Nuclear penalized multinomial regression with an application to predicting at bat outcomes in baseball

Scott Powers [1], **Trevor Hastie** [1], **and Robert Tibshirani** [1]

[1] Department of Statistics, Stanford University, Stanford, CA, USA

Address for correspondence: Scott Powers, Research & Development, Los Angeles Dodgers, 1000 Vin Scully Ave., Los Angeles, CA, USA.

E-mail: saberpowers@gmail.com.

Phone: (+1) 708 828 2234.

Fax: .

Abstract: We propose the nuclear norm penalty as an alternative to the ridge penalty for regularized multinomial regression. This convex relaxation of reduced-rank multinomial regression has the advantage of leveraging underlying structure among the response categories to make better predictions. We apply our method, nuclear penalized multinomial regression (NPMR), to Major League Baseball play-by-play data to predict outcome probabilities based on batter-pitcher matchups. The interpretation of the results meshes well with subject-area expertise and also suggests a novel understanding of what differentiates players.

Key words: multinomial regression; reduced-rank regression; baseball; nuclear norm; proximal gradient descent

1 Introduction

A baseball game comprises a sequence of matchups between one batter and one pitcher. Each matchup, or *plate appearance* (PA), results in one of several outcomes. Disregarding some obscure possibilities, we consider nine categories for PA outcomes: flyout (F), groundout (G), strikeout (K), base on balls (BB), hit by pitch (HBP), single (1B), double (2B), triple (3B) and home run (HR).

A problem which has received a prodigious amount of attention in sabermetric (the study of baseball statistics) literature is determining the value of each of the above outcomes, as it leads to scoring runs and winning games. But that is only half the battle. Much less work in this field focuses on an equally important problem: optimally estimating the probabilities with which each batter and pitcher will produce each PA outcome. Even for "advanced metrics" this second task is usually done by taking simple empirical proportions, perhaps shrinking them toward a population mean using a Bayesian prior.

In statistics literature, on the other hand, many have developed shrinkage estimators for a set of probabilities with application to batting averages, starting with Stein's estimator (Efron and Morris, 1975). Since then, Bayesian approaches to this problem have been popular. Morris (1983) and Brown (2008) used empirical Bayes for estimating batting averages, which are binomial probabilities. We are interested in

estimating multinomial probabilities, like the nested Dirichlet model of Null (2009) and the hierarchical Bayesian model of Albert (2016). What all of the above works have in common is that they do not account for the "strength of schedule" faced by each player: How skilled were his opponents?

The state-of-the-art approach, Deserved Run Average (Judge and BP Stats Team, 2015, DRA), is similar to the adjusted plus-minus model from basketball and the Rasch model used in psychometrics. The latter models the probability (on the logistic scale) that a student correctly answers an exam question as the difference between the student's skill and the difficulty of the question. DRA models players' skills as random effects and also includes fixed effects like the identity of the ballpark where the PA took place. Each category of the response has its own binomial regression model. Taking HR as an example, each batter B has a propensity β_B^{HR} for hitting home runs, and each pitcher P has a propensity γ_P^{HR} for allowing home runs. Distilling the model to its elemental form, if Y denotes the outcome of a PA between batter B and pitcher P,

$$\mathbb{P}(Y = \text{HR}|B, P) = \frac{e^{\alpha^{\text{HR}}+\beta_B^{\text{HR}}+\gamma_P^{\text{HR}}}}{1 + e^{\alpha^{\text{HR}}+\beta_B^{\text{HR}}+\gamma_P^{\text{HR}}}}.$$

(Actually, in detail DRA uses the probit rather than the logit link function.)

