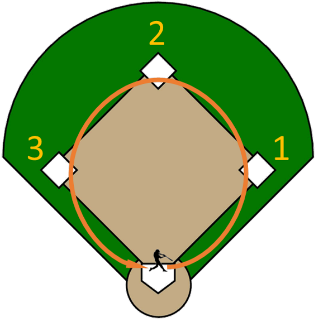
Nuclear penalized multinomial regression

Scott Powers
(Joint work with Trevor Hastie and Rob Tibshirani)

Rice Statistics Colloquium
April 17, 2023

RICE SOCIAL SCIENCES
Department of Sport Management

Background



- Game is a sequence of matchups between one pitcher and one batter
- Batters try to advance along bases before recording 3 outs
- Markov chain model works well
  - State space: {bases, outs}

(Essentially) 9 possible outcomes of each batter-pitcher matchup:

HR	3B	2B	1B	HBP	BB	K	G	F
1.38	1.06	0.76	0.46	0.35	0.33	-0.27	-0.27	-0.27

Predicting player performance

Baseball analysis blogs (early 2010s):

- "Stabilization rate":
  Sample size at which between-sample correlation is 0.7
  - wOBA: $n = 350$
  - K%: $n = 60$
  - BABIP: $n = 1200$

$$\text{BABIP} = \frac{1\text{B} + 2\text{B} + 3\text{B}}{1\text{B} + 2\text{B} + 3\text{B} + \text{G} + \text{F}}$$

Powers and Shayer (SABR Analytics 2016):

- Multinomial regression with ridge penalty
  - Two categorical variables (batter and pitcher), plus controls
  - Adjusts player performance for competition and sample size
  - Similar to random-effect models and Bayesian models

Notation

- $i = 1, ..., n$, indexes plate appearances (PA)
- Within the i^{th} PA, ...
  - Batter $B_i \in \mathcal{B} = \{$Mike Trout, ..., Zach Cozart$\}$
  - Pitcher $P_i \in \mathcal{P} = \{$Clayton Kershaw, ..., Zack Britton$\}$
  - Outcome $y_i \in \mathcal{O} = \{$F, G, K, BB, HBP, 1B, 2B, 3B, HR$\}$

Model:

$$\mathbb{P}(Y_i = k) = \frac{e^{\eta_{ik}}}{\sum_{k' \in \mathcal{O}} e^{\eta_{ik'}}}$$

$$\eta_{ik} = \alpha_k + \beta_{k:B_i} + \gamma_{k:P_i}$$

Objective:

$$\underset{\alpha,\beta,\gamma}{\text{minimize}} - \sum_{i=1}^{n} \log \mathbb{P}(Y_i = y_i) + \lambda \sum_{k \in \mathcal{O}} \left(\sum_{B \in \mathcal{B}} \beta_{k:B}^2 + \sum_{P \in \mathcal{P}} \gamma_{k:P}^2 \right)$$

Matrix notation

$$\mathbf{X} = \begin{array}{c} \begin{array}{cc} \overbrace{\hspace{3cm}}^{\text{Batters}} & \overbrace{\hspace{3cm}}^{\text{Pitchers}} \end{array} \\ \begin{pmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 1 \\ 0 & \dots & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix} \end{array} \qquad \mathbf{B} = \begin{pmatrix} \beta_{F:MT} & \dots & \beta_{HR:MT} \\ .. & \dots & .. \\ \beta_{F:ZC} & \dots & \beta_{HR:ZC} \\ \gamma_{F:CK} & \dots & \gamma_{HR:CK} \\ .. & \dots & .. \\ \gamma_{F:ZB} & \dots & \gamma_{HR:ZB} \end{pmatrix}$$

$$\underbrace{n \times (|\mathcal{B}| + |\mathcal{P}|)}_{n \times p} \qquad \underbrace{(|\mathcal{B}| + |\mathcal{P}|) \times |\mathcal{O}|}_{p \times K}$$

$$\underset{\alpha \in \mathbb{R}^K,\ \mathbf{B} \in \mathbb{R}^{p \times K}}{\text{minimize}} \underbrace{-\sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \frac{e^{\alpha_k + \mathbf{X}\mathbf{b}_k}}{\sum_{k'=1}^{K} e^{\alpha_{k'} + \mathbf{X}\mathbf{b}_{k'}}} \mathbb{I}_{\{y_i = k\}} \right)}_{= \ell(\alpha, \mathbf{B}; \mathbf{X}, \mathbf{Y})} + \lambda ||\mathbf{B}||_F^2$$

