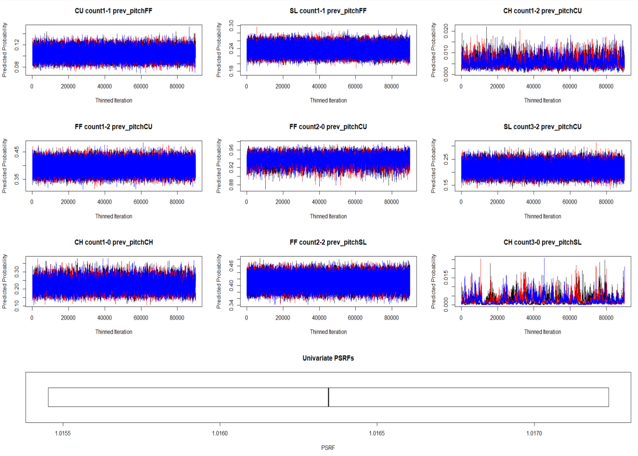


Figure: Convergence Diagnostics - Model 1

Convergence Diagnostics - Model 2

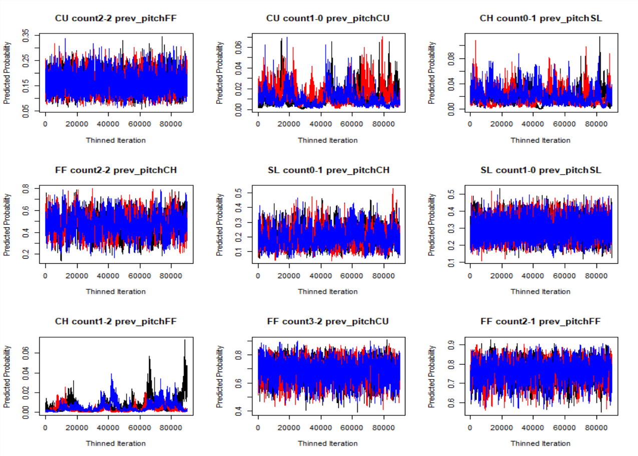


Figure: Convergence Diagnostics - Model 2

Convergence Diagnostics - Model 2 Cont.

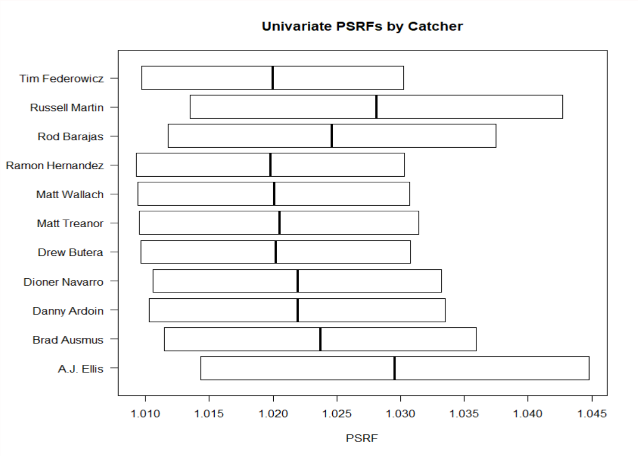


Figure: Convergence Diagnostics - Model 2