Take Me Out to (Analyze) the Ballgame

Visualization and Analysis Techniques for Big Spatial Data

Chris Comiskey

Oregon State University

August 30, 2017

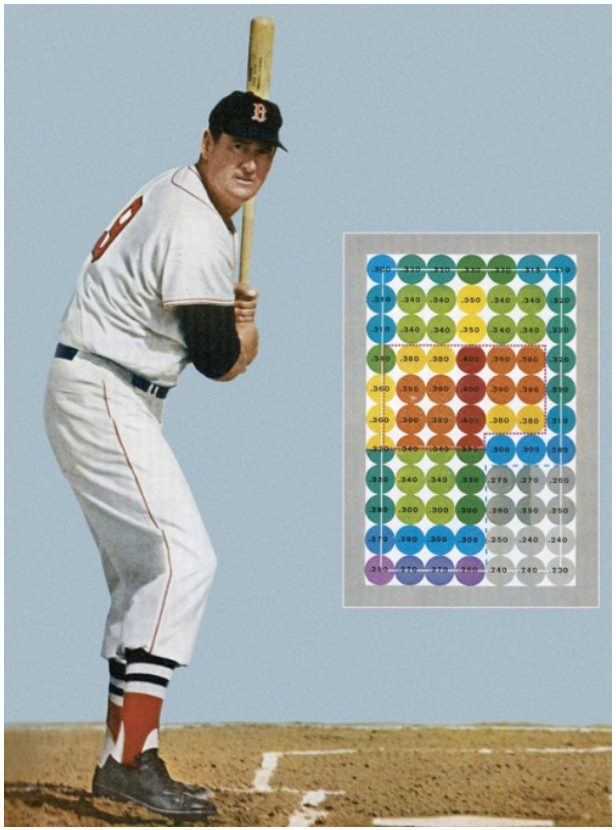
Baseball, Baseball, Baseball

- Chris, rookie year

- Chris, Boston Red Sox

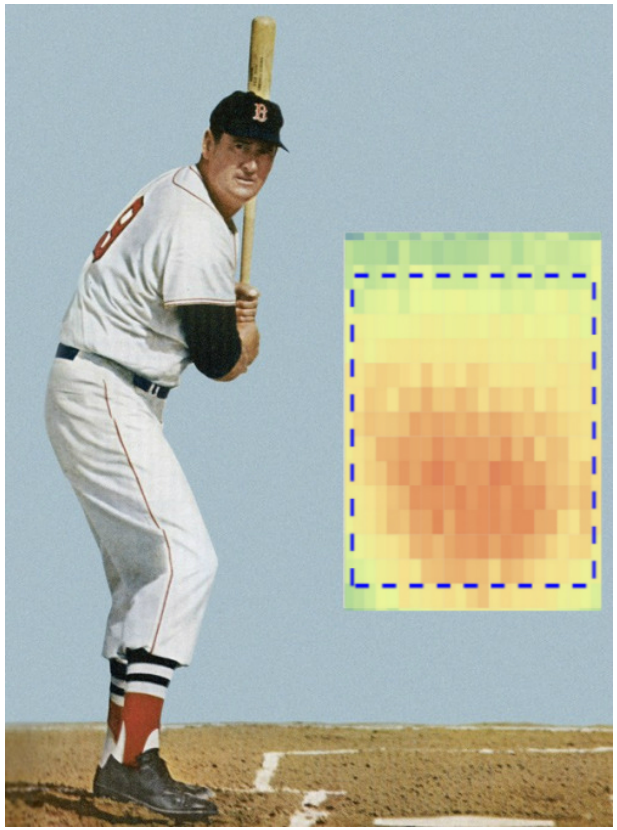
Hitting Analytics

- "The Science of Hitting"$_{1970}$



- Iconic breakthrough, hitter
- No data

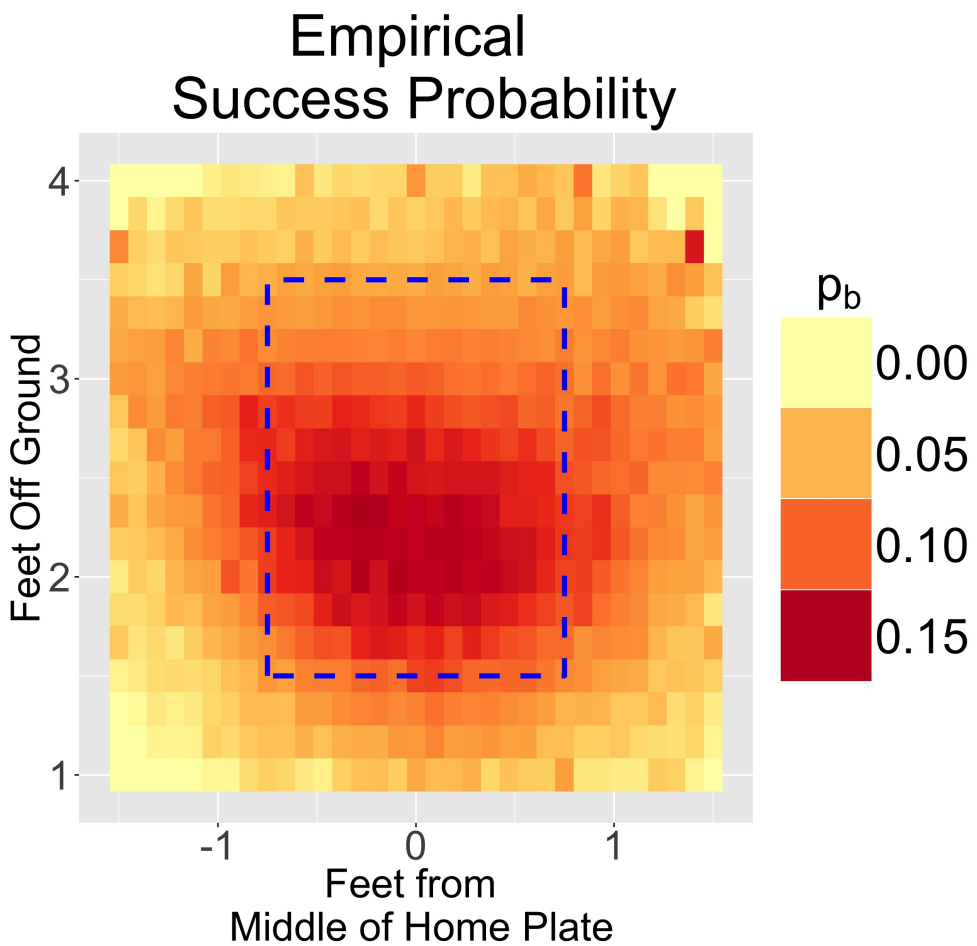
- The Statistics of Hitting



- PITCHf/x data, R, heat maps
- SGLMMs, Stan, PPMs, INLA

Outline

Empirical Success Probability Heat Map



Empirical Success Probability Heat Map

- Swings: $i = 1, 2, \ldots, N$
- Bernoulli(π_i) trials:

$$S_i = \begin{cases} 1; & \text{swing success} \\ 0; & \text{swing failure} \end{cases}$$