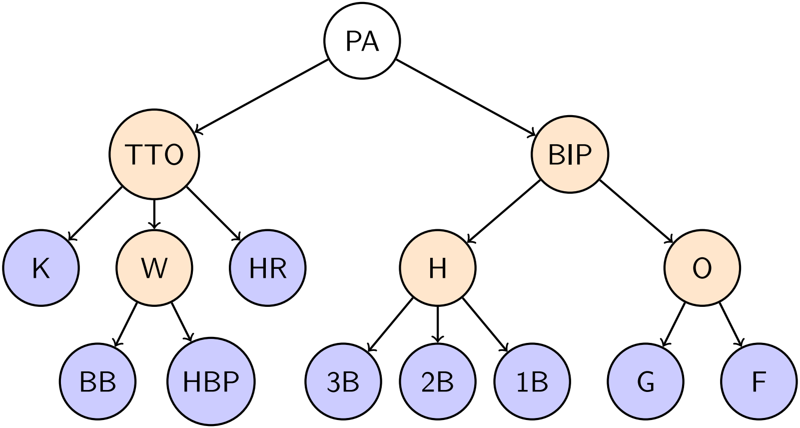
- \# parameters: $K + p \times K = 9 + 764 \times 9 = 6885$

Structure between plate appearance outcomes

Ordering:

$$K < G < F < BB < HBP < 1B < 2B < 3B < HR$$

Structure:



Adapted from Baumer & Zimbalist (2014)

Principal component analysis of observed rates

Batters:

Principal component	1	2	3	4	5	6	7	8	9
F	-0.2	0.7	0.5	-0.1	0.3	0.0	-0.1	0.1	-0.3
G	-0.5	-0.6	0.4	-0.3	0.1	-0.0	-0.1	0.1	-0.3
K	0.8	-0.3	0.3	0.2	0.2	0.1	-0.1	0.1	-0.3
BB	0.1	0.1	-0.6	-0.6	0.4	0.0	-0.1	0.1	-0.3
HBP	0.0	0.0	-0.0	0.0	-0.1	-0.1	0.9	0.1	-0.3
1B	-0.3	-0.0	-0.4	0.7	0.3	-0.1	-0.1	0.1	-0.3
2B	-0.0	0.1	-0.1	0.0	-0.5	0.7	-0.1	0.1	-0.3
3B	-0.0	-0.0	-0.0	0.0	-0.0	0.0	0.0	-0.9	-0.3
HR	0.1	0.1	-0.0	-0.1	-0.6	-0.6	-0.3	0.1	-0.3
% Variance explained	51.1	29.0	8.7	7.2	2.2	1.0	0.6	0.2	0.0

Pitchers:

Principal component	1	2	3	4	5	6	7	8	9
F	-0.3	-0.7	0.3	0.3	0.3	0.1	0.1	0.1	-0.3
G	0.7	0.2	0.4	0.3	0.1	0.1	0.1	0.1	-0.3
K	-0.6	0.7	0.3	-0.0	0.1	0.1	0.1	0.1	-0.3
BB	-0.0	0.1	-0.8	0.3	0.3	0.1	0.2	0.1	-0.3
HBP	0.0	0.0	-0.0	0.0	-0.0	-0.0	-0.9	0.1	-0.3
1B	0.2	-0.1	-0.0	-0.8	0.3	0.1	0.1	0.1	-0.3
2B	0.0	-0.1	-0.1	-0.1	-0.8	0.4	0.1	0.1	-0.3
3B	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.9	-0.3
HR	-0.0	-0.1	-0.0	0.0	-0.2	-0.9	0.2	0.1	-0.3
% Variance explained	52.9	32.7	6.7	4.9	1.5	0.6	0.3	0.2	0.0