One bothersome aspect of DRA is that the probability estimates do not sum to one; a natural solution is to use a single multinomial regression model instead of several independent binomial regression models. Making this adjustment would result in a model very similar to ridge multinomial regression (described in Section 3.3), and we will compare the results of our model with the results of ridge regression as a proxy for DRA. Ridge multinomial regression applied to this problem has about 8,000 parameters to estimate (one for each outcome for each player) on the basis of

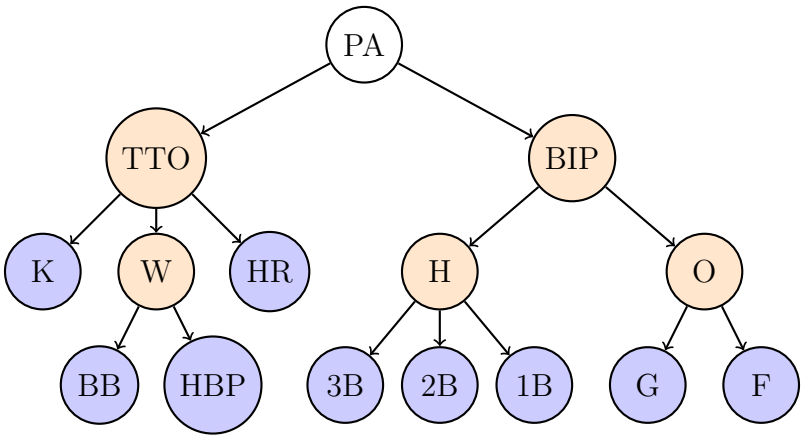
about 160,000 PAs in a season, bound together only by the restriction that probability estimates sum to one. One may seek to exploit the structure of the problem to obtain better estimates, as in ordinal regression. The categories have an ordering, from least to most valuable to the batting team:

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

with the ordering of the first three categories depending on the game situation. But the proportional odds model used for ordinal regression assumes that when one outcome is more likely to occur, the outcomes close to it in the ordering are also more likely to occur. That assumption is woefully off-base in this setting because as we show below, players who hit a lot of home runs tend to strike out often, and they tend not to hit many triples. The proportional odds model is better suited for response variables on the Likert scale (Likert, 1932), for example.

The actual relationships among the outcome categories are more similar to the hierarchical structure illustrated by Figure 1. The outcomes fall into two categories: balls in play (BIP) and the "three true outcomes" (TTO). The three true outcomes, as they have become traditionally known in sabermetric literature, include home runs, strikeouts and walks (which itself includes BB and HBP). The distinction between BIP and TTO is important because the former category involves all eight position players in the field on defense whereas the latter category involves only the batter and the pitcher. Figure 1 has been designed (roughly) by baseball experts so that terminal nodes close to each other (by the number of edges separating them) are likely to co-occur. Players who hit a lot of home runs tend to strike out a lot, and the outcomes K and HR are only two edges away from each other. Hence, the graph

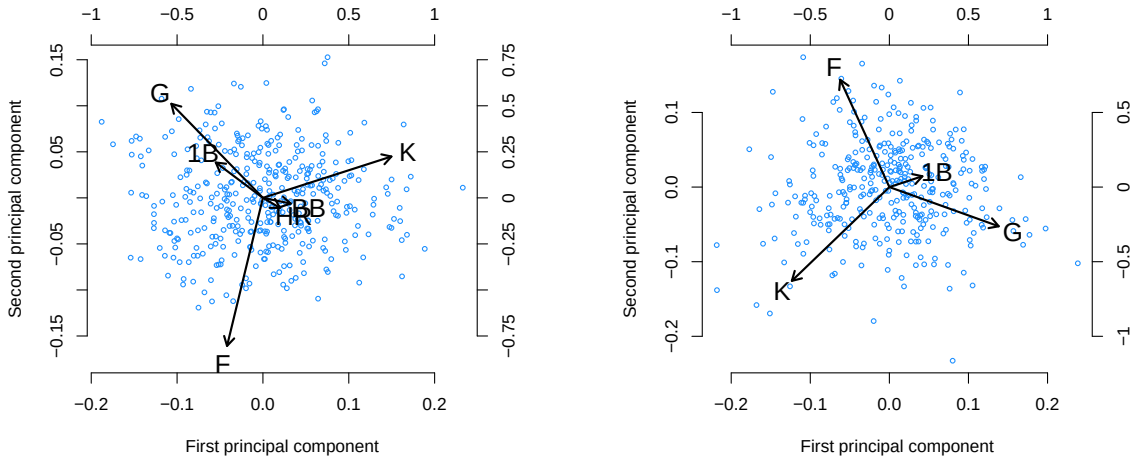
Figure 1: *Illustration of the hierarchical structure among the PA outcome categories, adapted from Baumer and Zimbalist (2014). Blue terminal nodes correspond to the nine outcome categories in the data. Orange internal nodes have the following meaning: TTO, three true outcomes; BIP, balls in play; W, walks; H, hits; O, outs. Outcomes close to each other (in terms of number of edges separating them) tend to occur in similar circumstances.*