- Location (x_i, y_i)
- Grid boxes $G_1, G_2, \ldots, G_B$
- Box b totals:

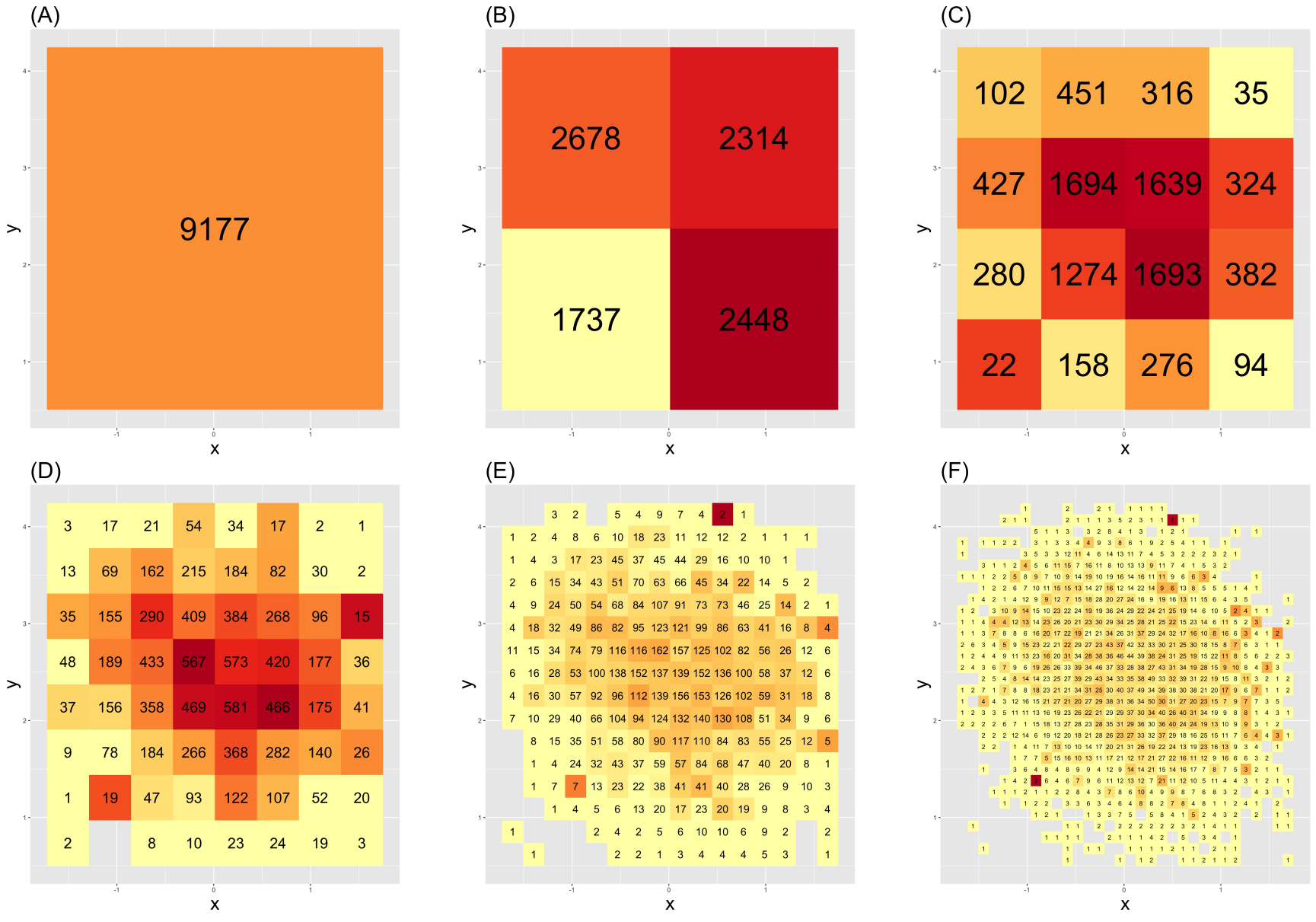
$$N_b = \sum_{i=1}^{N} I_{(x_i, y_i) \in G_b}$$

- Box b empirical success:

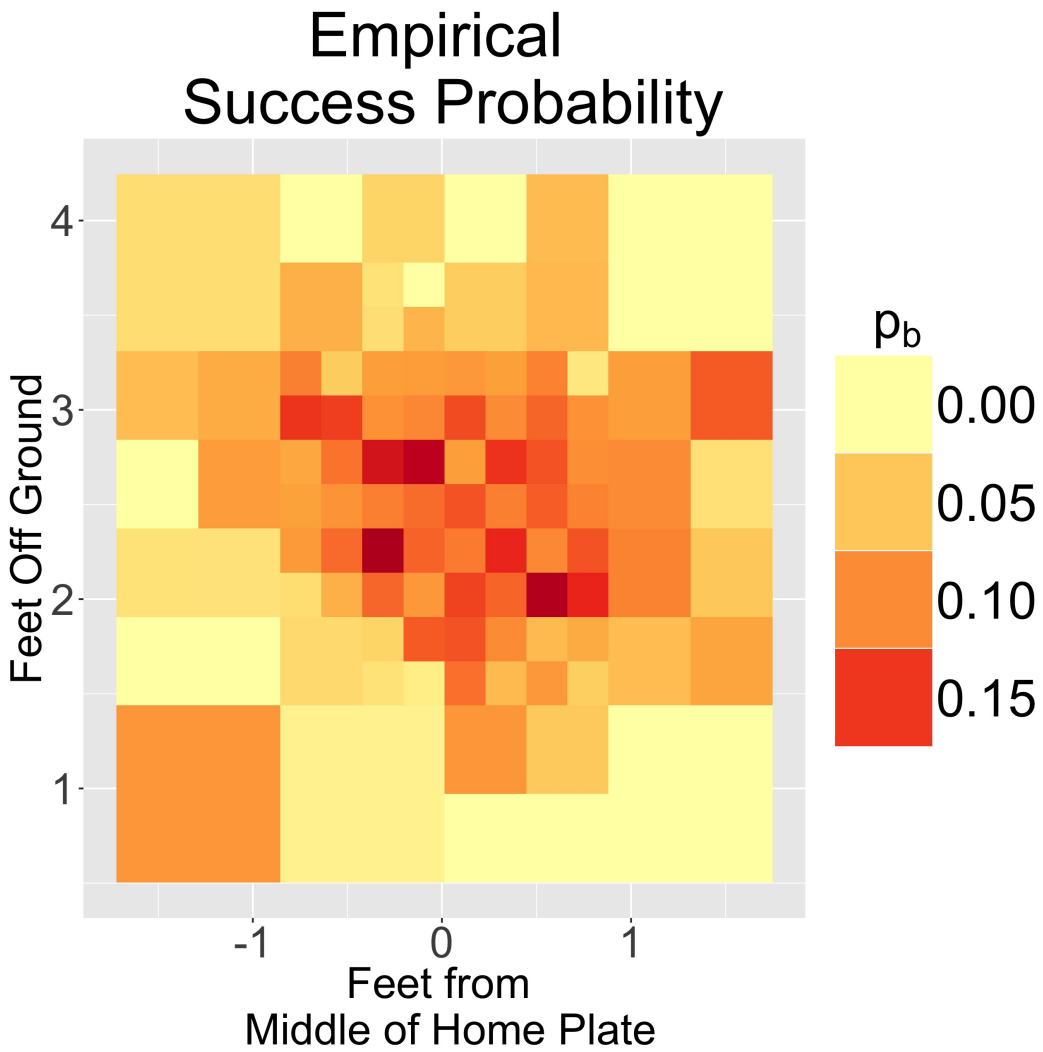
$$p_b = \frac{1}{N_b} \sum_{i=1}^{N} S_i I_{(x_i, y_i) \in G_b}$$

Heat Map Resolution Selection

Jhonny Peralta



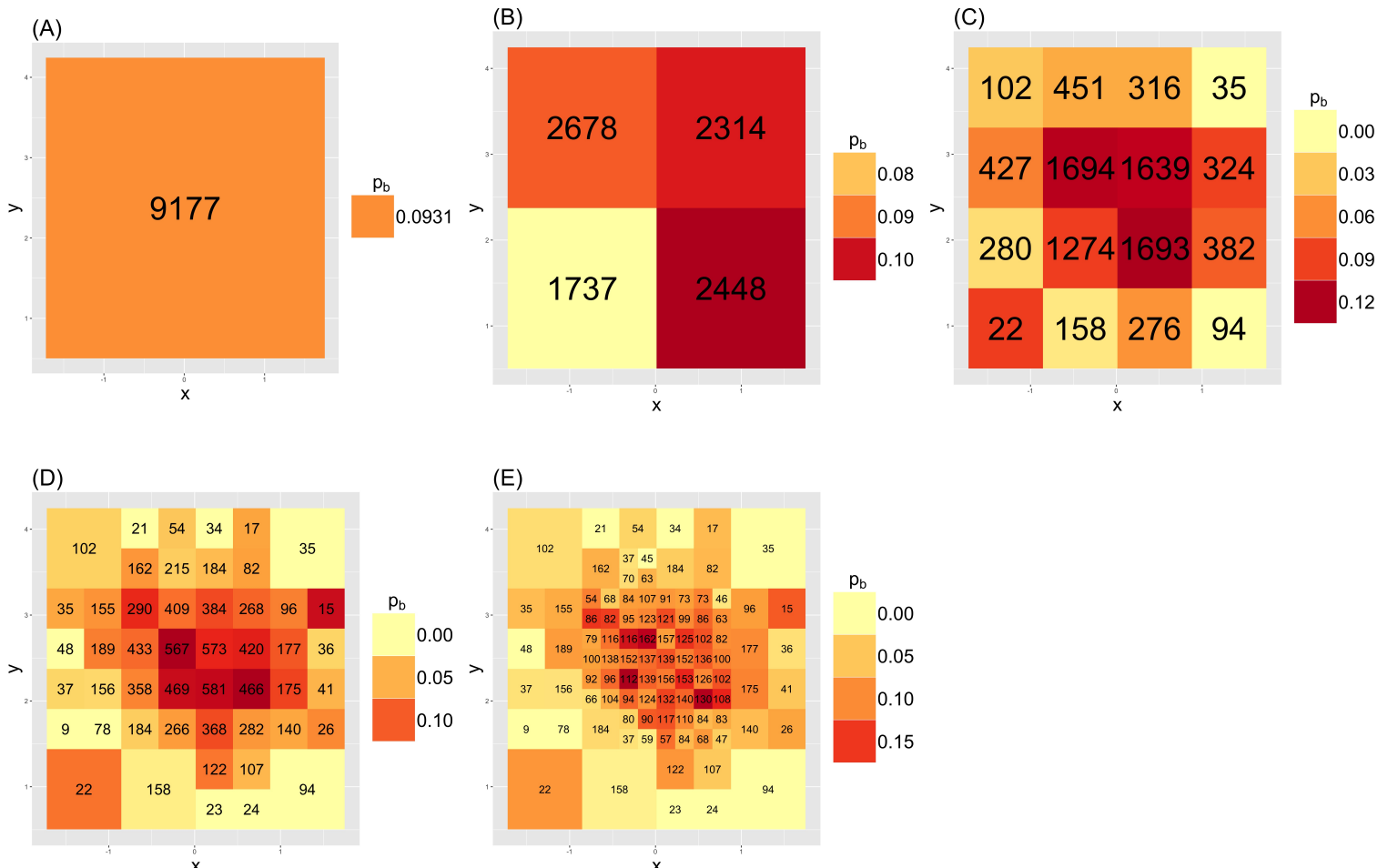
A Thing of Beauty



Variable-Resolution Heat Maps

Jhonny Peralta

- Stopping rule
- Subdivision method
- VR algorithm



Variable-Resolution Heat Maps

Combine Resolutions

- Stopping rule: sample size threshold
- Combine resolutions
- Resolution conveys data density
- Improvements:
  - Subdivision methods
  - Subdivision criteria

varyres(...) in **varyres**

varyres {varyres} R Documentation

A variable-resolution heat map generator

Description

This function creates variable resolution heat maps accroding to a stopping rule

Usage

```
varyres(dataset, cutoff, fun = mean, max = 6)
```

Arguments

`dataset`	data frame with spatial data: x-coordinates (x), y coordinates (y), and Bernoulli responses at those locations (res)
`cutoff`	Box subdivisions cease when a box sample size drops below the cutoff
`fun`	Function to apply to responses in each box
`max`	The maximum number of subdivision iterations the algorithm will perform