Reduced-rank multinomial regression

$$\begin{aligned} \underset{\alpha \in \mathbb{R}^K,\ \mathbf{B} \in \mathbb{R}^{p \times K}}{\text{minimize}} \quad & -\ell(\alpha, \mathbf{B}; \mathbf{X}, \mathbf{Y}) \\ \text{subject to} \quad & \text{rank}(\mathbf{B}) \le r \end{aligned}$$

$$\Rightarrow \exists\ \mathbf{A} \in \mathbb{R}^{p \times r}, \mathbf{C} \in \mathbb{R}^{K \times r} \text{ s.t. } \mathbf{B} = \mathbf{A}\mathbf{C}^T$$

- CRAN package `VGAM` implements reduced-rank vector generalized linear models (RR-VGLMs, Yee and Hastie, 2003)
- Far too slow to fit on full season of MLB play-by-play data
    - Not a convex optimization problem

Nuclear penalized multinomial regression

NPMR:

$$\underset{\alpha\in\mathbb{R}^K,\ \mathbf{B}\in\mathbb{R}^{p\times K}}{\text{minimize}} -\ell(\alpha, \mathbf{B}; \mathbf{X}, \mathbf{Y}) + \lambda||\mathbf{B}||_*$$

$$||\mathbf{B}||_* = \sum_{r=1}^{\text{rk}(\mathbf{B})} \sigma_r$$

- Convex relaxation of reduced-rank regression
  (Just as the lasso is a convex relaxation of best subset regression!)

How to solve it?

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times K}}{\text{minimize}} - \ell(\mathbf{B};\mathbf{X},\mathbf{Y}) + \lambda||\mathbf{B}||_*$$

1. Alternating direction method of multipliers (ADMM)

   Variable splitting:

$$\begin{aligned} \underset{\mathbf{B},\mathbf{C}\in\mathbb{R}^{p\times K}}{\text{minimize}} & - \ell(\mathbf{B};\mathbf{X},\mathbf{Y}) + \lambda||\mathbf{C}||_* \\ \text{subject to} & \;\; \mathbf{B} - \mathbf{C} = 0 \end{aligned}$$

2. Proximal gradient descent (PGD)

Proximal gradient descent

- Gradient descent:

$$\mathbf{B}^{(t+1)} = \mathbf{B}^{(t)} - s\nabla f(\mathbf{B}^{(t)})$$
$$= \underset{\mathbf{B}\in\mathbb{R}^{p\times K}}{\arg\min} \left\{ f(\mathbf{B}^{(t)}) + \left\langle \nabla f(\mathbf{B}^{(t)}), \mathbf{B} - \mathbf{B}^{(t)} \right\rangle + \frac{1}{2s}||\mathbf{B} - \mathbf{B}^{(t)}||_F^2 \right\}$$

- **Proximal** gradient descent (PGD)

$$\mathbf{B}^{(t+1)} = \underset{\mathbf{B}\in\mathbb{R}^{p\times K}}{\arg\min} \left\{ g(\mathbf{B}^{(t)}) + \left\langle \nabla g(\mathbf{B}^{(t)}), \mathbf{B} - \mathbf{B}^{(t)} \right\rangle + \frac{1}{2s}||\mathbf{B} - \mathbf{B}^{(t)}||_F^2 + h(\mathbf{B}) \right\}$$

Proximal operator: $\mathbf{prox}_h(z) \equiv \underset{\theta}{\arg\min} \left\{ \frac{1}{2}||z - \theta||_2^2 + h(\theta) \right\}$

$$= \mathbf{prox}_{sh}\left(\mathbf{B}^{(t)} - s\nabla g(\mathbf{B}^{(t)})\right)$$

Proximal gradient descent

Proximal map for nuclear norm: soft-thresholding singular values

$$\mathbf{B} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \qquad [\mathcal{S}_{s\lambda}(\boldsymbol{\Sigma})]_{rr} = (\sigma_r - s\lambda)_+$$

$$\mathbf{prox}_{s\lambda||\cdot||_*}(\mathbf{B}) = \mathbf{U}\mathcal{S}_{s\lambda}(\boldsymbol{\Sigma})\mathbf{V}^T$$