reveals something of the underlying structure in the outcome categories.

This structure is further evidenced by principal component analysis of the player-outcome matrix, illustrated in Figure 2 and Table 1. The player-outcome matrix has a row for each player giving the proportion of PAs which have resulted in each of the nine outcomes in the dataset. For batters, the principal component (PC) which describes most of the variance in observed outcome probabilities has negative loadings on all of the BIP outcomes and positive loadings on all of the TTO outcomes. For both batters and pitchers, the percentage of variance explained after two PCs drops off precipitously.

Principal component analysis is useful for illustrating the relationships between the outcome categories. For example, Table 1(a) suggests that batters who tend to hit



(a) *Batters* (b) *Pitchers*

Figure 2: *Biplots of the principal component analyses of player outcome matrices, separate for batters and pitchers. The blue dots represent players, and the black arrows (corresponding to the top and right axes) show the loadings for the first two principal components on each of the outcomes. We exclude outcomes for which the loadings of both of the first two principal components are sufficiently small.*

singles (1B) more than average also tend to ground out (G) more than average. So an estimator of a batter's groundout rate could benefit from taking into account the batter's single rate, and *vice versa*. This is an example of the type of structure in outcome categories that motivates our proposal, which aims to leverage this structure to obtain better regression coefficient estimates in multinomial regression.

In Section 2 we review reduced-rank multinomial regression, a first attempt at leveraging this structure. We improve on this in Section 3 by proposing nuclear penalized multinomial regression, a convex relaxation of the reduced rank problem. We compare our method with ridge regression in a simulation study in Section 4. In Section 5 we apply our method and interpret the results on the baseball data, as well as another

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

(a) *Principal components of batter outcome matrix*

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

(b) *Principal components of pitcher outcome matrix*

Table 1: *Visualization of principal component analysis of player-outcome matrices, separate for batters and for pitchers. The visualization shows the loadings for each PC, along with a green bar plot corresponding to the percentage of variance explained by each PC, which is also printed in the row below the matrix of PC loadings.*

application. The manuscript concludes with a discussion in Section 6.

2 Multinomial logistic regression and reduced rank

Suppose that we observe data $\mathbf{x}_i \in \mathbb{R}^p$ and $Y_i \in \{1, ..., K\}$ for $i = 1, ..., n$. We use $\mathbf{X}$ to denote the matrix with rows $\mathbf{x}_i$, specifically $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$. The multinomial logistic regression model assumes that the Y_i are, conditional on $\mathbf{X}$, independent, and

that for $k = 1, ..., K$:

$$\mathbb{P}(Y_i = k|\mathbf{x}_i) = \frac{e^{\alpha_k + \mathbf{x}_i^T \boldsymbol{\beta}_k}}{\sum_{\ell=1}^{K} e^{\alpha_\ell + \mathbf{x}_i^T \boldsymbol{\beta}_\ell}}, \tag{2.1}$$

where $\alpha_k \in \mathbb{R}$ and $\boldsymbol{\beta}_k \in \mathbb{R}^p$ are fixed, unknown parameters. The model (2.1) is not identifiable because an equal increase in the same element of each of the $\boldsymbol{\beta}_k$ (or in each of the α_k) does not the change the estimated probabilities. That is, for each choice of parameter values there is an infinite set of choices which have the same likelihood as the original choice, for any observed data. This problem may readily be resolved by adopting the restriction for some $k_0 \in \{1, ..., K\}$ that $\alpha_{k_0} = 0$ and $\boldsymbol{\beta}_{k_0} = \mathbf{0}_p$. However, depending on the method used to fit the model, this identifiability issue may not interfere with the existence of a unique solution; in such a case, we do not adopt this restriction. For example, fitting the model with ridge regression would involve minimizing the sum of the negative log-likelihood and the sum of squares of the regression coefficients. Adding this strictly convex function to the objective leads to a unique solution. See the appendix for a detailed discussion.