Value

A list containing a data frame for each iteration of the subdivision algorithm; and a vector of the number of boxes eligible for subdivision at each iteration.

Examples

```
data(hitter)
data <- varyres(hitter, mean, cutoff = 200, max = 6)
mapit(data[[4]])
```

Outline

Generalized Linear Model

$$Y_i|\mathbf{X}_i(\mathbf{s}_i) \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$$
$$\text{logit}(\pi_i|\boldsymbol{s}_i) = \mathbf{X}_i(\mathbf{s}_i)\beta$$

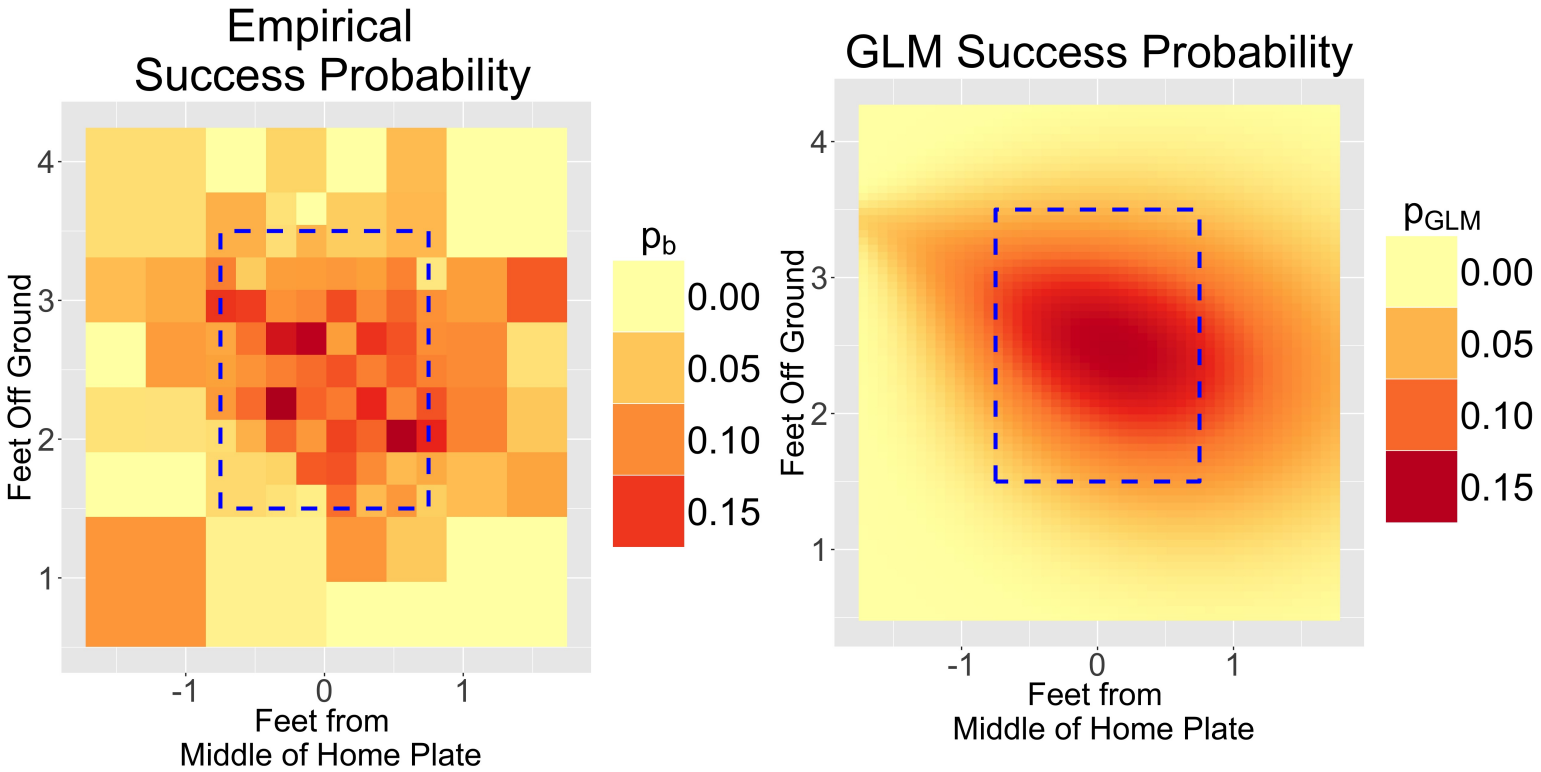
- Swings: $i = 1, 2, \ldots, N$
- Pitch location: $\mathbf{s}_i = (x_i, y_i)$
- Biomechanical covariates: $\mathbf{X}_i(\mathbf{s}_i)$
- Fit to Jhonny Peralta data