Repeat until convergence:

1. $\alpha^{(t+1)} = \alpha^{(t)} + s\mathbf{1}^T\left(\mathbf{Y} - \hat{\mathbf{P}}(\alpha^{(t)}, \mathbf{B}^{(t)})\right)$
2. $\mathbf{B}^{(t+1)} = \mathbf{prox}_{s\lambda||\cdot||_*}\left(\mathbf{B}^{(t)} + s\mathbf{X}^T\left(\mathbf{Y} - \hat{\mathbf{P}}(\alpha^{(t)}, \mathbf{B}^{(t)})\right)\right)$

sublinear convergence (Nesterov, 2007) b/c $\nabla\ell$ is Lipschitz

Accelerated PGD

Initialize $\alpha^{(0)}, \mathbf{A}^{(0)}, \mathbf{B}^{(0)}$, and iterate until convergence:

1. $\alpha^{(t+1)} = \alpha^{(t)} + s\mathbf{1}^T \left(\mathbf{Y} - \hat{\mathbf{P}}(\alpha^{(t)}, \mathbf{A}^{(t)}) \right)$
2. $\mathbf{A}^{(t+1)} = \mathbf{B}^{(t)} + \frac{t}{t+3}(\mathbf{B}^{(t)} - \mathbf{B}^{(t-1)})$
3. $\mathbf{B}^{(t+1)} = \mathbf{prox}_{s\lambda ||\cdot||_*} \left(\mathbf{A}^{(t+1)} + s\mathbf{X}^T \left(\mathbf{Y} - \hat{\mathbf{P}}(\alpha^{(t+1)}, \mathbf{A}^{(t+1)}) \right) \right)$

Much faster! Implemented on CRAN in `npmr` package.

Simulation study

Y_i simulated independently from:

$$\mathbb{P}(Y_i = k) = \frac{e^{\mathbf{X}\beta_k}}{\sum_{\ell=1}^{8} e^{\mathbf{X}\beta_\ell}} \text{ for } i = 1, ..., n \text{ and } k = 1, ..., 8$$

$$\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \text{Normal}(\vec{0}_{12}, \mathbb{I}_{12})$$

Full rank setting

$$B_{jk} \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, 1)$$

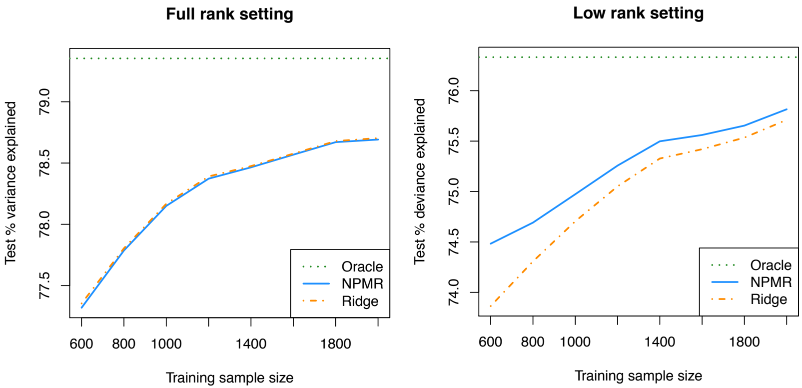
Low rank setting

$$\mathbf{B}_{12\times 8} = \mathbf{A}_{12\times 2}\mathbf{C}_{2\times 8}$$

$$A_{j\ell} \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, 1)$$

$$C_{\ell k} \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, 1)$$

Simulation results



Baseball application details

- $n = 181{,}577$ PA. For i^{th} PA, observe:
  - B_i: **B**atter (403 unique batters)
  - P_i: **P**itcher (361 unique pitchers)
  - S_i: **S**tadium
  - H_i: indicator batter is on **H**ome team
  - O_i: indicator batter has **O**pposite handedness of pitcher's