In contrast with logistic regression, multinomial regression involves estimating not a vector but a matrix of regression coefficients: one for each independent variable, for each class. We denote this matrix by $\mathbf{B} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K)$. Motivated by the principal component analysis from Section 1, instead of learning a coefficient vector for each class, we might do better by learning a coefficient vector for each of a smaller number r of latent variables, each having a loading on the classes. For $r = 1$, this is the *stereotype model* originally proposed by Anderson (1984), who observed its applicability to multinomial regression problems with structure between the response categories, including ordinal structure. Greenland (1994) argued for the stereotype model as an alternative in medical applications to the standard techniques for ordinal categorical

regression: the cumulative-odds and continuation-ratio models.

Yee and Hastie (2003) generalized the model to reduced-rank vector generalized linear models. In detail, the reduced-rank multinomial logistic model (RR-MLM) fits (2.1) by solving, for some $r \in \{1, ..., K-1\}$, the optimization problem:

$$\begin{aligned} &\underset{\alpha \in \mathbb{R}^K,\ \mathbf{B} \in \mathbb{R}^{p \times K}}{\text{minimize}} \ -\sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \frac{e^{\alpha_k + \mathbf{x}_i^T \boldsymbol{\beta}_k}}{\sum_{\ell=1}^{K} e^{\alpha_\ell + \mathbf{x}_i^T \boldsymbol{\beta}_\ell}} \mathbb{I}_{\{Y_i = k\}} \right) \\ &\text{subject to} \quad \text{rank}(\mathbf{B}) \leq r, \alpha_1 = 0, \boldsymbol{\beta}_1 = \mathbf{0}_p. \end{aligned} \tag{2.2}$$

If $\text{rank}(\mathbf{B}) < r$, then there exist $\mathbf{A} \in \mathbb{R}^{p \times r}$, $\mathbf{C} \in \mathbb{R}^{K \times r}$ such that $\mathbf{B} = \mathbf{A}\mathbf{C}^T$. Under this factorization, the r columns of $\mathbf{C}$ can be interpreted as defining latent outcome variables, each with a loading on each of the K outcome classes. The r columns of $\mathbf{A}$ give regression coefficient vectors for these latent outcome variables, rather than the outcome classes.

The optimization problem (2.2) is not convex because $\text{rank}(\cdot)$ is not a convex function. Yee (2010) implemented an alternating algorithm to solve it in the R (R Core Team, 2016) package `VGAM`. However, this algorithm is too slow for feasible application to datasets as large as the one motivating Section 1. See Section 5.1 for a detailed description of the dataset.

3 Nuclear penalized multinomial regression

Because of the computational difficulty of solving (2.2), we propose replacing the rank restriction with a restriction on the nuclear norm $||\cdot||_*$ (defined below) of the

regression coefficient matrix. For some $\rho > 0$, this convex optimization problem is:

$$\begin{aligned} &\underset{\alpha\in\mathbb{R}^K,\ \mathbf{B}\in\mathbb{R}^{p\times K}}{\text{minimize}} \ -\sum_{i=1}^{n}\log\left(\sum_{k=1}^{K}\frac{e^{\alpha_k+\mathbf{x}_i^T\boldsymbol{\beta}_k}}{\sum_{\ell=1}^{K}e^{\alpha_\ell+\mathbf{x}_i^T\boldsymbol{\beta}_\ell}}\mathbb{I}_{\{Y_i=k\}}\right) \\ &\text{subject to} \quad ||\mathbf{B}||_* \leq \rho. \end{aligned} \tag{3.1}$$