Logistic Regression Model

Jhonny Peralta

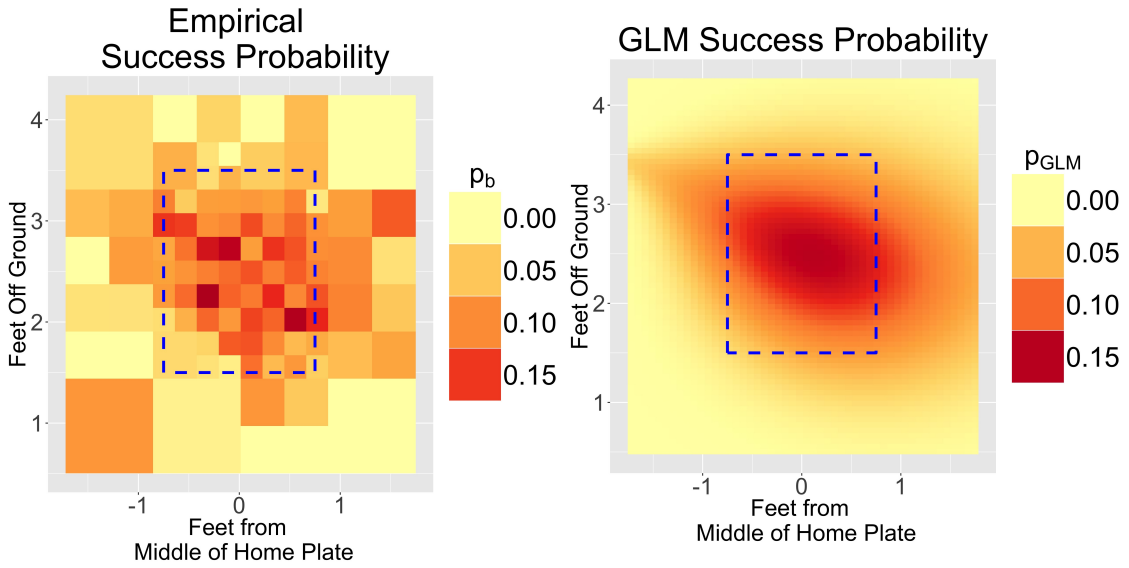
$$Y_i|\mathbf{X}_i(\mathbf{s}_i) \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$$

$$\text{logit}(\pi_i|\boldsymbol{s}_i) = \mathbf{X}_i(\mathbf{s}_i)\beta,$$



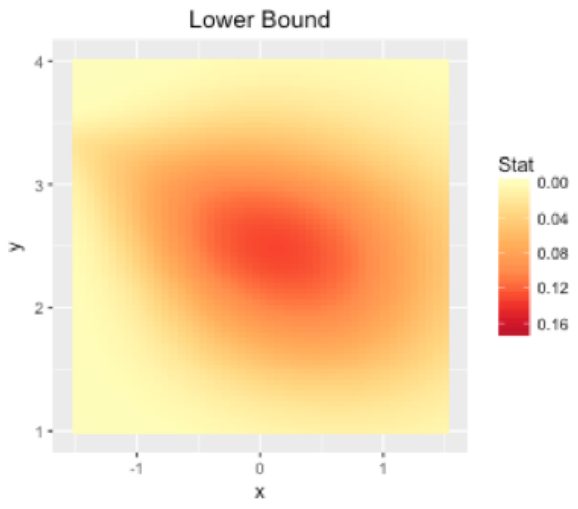
Logistic Regression Model

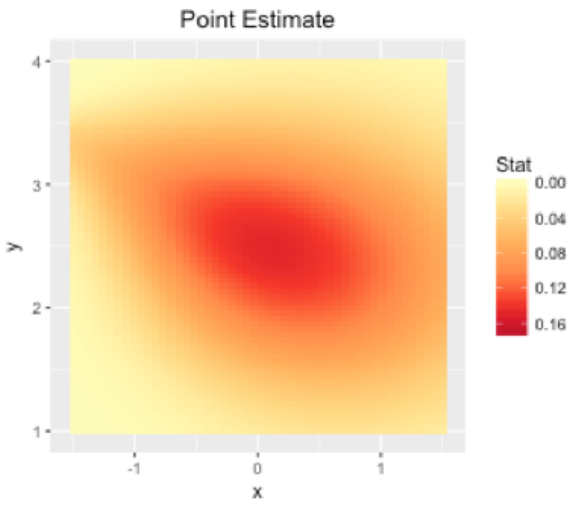
Jhonny Peralta

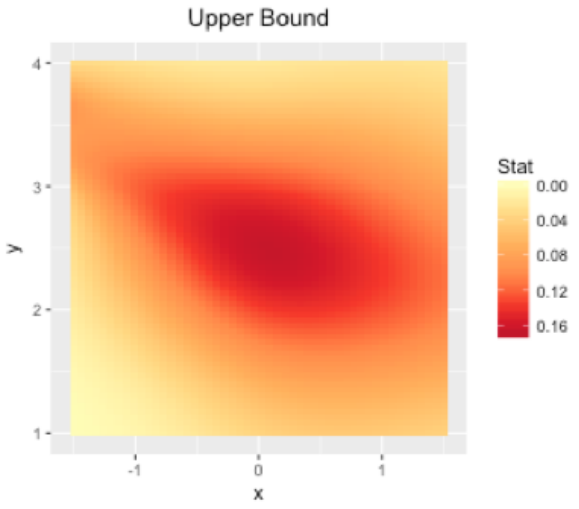


- Hosmer-Lemeshow (logistic regression) GOF test
  - H_0: Well fit
    H_A: Lack of fit
  - p-value = 0.8217
- Confidence intervals

Interactive Confidence Intervals

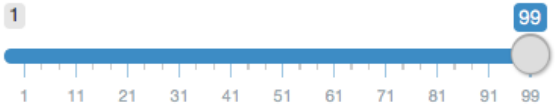






This is the 99 % confidence interval layer.

Confidence Interval



mapapp: An R Package

mapapp: An R Package

- Challenge: Where does user stop and package begin?
- Future improvement

```
all_in_one <- get_CI(model, x, y, levels)
shinyHMCI(all_in_one)
```

1. `get_CI(...)` — creates proper data structure
2. `shinyHMCI(all_in_one)` — creates Shiny app

Outline

Spatial Generalized Linear Mixed Model (SGLMM)

$$Y_i|\mathbf{X}_i(\mathbf{s}_i) \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$$
$$\text{logit}(\pi_i|\boldsymbol{s}_i) = \mathbf{X}_i(\mathbf{s}_i)\beta + w(\boldsymbol{s}_i)$$

- $w(\boldsymbol{s}_i)$ — spatial random effect, location $\boldsymbol{s}_i$.
- $\boldsymbol{w}(\boldsymbol{s}) = (w(\boldsymbol{s}_1), w(\boldsymbol{s}_2), \ldots, w(\boldsymbol{s}_N))$ — vector
- $\boldsymbol{w}(\boldsymbol{s})$ - Gaussian Random Field (GRF)