Model:

$$\mathbb{P}(Y_i = k) = \frac{e^{\eta_{ik}}}{\sum_{k' \in \mathcal{O}} e^{\eta_{ik'}}} \text{for } k \in \mathcal{O}, \text{ where}$$

$$\eta_{ik} = \alpha_k + \beta_{k:B_i} + \gamma_{k:P_i} + \delta_{k:S_i} + \zeta_k H_i + \theta_k O_i$$

Objective:

$$\underset{\alpha \in \mathbb{R}^9,\ \mathbf{B} \in \mathbb{R}^{796 \times 9}}{\text{minimize}} - \ell(\alpha, \mathbf{B}; \mathbf{X}, \mathbf{Y}) + \lambda(||\mathbf{B}_{\mathcal{B}}||_* + ||\mathbf{B}_{\mathcal{P}}||_* + ||\mathbf{B}_{\mathcal{S}}||_*)$$

Results

Batters:

Latent variable	1	2	3	4	5	6	7	8	9
1B	0.38	-0.28	-0.68	0.42	-0.14	-0.07	0.34	-0.03	-0.03
2B	0.03	-0.02	-0.06	-0.46	0.03	-0.77	0.31	0.26	0.17
3B	0.01	-0.00	-0.00	-0.27	0.16	0.09	0.31	0.00	-0.89
BB	-0.16	-0.10	-0.06	-0.45	-0.40	0.31	0.42	-0.52	0.24
F	0.14	0.87	0.09	0.25	-0.12	-0.07	0.35	-0.09	0.02
G	0.43	-0.36	0.72	0.22	-0.12	-0.02	0.33	0.02	0.03
HBP	-0.01	-0.01	-0.03	-0.01	0.85	0.22	0.36	-0.09	0.31
HR	-0.04	0.05	-0.06	-0.14	-0.19	0.47	0.23	0.80	0.14
K	-0.79	-0.15	0.09	0.45	-0.07	-0.17	0.33	0.06	-0.06
Corresponding diagonal	3.66	2.20	1.23	0.00	0.00	0.00	0.00	0.00	0.00

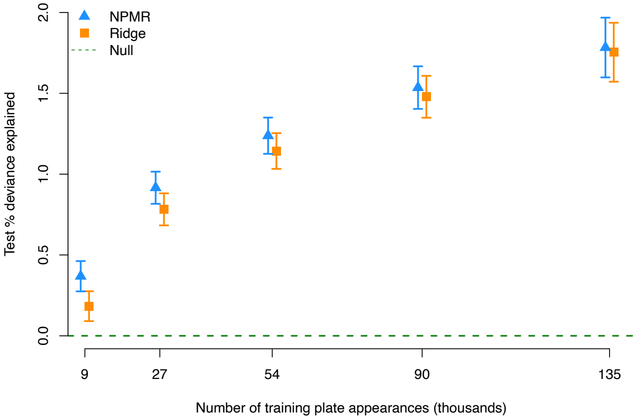
Pitchers:

Latent variable	1	2	3	4	5	6	7	8	9
1B	0.16	0.24	-0.34	0.48	-0.46	-0.27	0.42	-0.34	0.05
2B	0.01	0.03	-0.01	0.57	0.71	0.23	0.27	0.00	-0.20
3B	-0.00	-0.01	-0.05	-0.17	-0.12	0.38	-0.14	-0.61	-0.65
BB	0.07	-0.04	-0.69	-0.46	0.12	0.23	0.43	0.22	-0.01
F	0.37	-0.74	0.33	-0.01	-0.14	0.07	0.41	-0.04	0.00
G	0.26	0.62	0.51	-0.27	-0.03	0.19	0.42	0.07	-0.01
HBP	-0.01	0.01	0.00	0.19	-0.31	-0.10	-0.00	0.65	-0.66
HR	0.01	-0.00	0.05	-0.30	0.35	-0.79	0.16	-0.19	-0.31
K	-0.87	-0.09	0.18	-0.03	-0.13	0.05	0.42	-0.05	0.00
Corresponding diagonal	1.98	1.54	0.32	0.00	0.00	0.00	0.00	0.00	0.00

