Nuclear penalized multinomial regression for predicting at bat outcomes in baseball

Scott Powers, Trevor Hastie and Robert Tibshirani

Stanford University

August 1, 2016

JSM2016 CHICAGO July 30–August 4

Outline

1. Motivation
    - Ridge regression
    - Relationships between outcome classes
2. Leveraging structure between outcome classes
    - Reduced rank regression
    - Nuclear penalized multinomial regression (NPMR)
3. Results
    - Validation
    - Interpretation

The Problem

- Baseball = sequence of matchups between 1 batter, 1 pitcher
- Each matchup results in F, G, K, BB, HBP, 1B, 2B, 3B or HR

Nuccio DiNuzzo | Chicago Tribune

- If Kris Bryant bats against Chris Sale, what is probability of each possible outcome?

Multinomial regression

What is probability Kris Bryant hits home run against Chris Sale?

$$\eta_{\text{HR}} = \alpha_{\text{HR}} + \beta_{\text{HR:Kris Bryant}} + \gamma_{\text{HR:Chris Sale}}$$

$$\mathbb{P}(\text{HR}) = \frac{e^{\eta_{\text{HR}}}}{\sum_{k \in \mathcal{O}} e^{\eta_k}}$$

Multinomial regression

What is probability Kris Bryant hits home run against Chris Sale?

$$\eta_{\text{HR}} = \alpha_{\text{HR}} + \beta_{\text{HR:Kris Bryant}} + \gamma_{\text{HR:Chris Sale}}$$

$$\mathbb{P}(\text{HR}) = \frac{e^{\eta_{\text{HR}}}}{\sum_{k \in \mathcal{O}} e^{\eta_k}}$$

More generally,

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

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

Ridge multinomial regression

- $i = 1, ..., n$, indexes plate appearances (PA)
- $\mathcal{O} = \{$F, G, K, BB, HBP, 1B, 2B, 3B, HR$\}$
- $\mathcal{B} = \{$Kris Bryant, ..., Zach Cozart$\}$
- $\mathcal{P} = \{$Chris Sale, ..., Zack Britton$\}$

For some $\lambda > 0$,

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

Matrix notation

$$\mathbf{X} = \begin{array}{c} \begin{array}{cc} \overbrace{\hspace{6em}}^{\text{Batters}} & \overbrace{\hspace{6em}}^{\text{Pitchers}} \end{array} \\ \begin{pmatrix} 1 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & \ldots & 1 \\ 0 & \ldots & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & \ldots & 0 \end{pmatrix} \end{array} \qquad \mathbf{B} = \begin{pmatrix} \beta_{F:KB} & \cdots & \beta_{HR:KB} \\ \because & \cdots & \because \\ \beta_{F:ZC} & \cdots & \beta_{HR:ZC} \\ \gamma_{F:CS} & \cdots & \gamma_{HR:CS} \\ \because & \cdots & \because \\ \gamma_{F:ZB} & \cdots & \gamma_{HR:ZB} \end{pmatrix}$$

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

Matrix notation

$$\mathbf{X} = \begin{array}{c} \quad \overbrace{\qquad\qquad\qquad}^{\text{Batters}} \quad \overbrace{\qquad\qquad\qquad}^{\text{Pitchers}} \\ \begin{pmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 1 \\ 0 & \dots & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix} \end{array} \qquad \mathbf{B} = \begin{pmatrix} \beta_{F:KB} & \dots & \beta_{HR:KB} \\ .. & \dots & .. \\ \beta_{F:ZC} & \dots & \beta_{HR:ZC} \\ \gamma_{F:CS} & \dots & \gamma_{HR:CS} \\ .. & \dots & .. \\ \gamma_{F:ZB} & \dots & \gamma_{HR:ZB} \end{pmatrix}$$

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

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

Structure between plate appearance outcomes

Ordering:

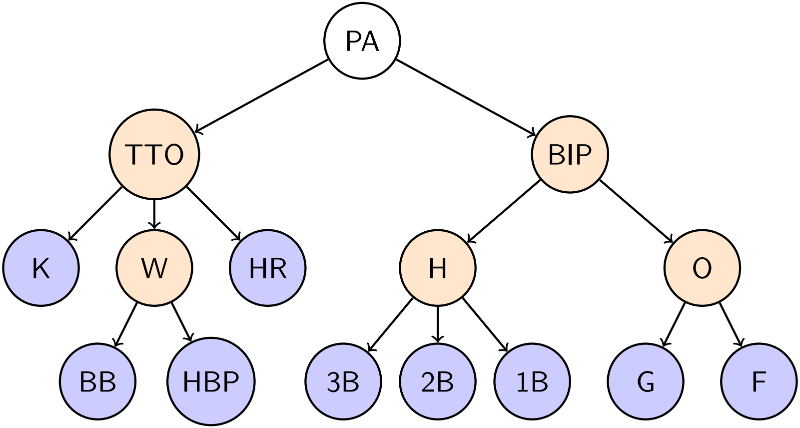
$$\text{K} < \text{G} < \text{F} < \text{BB} < \text{HBP} < \text{1B} < \text{2B} < \text{3B} < \text{HR}$$

Structure between plate appearance outcomes

Ordering:

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

Structure:



Adapted from Baumer & Zimbalist (2014)

Reduced-rank multinomial regression

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

$$\mathbf{B} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_{r=1}^{\text{rk}(\mathbf{B})} \sigma_r \mathbf{u}_r \mathbf{v}_r^T$$

- 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
- Solved via **accelerated** proximal gradient descent
- Implemented on CRAN in `npmr`

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:

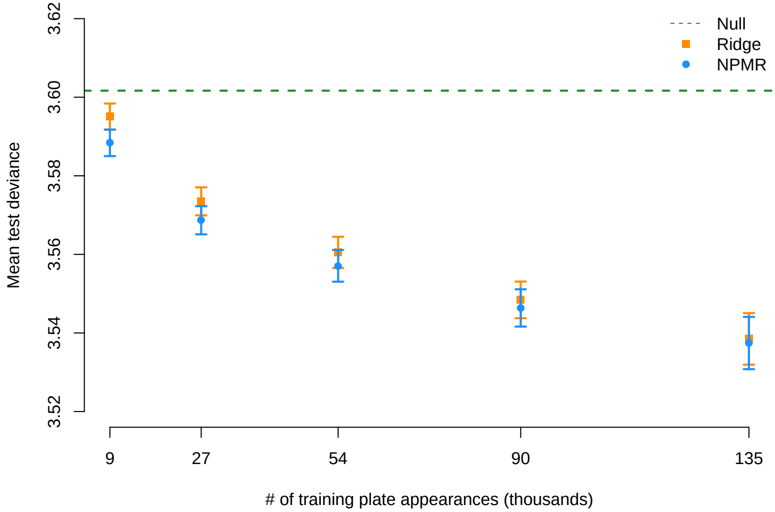
$$\mathbb{P}(Y_i = k) = \frac{e^{\eta_{ik}}}{\sum_{k' \in \mathcal{O}} e^{\eta_{ik'}}} \text{for } k \in \mathcal{O}, \text{ where}$$

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

Fitting:

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

Validation of NPMR v ridge regression



Results on 5% of season

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 on full season

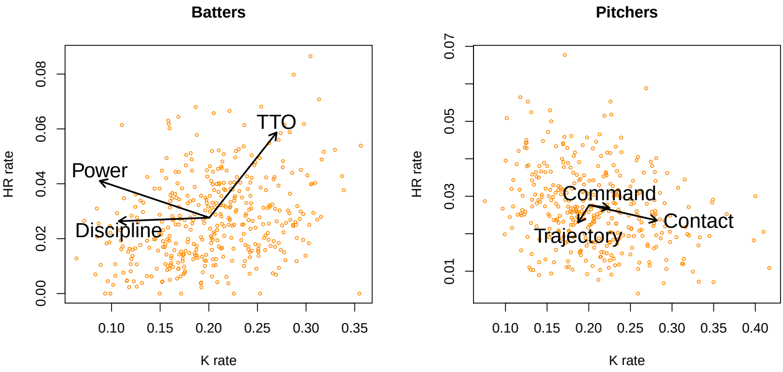
Batters

Tool	TTO	Power	Discipline
	More K, BB	More F, HR	More BB
Top 5	J Bautista	A Pujols	M Cabrera
	B Harper	D Murphy	B Zobrist
	C Carter	N Arenado	J Votto
	C Davis	S Perez	B Posey
	G Stanton	W Flores	M Brantley
Bot 5	B Pena	M Bourne	Y Gomes
	I Suzuki	A Gose	S Gennett
	E Aybar	S Peterson	S Rodriguez
	D Gordon	D DeShields	D Santana
	B Revere	R Perez	E Rosario
	More G, 1B	More K, BB	More K

Pitchers

Tool	Contact	Trajectory	Command
	More K, BB	More G, 1B	More K, G
Top 5	C Allen	S Dyson	C Kershaw
	Z Britton	J Petricka	E Scribner
	C Kimbrel	B Treinen	L Gregerson
	A Miller	B Anderson	P Hughes
	D Betances	B Ziegler	M Pineda
Bot 5	P Hughes	Y Petit	J Grimm
	M Buehrle	D Haren	M Lorenzen
	J Collmenter	T Clippard	E Butler
	J Nicolino	Y Garcia	S Oberg
	S O'Sullivan	C Young	R Detweiler
	More F, G	More F, HR	More BB, HR

Results on full season



Conclusions

- Results not significant improvement over ridge **but**
    - Interpretation of results provides interesting insight
    - NPMR natural for structed outcome space
- Method applicable to other problems
    - e.g. Robinson (1989) vowel data
- Simulation: NPMR beats ridge in low-rank regime, not much worse in full rank regime

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*

Backup slides

Proximal gradient descent

Initialize $\mathbf{B}^{(0)}$

Gradient descent: $\mathbf{B}^{(t+1)} = \mathbf{B}^{(t)} - s\nabla\ell(\mathbf{B}^{(t)})$

Proximal gradient descent

Initialize $\mathbf{B}^{(0)}$

Gradient descent: $\mathbf{B}^{(t+1)} = \mathbf{B}^{(t)} - s\nabla\ell(\mathbf{B}^{(t)})$

Proximal gradient descent (PGD):

$$\mathbf{B}^{(t+1)} = \mathbf{prox}_{s\lambda||\cdot||_*}\left(\mathbf{B}^{(t)} - s\nabla\ell(\mathbf{B}^{(t)})\right)$$

Proximal gradient descent

Proximal map for nuclear norm: soft-thresholding singular values

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

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

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 b/c gradient is Lipschitz (Nesterov, 2007)

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:

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

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

Low rank setting

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

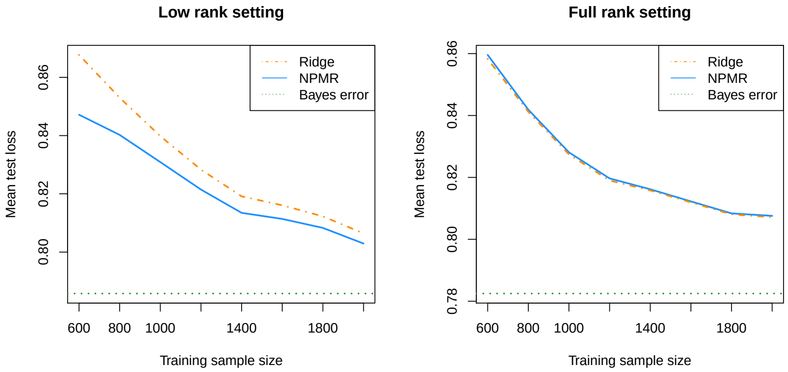
$$A_{j\ell} \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, 1)$$

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

Full rank setting

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

Simulation results



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