Gaussian Random Field: $\mathbf{w}(\mathbf{s})$

$$\boldsymbol{w}(\boldsymbol{s})|\boldsymbol{\theta} \sim \text{MVN}(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$$

$$\Sigma(\phi, \sigma^2)_{i,k} = \sigma^2 \exp(-||\boldsymbol{s}_i - \boldsymbol{s}_k||/\phi)$$

- Spatial exponential covariance
  - $||\boldsymbol{s}_i - \boldsymbol{s}_k||$ - Euclidean distance
  - σ^2 - scale parameter
  - ϕ - range parameter.
- Notice: $\Sigma(\boldsymbol{\theta})$ — $n \times n$

Computational Cost: “Big N” Problem

- Peralta: $n = 9177$
- $\boldsymbol{w}(\boldsymbol{s})$: 9177×9177 cov. matrix
- MCMC iterations require: Σ^{-1}, determinant of Σ
- $\mathcal{O}(n^3)$ rate of increase:

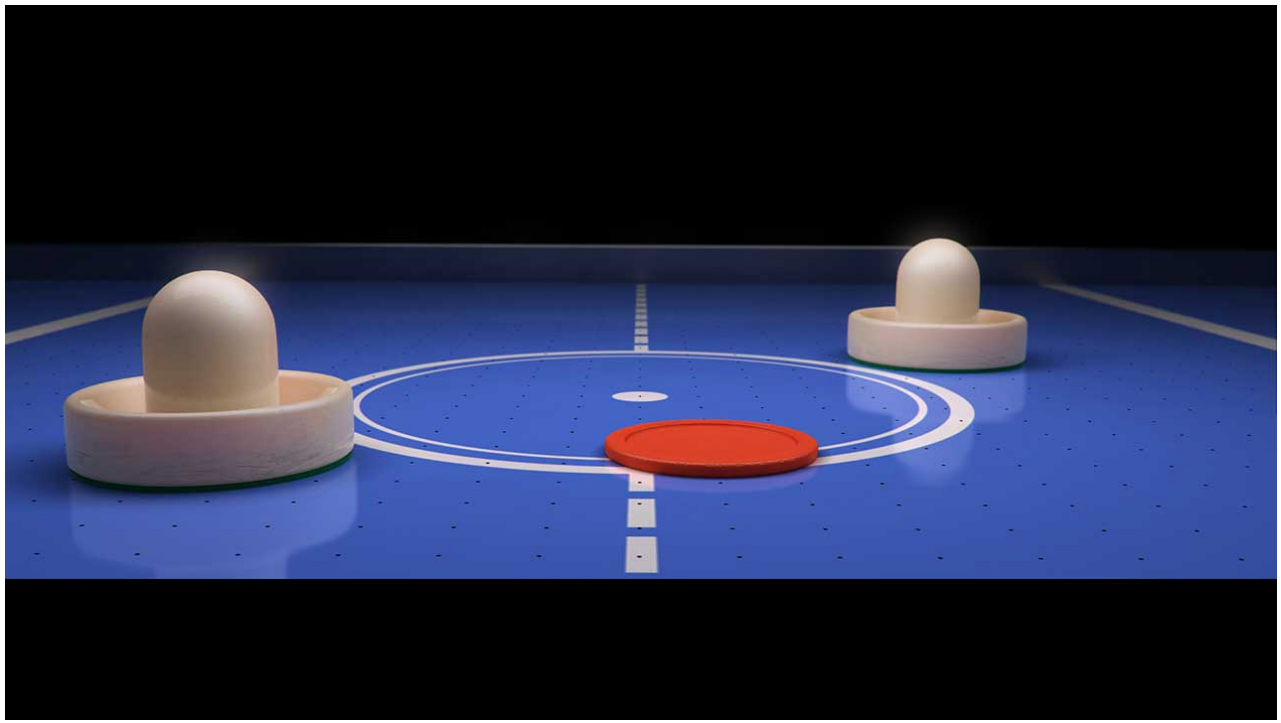
$$t(n) \leq M \cdot n^3$$

$$\text{as } n \to \infty$$

Outline

Hamiltonian Monte Carlo (HMC)

- Stan uses Hamiltonian proposal mechanism
- **Short version**: disk on surface, randomly sample momentum (auxiliary), calculate new position (parameters) — that's your Metropolis proposal.



Stan Computational Optimization

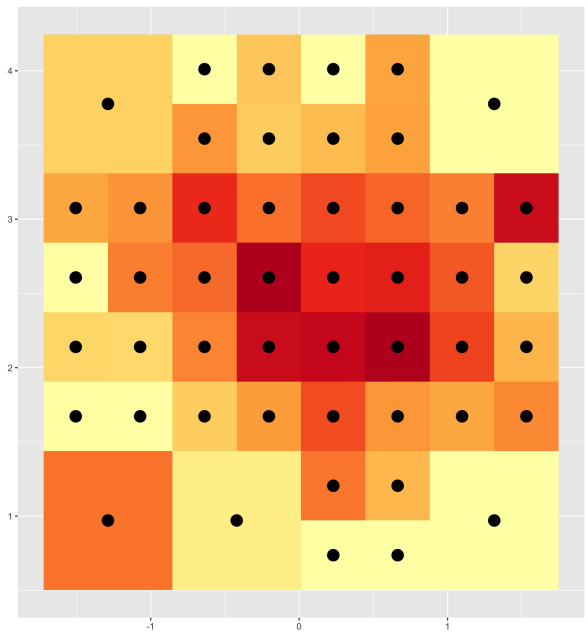
- ϕ, β: informative/proper priors $\rightarrow$ identifiability/cost/convergence
- QR factorization of X, Cholesky decomposition of $\Sigma(\theta)$
- Matrices & vectors faster than loops & scalars
- n = 25: 40 seconds $\rightarrow$ 3 seconds
- n = 2000 — overnight, 350 draws

Outline

Predictive Process Models (PPMs)

[Banerjee et al., 2008]

- Knots: $\boldsymbol{S}^* = \{\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_m^*\}$
  - $m << n$



$$\text{logit}(\pi_i|\boldsymbol{s}_i) = \boldsymbol{X}_i(\boldsymbol{s}_i)\beta + \widetilde{w}(\boldsymbol{s}_i)$$