We prefer to frame the problem in its equivalent Lagrangian form: For some $\lambda > 0$,

$$\begin{aligned} (\alpha^*, \mathbf{B}^*) &= \underset{\alpha\in\mathbb{R}^K,\ \mathbf{B}\in\mathbb{R}^{p\times K}}{\arg\min} \ -\sum_{i=1}^{n}\log\left(\sum_{k=1}^{K}\frac{e^{\alpha_k+\mathbf{x}_i^T\boldsymbol{\beta}_k}}{\sum_{\ell=1}^{K}e^{\alpha_\ell+\mathbf{x}_i^T\boldsymbol{\beta}_\ell}}\mathbb{I}_{\{Y_i=k\}}\right) + \lambda||\mathbf{B}||_* \\ &\equiv \underset{\alpha\in\mathbb{R}^K,\ \mathbf{B}\in\mathbb{R}^{p\times K}}{\arg\min} \ -\ell(\alpha, \mathbf{B}; \mathbf{X}, Y) + \lambda||\mathbf{B}||_*. \end{aligned} \tag{3.2}$$

This optimization problem (3.2) is what we call nuclear penalized multinomial regression (NPMR). We use $\ell(\alpha, \mathbf{B}; \mathbf{X}, Y)$ to denote the log-likelihood of the regression coefficients α and $\mathbf{B}$ given the data $\mathbf{X}$ and Y. The nuclear norm of a matrix is defined as the sum of its singular values, that is, the ℓ_1-norm of its vector of singular values. Explicitly, consider the singular value decomposition of $\mathbf{B}$ given by $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, with $\mathbf{U} \in \mathbb{R}^{p\times p}$ and $\mathbf{V} \in \mathbb{R}^{K\times K}$ orthogonal and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times K}$ having values $\sigma_1, ..., \sigma_{\min\{p,K\}}$ along the main diagonal and zeros elsewhere. Then

$$||\mathbf{B}||_* = \sum_{d=1}^{\min\{p,K\}} \sigma_d.$$

In the same way that the lasso induces sparsity of the estimated coefficients in a regression, solving (3.2) drives some of the singular values to exactly zero. Because the number of nonzero singular values is the rank of a matrix, the result is that the estimated coefficient matrix $\mathbf{B}^*$ tends to have less than full rank. Thus, (3.2) is a convex relaxation of the reduced-rank multinomial logistic regression problem, in much

the same way as the lasso is a convex relaxation of best subset regression (Tibshirani, 1996). The convexity of (3.2) makes it easier to solve than (2.2), and we discuss algorithms for solving it in Sections 3.1 and 3.2. In practice, we recommend using standard cross-validation (CV) techniques for selecting the regularization parameter λ, which controls the rank of the solution. For CV loss we use multinomial deviance; other choices are valid. Throughout this manuscript we use 10-fold CV.

Consider the singular value decomposition $\mathbf{U}^*\mathbf{\Sigma}^*\mathbf{V}^{*T}$ of the $p \times K$ estimated coefficient matrix $\mathbf{B}^*$. Each column of the $K \times K$ orthogonal matrix $\mathbf{V}^*$ represents a latent variable as a linear combination of the K outcome categories. Meanwhile, each row of $\mathbf{U}^*\mathbf{\Sigma}^*$ specifies for each predictor variable a coefficient for each *latent* variable, rather than for each outcome category. By estimating some of the singular values of $\mathbf{B}^*$ (the entries of the diagonal $p \times K$ matrix $\mathbf{\Sigma}^*$) to be zero, we reduce the number of coefficients to be estimated for each predictor variable from (a) one for each of K outcome categories; to (b) one for each of some smaller number of latent variables. These latent variables learned by the model express relationships between the outcomes because two categories for which a latent variable has both large positive coefficients are both likely to occur for large values of that latent variable. Similarly, if a latent variable has a large positive coefficient for one category and a large negative coefficient for another, those two categories oppose each other diametrically with respect to that latent variable.