Results

Tool	Batters: Patience	Batters: Trajectory	Batters: Speed	Pitchers: Power	Pitchers: Trajectory	Pitchers: Command
	More K, BB	More F	More 1B	More K	More F	More F, G, K
Top 5	P Bourjos	I Kinsler	Y Cespedes	J Quintana	J Chavez	M Scherzer
	E Rosario	F Freeman	L Cain	C Kluber	J Verlander	M Tanaka
	C Santana	O Infante	J Iglesias	M Bumgarner	J Peavy	J deGrom
	G Springer	K Wong	K Kiermaier	M Scherzer	J Cueto	R de la Rosa
	M Napoli	J Altuve	D DeShields Jr	C Kershaw	C Young	M Harvey
Bot 5	J Reddick	D Gordon	E Longoria	J Danks	D Keuchel	M Pelfrey
	JT Realmuto	A Rodriguez	R Howard	D Haren	G Richards	C Tillman
	AJ Pollock	C Maybin	O Herrera	C Hamels	S Dyson	E Butler
	K Pillar	S Choo	S Smith	A Simón	B Anderson	G Gonzalez
	E Aybar	F Cervelli	J Lamb	RA Dickey	M Pineda	J Samardzija
	More F, G, 1B	More G, 1B	More G	More F, G	More G	More BB, 1B

Validation of NPMR v ridge regression



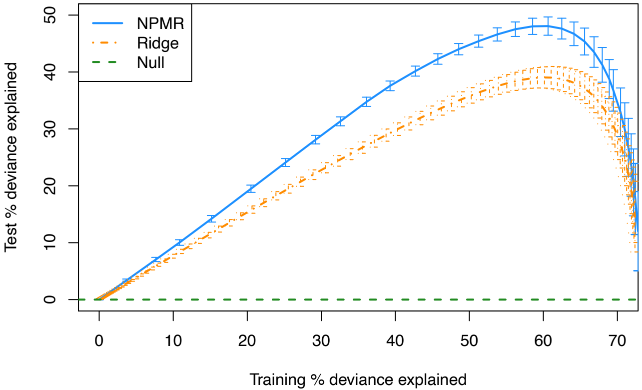
Vowel data set

Robinson (1989) vowel data:

Vowel	Word	Vowel	Word
i	heed	O	hod
I	hid	C:	hoard
E	head	U	hood
A	had	u:	who'd
a:	hard	3:	heard
Y	hud		

- 15 subjects (8 in training set, 7 in test set)
- $K = 11$, $n = 528$, $p = 10$ and $m = 462$

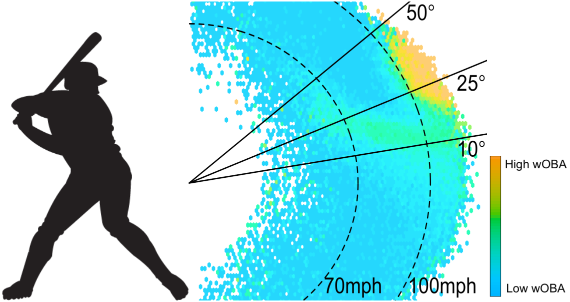
Results on vowel data



Results on vowel data

Latent variable	1	2	3	4	5	6	7	8	9	10
i (heed)	-0.13	0.51	0.66	0.08	-0.41	-0.00	0.09	-0.05	-0.07	0.00
I (hid)	-0.03	0.44	-0.30	-0.44	0.11	0.33	-0.18	0.18	0.17	-0.46
E (head)	0.35	0.32	-0.43	0.18	-0.16	-0.01	0.02	0.20	0.06	0.63
A (had)	0.52	-0.08	-0.14	0.41	-0.08	-0.11	0.22	-0.19	-0.22	-0.55
a: (hard)	0.23	-0.35	0.35	-0.13	0.20	-0.00	0.34	0.51	0.41	0.01
Y (hud)	0.22	-0.14	0.25	0.04	0.37	0.51	-0.32	-0.47	-0.00	0.24
O (hod)	0.02	-0.34	0.06	-0.17	-0.22	-0.17	-0.57	0.36	-0.49	0.00
C: (hoard)	-0.30	-0.41	-0.23	-0.02	-0.58	0.14	0.03	-0.29	0.40	-0.02
U (hood)	-0.34	-0.09	-0.15	-0.21	0.17	0.18	0.58	-0.04	-0.55	0.14
u: (who'd)	-0.53	0.05	-0.07	0.62	0.37	-0.13	-0.18	0.18	0.13	-0.08
3: (heard)	0.01	0.08	-0.01	-0.36	0.24	-0.73	-0.03	-0.40	0.15	0.07
Corresponding diagonal	9.37	7.97	2.65	1.98	1.77	0.78	0.39	0.00	0.00	0.00