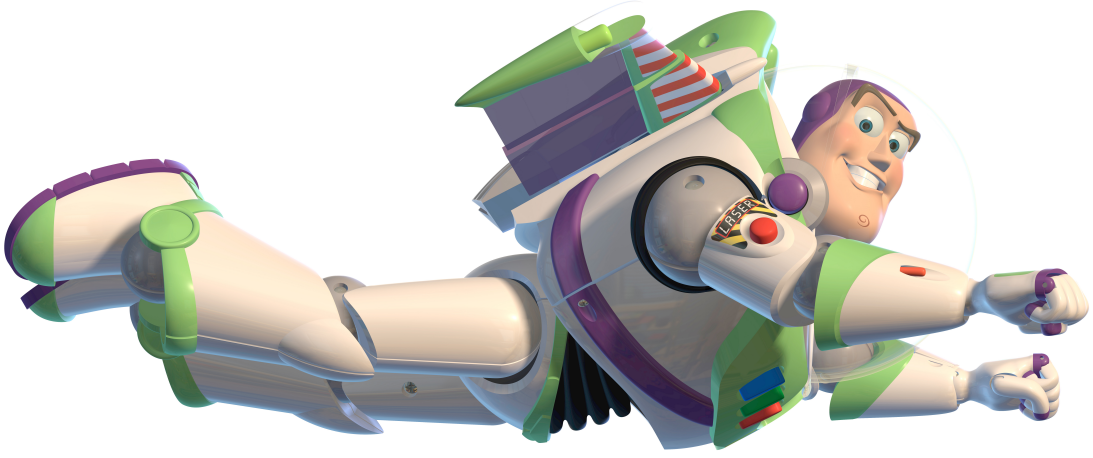
$$\widetilde{\boldsymbol{w}}(\boldsymbol{s}) \sim \text{MVN}\{0, \widetilde{\Sigma}(\boldsymbol{\theta})\}$$

$$\widetilde{\Sigma}(\boldsymbol{\theta})_{i,j} = \boldsymbol{\sigma}^{*T}(\boldsymbol{s}_i;\boldsymbol{\theta}) \cdot \boldsymbol{\Sigma}^{*-1}(\boldsymbol{\theta}) \cdot \boldsymbol{\sigma}^*(\boldsymbol{s}_j;\boldsymbol{\theta})$$

- $\boldsymbol{\sigma}^*(\boldsymbol{s}_i;\boldsymbol{\theta}) = \text{Cov}(\boldsymbol{s}_i, \boldsymbol{S}^*)$
- $\boldsymbol{\Sigma}^*(\boldsymbol{\theta}) = \text{Var}(\boldsymbol{S}^*)$

PPM Results

- Implement in **spBayes** [Finley et al., 2013].
- MCMC chains did not converge.
  - n = 1000, knots = 97, 10K samples, $\approx$ 6.7 mins
  - n = 1000, knots = 49, 30K samples, $\approx$ 7 mins
  - n = 3000, knots = 49, 80K samples, $\approx$ 54 mins
- Speed issue for extending MCMC chains
- Buzz Lightyear is fast.

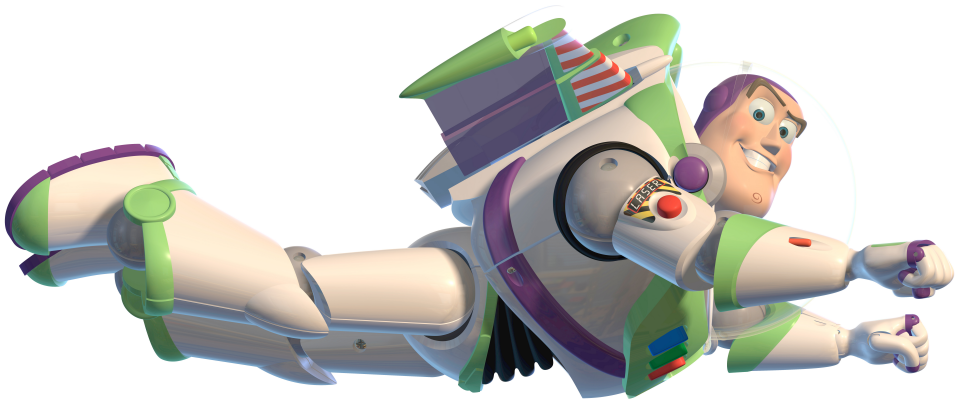


Outline

INLA Overview

$$\text{logit}(\pi_i) = \boldsymbol{X}_i(\boldsymbol{s}_i)\beta + \boldsymbol{w}(\boldsymbol{s}_i)$$
$$\boldsymbol{w}(\boldsymbol{s})|\boldsymbol{\theta} \sim \text{MVN}(\boldsymbol{0}, \Sigma(\boldsymbol{\theta}))$$

- Bayesian hierarchical models w/ latent GRF
- Assume Matérn covariance
- Two parts (continuous domain)
  - Part 1: Represent GRF as GMRF (SPDE)
  - Part 2: Integrated Nested Laplace Approximation (INLA)



Integrated Nested Laplace Approximation

[Rue et al., 2009]

- Parameter vector: $\boldsymbol{\rho} = (\beta^{\mathsf{T}}, \widetilde{\boldsymbol{w}}^{\mathsf{T}})^{\mathsf{T}}$
  - $\boldsymbol{Q}$: precision matrix of $\boldsymbol{\rho}$
- Hyperparameter vector: $\boldsymbol{\theta} = (\kappa, \sigma)$

1. Gaussian approximation:

$$p(\boldsymbol{\rho}|\boldsymbol{\theta}, \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\rho}, \boldsymbol{\theta}) p(\boldsymbol{\rho}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

$$p(\boldsymbol{\rho}|\boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\rho}^T \boldsymbol{Q} \boldsymbol{\rho} + \sum_i \log p(y_i|\boldsymbol{\rho}, \boldsymbol{\theta})\right)$$

$$p_G(\boldsymbol{\rho}|\boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\rho} - \boldsymbol{\mu})^T (\boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c}))(\boldsymbol{\rho} - \boldsymbol{\mu})\right)$$

   - $\boldsymbol{c}$ and $\boldsymbol{\mu}$ depend on second order Taylor expansions of $f(\boldsymbol{\rho}) = \sum_i \log p(y_i|\boldsymbol{\rho}, \boldsymbol{\theta})$

Integrated Nested Laplace Approximation

[Rue et al., 2009]