3.1 Proximal gradient descent

NPMR relies on solving (3.2). The objective is convex but non-differentiable where any singular values of $\mathbf{B}$ are zero, so we cannot use gradient descent. Generally, when

minimizing a function $f : \mathbb{R}^d \to \mathbb{R}$ of a vector $\mathbf{x} \in \mathbb{R}^d$, the gradient descent update of step size s takes the form

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - s\nabla f(\mathbf{x}^{(t)}),$$

or equivalently (Hastie et al., 2015),

$$\mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ f(\mathbf{x}^{(t)}) + \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{x} - \mathbf{x}^{(t)} \rangle + \frac{1}{2s}||\mathbf{x} - \mathbf{x}^{(t)}||_2^2 \right\}.$$

Still, if f is non-differentiable, as it is in (3.2), then ∇f is undefined. However, if f is the sum of two convex functions g and h, with g differentiable, we can instead apply the generalized gradient update step (Hastie et al., 2015):

$$\mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ g(\mathbf{x}^{(t)}) + \langle \nabla g(\mathbf{x}^{(t)}), \mathbf{x} - \mathbf{x}^{(t)} \rangle + \frac{1}{2s}||\mathbf{x} - \mathbf{x}^{(t)}||_2^2 + h(\mathbf{x}) \right\}. \tag{3.3}$$

Repeatedly applying this update is known as proximal gradient descent (PGD). In (3.2), we have $\mathbf{x} = (\alpha, \mathbf{B})$, $g = -\ell$ and $h = ||\cdot||_*$. So the PGD update step is:

$$\begin{aligned}\left(\alpha^{(t+1)},\ \mathbf{B}^{(t+1)}\right) = \arg\min_{\alpha,\ \mathbf{B}} \big\{ & -\ell(\alpha^{(t)}, \mathbf{B}^{(t)}; \mathbf{X}, \mathbf{Y}) \\ & + \langle -\mathbf{X}^T(\mathbf{Y} - \mathbf{P}^{(t)}), \mathbf{B} - \mathbf{B}^{(t)} \rangle + \langle -1_n^T(\mathbf{Y} - \mathbf{P}^{(t)}), \alpha - \alpha^{(t)} \rangle \\ & + \frac{1}{2s}||\mathbf{B} - \mathbf{B}^{(t)}||_F^2 + \frac{1}{2s}\,||\alpha - \alpha^{(t)}||_2^2 + \lambda||\mathbf{B}||_* \big\},\end{aligned}$$

where $\mathbf{Y} \in \{0,1\}^{n\times K}$ is the matrix containing the response variable and $\mathbf{P} \in (0,1)^{n\times K}$

is the matrix containing the fitted values. That is, for $i = 1, ..., n$, and $k = 1, ..., K$,

$$\{\mathbf{Y}\}_{ik} = \mathbb{I}_{\{Y_i = k\}}, \quad \text{and} \quad \{\mathbf{P}\}_{ik} = \frac{e^{\alpha_k + \mathbf{x}_i^T \boldsymbol{\beta}_k}}{\sum_{\ell=1}^{K} e^{\alpha_\ell + \mathbf{x}_i^T \boldsymbol{\beta}_\ell}}. \tag{3.4}$$

The squared Frobenius norm $||\cdot||_F^2$ is the sum of the squares of the entries of a matrix.

The problem is separable in α and $\mathbf{B}$:

$$\begin{aligned} \alpha^{(t+1)} &= \arg\min_{\alpha} \left\{ \langle -1_n^T(\mathbf{Y} - \mathbf{P}^{(t)}), \alpha - \alpha^{(t)} \rangle + \frac{1}{2s} ||\alpha - \alpha^{(t)}||_2^2 \right\} \\ &= \alpha^{(t)} + s 1_n^T (\mathbf{Y} - \mathbf{P}^{(t)}), \text{ and} \end{aligned} \tag{3.5}$$