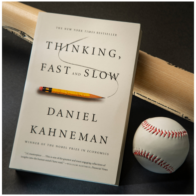
Back to baseball: Better data

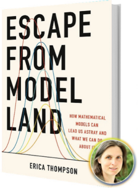


- In addition to outcome, we observe batted ball characteristics: exit velocity, launch angle, bearing
- Powers (Saberseminar 2016): Model the joint distribution of exit velocity and launch angle by batter-pitcher matchup

Lessons learned from 6 seasons in baseball

1. Start simple! Start with the data
2. Edge cases *really* matter
3. Recommended reading:

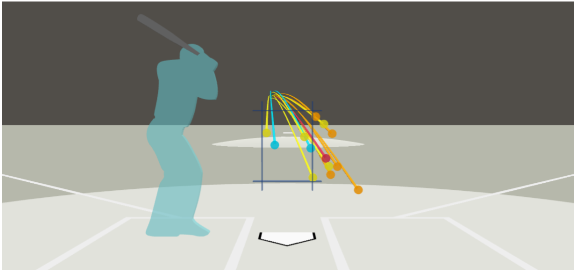




Upcoming projects

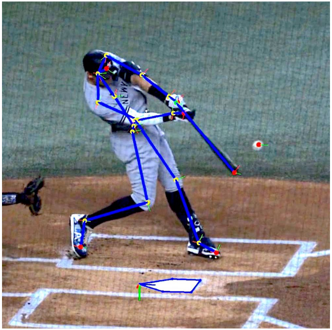
- Predicting future outcomes from pitch trajectories
- Swing biomechanics
- Volleyball analytics

Predicting future outcomes from pitch trajectories



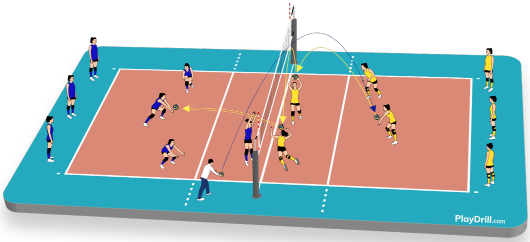
- **Data:** Each pitch generates $\vec{X} \in \mathbb{R}^9$ estimating flight path as a 3-dimensional quadratic in time
- **Problem:** Predict a pitcher's *future* results based on tracking data from past pitches thrown

Swing biomechanics



- **Data:** Each swing generates time series $\{\vec{Y}_t\} \in \mathbb{R}^{56}$ for $t = 1, \ldots, 300$
- **Problem:** Develop a statistic to measure swing adaptability

Volleyball analytics



- **Data:** Manually charted touch-by-touch data for over a decade of NCAA women's Division I volleyball
- **Problem:** What is the magnitude of the effect that individual actions have on team performance?

Thank You!

References

Anderson (1984) Regression and ordered categorical variables. *JRSS B*

Baumer and Zimbalist (2014) *The Sabermetric Revolution*

Hastie, Tibshirani and Wainwright (2015) *Statistical Learning with Sparsity: The Lasso and Generalizations*

Lu et al. (2009) Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Mathematical programming*

Toh and Yun (2009) An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*

Yee and Hastie (2003) Reduced-rank vector generalized linear models. *Statistical Modelling*

Yuan et al. (2007) Dimension reduction and coefficient estimation in multivariate linear regression. *JRSS B*