- Fact: $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{p(\boldsymbol{y}|\boldsymbol{\rho}, \boldsymbol{\theta})p(\boldsymbol{\rho}|\boldsymbol{\theta})}{p(\boldsymbol{\rho}|\boldsymbol{y}, \boldsymbol{\theta})}$$

- Bayes proportionality:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto \frac{p(\boldsymbol{y}|\boldsymbol{\rho}, \boldsymbol{\theta})p(\boldsymbol{\rho}|\boldsymbol{\theta})}{p(\boldsymbol{\rho}|\boldsymbol{y}, \boldsymbol{\theta})} \cdot p(\boldsymbol{\theta}) \end{aligned}$$

2. For a given $\boldsymbol{\theta}$, let $\boldsymbol{\rho}_0 = \text{argmax}_{\rho} p(\boldsymbol{\rho}|\boldsymbol{y}, \boldsymbol{\theta})$. Then,

$$\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{p(\boldsymbol{y}|\boldsymbol{\rho}_0, \boldsymbol{\theta})p(\boldsymbol{\rho}_0|\boldsymbol{\theta})}{p_G(\boldsymbol{\rho}_0|\boldsymbol{y}, \boldsymbol{\theta})} \cdot p(\boldsymbol{\theta})$$

Integrated Nested Laplace Approximation

[Rue et al., 2009]

3. Numerical integration:

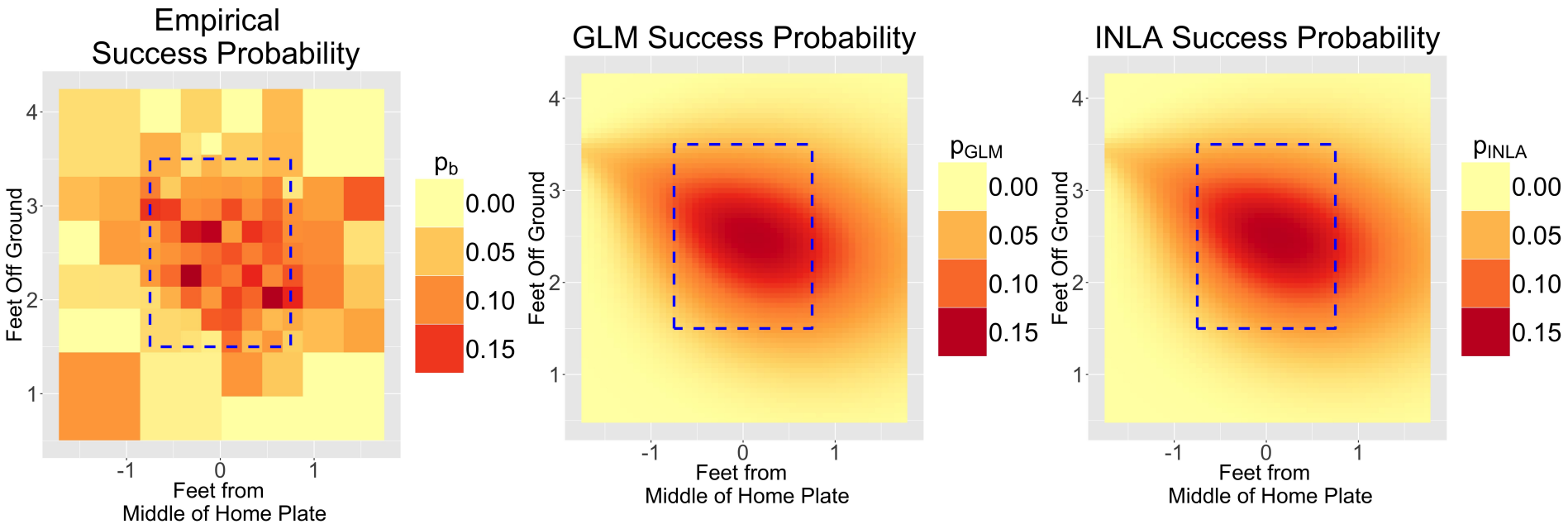
$$p(\rho_j|\boldsymbol{y}) \approx \int p_{\mathrm{G}}(\rho_j|\boldsymbol{\theta},\boldsymbol{y})\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

4. Numerical integration:

$$p(\theta_k|\boldsymbol{y}) \approx \int \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-k}$$

INLA Model Fit: 34 seconds

34 seconds!



Estimates

GLM			
Covariate	β_i	MLE	SE
1	β_0	-4.08	0.70
r	β_1	1.19	0.51
θ	β_2	-1.93	1.90
$r * \theta$	β_3	-1.64	0.70
r^2	β_4	-0.32	0.09
θ^2	β_5	-3.92	1.10
$r^2 * \theta^2$	β_6	-0.46	0.21

INLA			
Covariate	ρ_i	$\hat{\rho}_i$	SE
N/A	κ	3.23	$\pm$ 1 SE: (1.35, 7.54)
N/A	σ	0.11	$\pm$ 1 SE: (0.05, 0.26)
1	β_0	-4.14	0.76
r	β_1	1.25	0.55
θ	β_2	-1.90	1.96
$r * \theta$	β_3	-1.70	0.93
r^2	β_4	-0.33	0.10
θ^2	β_5	-3.93	1.14
$r^2 * \theta^2$	β_6	-0.48	0.22

- $\hat{\kappa}$ — long range correlation
- $\text{SE}(w(\boldsymbol{s})) = 0.11$.
- $\rightarrow 0.15 \pm 1 \cdot \text{SE} = (0.137, 0.165)$
- $\rightarrow 0.15 \pm 2 \cdot \text{SE} = (0.125, 0.181)$

Summary

- Resolution selection? Variable-resolution heat maps and **varyres**
- Heat map confidence intervals? Interactive HMCIs and **mapapp**
- Fitting big data SGLMMs to baseball data?
  - Stan - insufficient
  - PPM - Slow and did not converge; to be continued...
  - INLA - Fast, successful; to be continued...

Thanks for listening.

Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

Andrew O Finley, Sudipto Banerjee, and Alan E Gelfand. spbayes for large univariate and multivariate point-referenced spatio-temporal data models. *arXiv preprint arXiv:1310.8192*, 2013.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

Moving Forward

- Improve packages
- Submit packages to CRAN
- Biomechanists at OSU
- Compare GLM and SGLMM with scoring rules
- Exit velocity as response variable
- Independence assumption, pitch sequences
- Randomized pitch selection