$$\begin{aligned} \mathbf{B}^{(t+1)} &= \arg\min_{\mathbf{B}} \left\{ \langle -\mathbf{X}^T(\mathbf{Y} - \mathbf{P}^{(t)}), \mathbf{B} - \mathbf{B}^{(t)} \rangle + \frac{1}{2s} ||\mathbf{B} - \mathbf{B}^{(t)}||_F^2 + \lambda ||\mathbf{B}||_* \right\} \\ &= \mathcal{S}^*_{s\lambda}(\mathbf{B}^{(t)} + s\mathbf{X}^T(\mathbf{Y} - \mathbf{P}^{(t)})), \end{aligned} \tag{3.6}$$

where $\mathcal{S}^*_{s\lambda} : \mathbb{R}^{p \times K} \to \mathbb{R}^{p \times K}$ is the soft-thresholding operator on the singular values of a matrix. Explicitly, if a matrix $\mathbf{M} \in \mathbb{R}^{p \times K}$ has singular value decomposition $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, then $\mathcal{S}^*_{s\lambda}(\mathbf{M}) = \mathbf{U}\mathcal{S}_{s\lambda}(\boldsymbol{\Sigma})\mathbf{V}^T$, where

$$\{\mathcal{S}_{s\lambda}(\boldsymbol{\Sigma})\}_{jk} = \text{sign}(\boldsymbol{\Sigma}_{jk}) \max\{|\boldsymbol{\Sigma}_{jk}| - s\lambda, 0\}.$$

$\mathcal{S}^*_{s\lambda}$ is called the *proximal operator* of the nuclear norm, and in general solving (3.3) involves the proximal operator of h, hence the name proximal gradient descent.

So to solve (3.2), initialize α and $\mathbf{B}$, and iteratively apply the updates (3.5) and (3.6). Due to Nesterov (2007), this procedure converges with step size $s \in (0, 1/L)$ if the log-likelihood ℓ is continuously differentiable and has Lipschitz gradient with Lipschitz constant L. The appendix includes a proof that the gradient of ℓ is Lipschitz with constant $L = \sqrt{K}||\mathbf{X}||_F^2$, but in practice we recommend starting with step size $s = 0.1$ and halving the step size if any proximal gradient descent step would result in an increase of the objective function (3.2).

3.2 Accelerated PGD

In practice, we find that it helps to speed things up considerably to use an accelerated PGD method, also due to Nesterov (2007). Specifically, we iteratively apply the following updates:

1. $\alpha^{(t+1)} = \alpha^{(t)} + s\mathbf{1}_n^T\left(\mathbf{Y} - \mathbf{P}^{(t)}\right)$
2. $\mathbf{A}^{(t+1)} = \mathbf{B}^{(t)} + \frac{t}{t+3}(\mathbf{B}^{(t)} - \mathbf{B}^{(t-1)})$
3. $\mathbf{P}^{(t+1)} = \mathbf{P}(\alpha^{(t+1)}, A^{(t+1)})$
4. $\mathbf{B}^{(t+1)} = \mathcal{S}^*_{s\lambda}\left(\mathbf{A}^{(t+1)} + s\mathbf{X}^T\left(\mathbf{Y} - \mathbf{P}^{(t+1)}\right)\right)$

The function $\mathbf{P}(\cdot)$ in Step 3 returns the matrix of fitted probabilities based on the regression coefficients as described in (3.4). Step 2 is the key to the acceleration because it uses the "momentum" in $\mathbf{B}$ to push it further in the same direction it is heading. We strongly recommend using this accelerated version of PGD, and our implementation of NPMR is available on the Comprehensive R Archive Network as the R package `npmr`.

3.3 Related work

Tutz and Gertheiss (2016) provide a systematic review of regularized regression for categorical data. NPMR is a novel method in this category and would fit well in their section on categorical response variables. The authors describe penalties for variable selection multinomial logistic regression and in ordinal regression. None of these methods induces low rank in the solution, as NPMR does. The idea of using a

nuclear norm penalty as a convex relaxation to reduced-rank regression has previously been proposed in the Gaussian regression setting (Chen et al., 2013), but we are not aware of any attempt to do so in the multinomial setting.

The nearest competitor to NPMR that can feasibly be applied to the baseball matchup dataset is multinomial ridge regression, which penalizes the squared Frobenius norm (the sum of the squares of the entries) of the coefficient matrix, instead of the nuclear norm. In detail, ridge regression estimates the regression coefficients by solving the optimization problem

$$(\alpha^*, \mathbf{B}^*) = \underset{\alpha \in \mathbb{R}^K,\ B \in \mathbb{R}^{p \times K}}{\arg\min} - \ell(\alpha^{(t)}, \mathbf{B}^{(t)}; \mathbf{X}, Y) + \lambda ||\mathbf{B}||_F^2. \tag{3.7}$$

This model is very similar to the state of the art in public sabermetric literature for evaluating pitchers on the basis of outcomes while simultaneously controlling for sample size, opponent strength and ballpark effects (Judge and BP Stats Team, 2015). Software is available to solve this problem very quickly in the R package `glmnet` (Friedman et al., 2010). This is the standard approach used for regularized multinomial regression problems, so we use it as the benchmark against which to compare the performance of NPMR in Sections 4 and 5.

4 Simulation study

In this section we present the results of two different simulations, one using a full-rank coefficient matrix and the other using a low-rank coefficient matrix. In both settings we vary the training sample size n from 600 to 2000, and we fix the number

of predictor variables to be 12 and the number of levels of the response variable to be 8. Given design matrix $\mathbf{X} \in \mathbb{R}^{n\times 12}$ and coefficient matrix $\mathbf{B} \in \mathbb{R}^{12\times 8}$, we simulate the response according to the multinomial regression model. Explicity, for $i = 1, ..., n$, and $k = 1, ..., 8$,

$$\mathbb{P}(Y_i = k) = \frac{e^{\mathbf{X}\boldsymbol{\beta}_k}}{\sum_{\ell=1}^{8} e^{\mathbf{X}\boldsymbol{\beta}_\ell}}.$$

For both simulations the entries of $\mathbf{X}$ are i.i.d. standard normal:

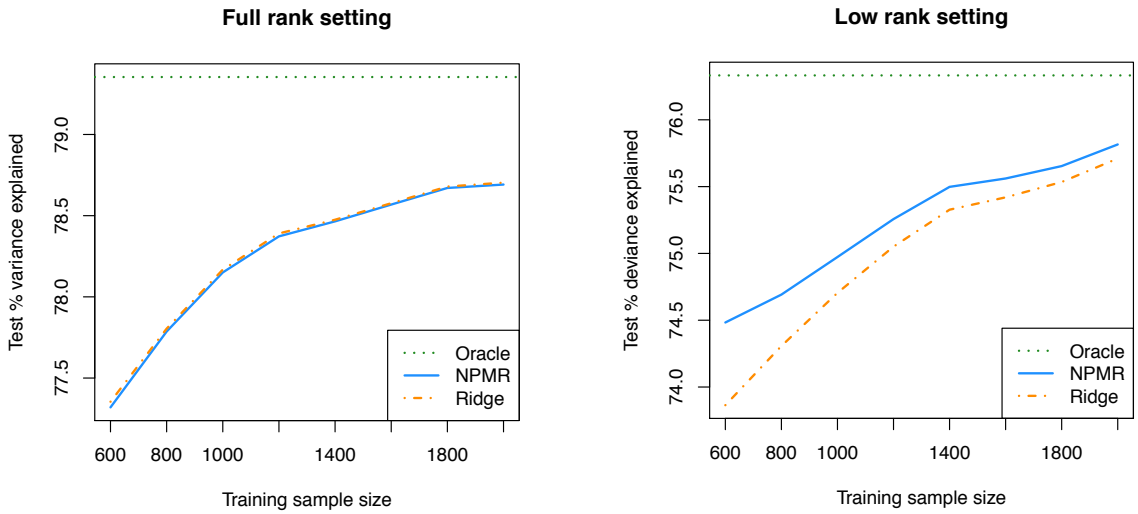
$$\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mathbf{0}_{12}, \mathbf{I}_{12})$$

for $i = 1, ..., n$. However the simulations differ in the generation of the coefficient matrix $\mathbf{B}$. In the *full rank* setting, the entries of $\mathbf{B}$ follow an i.i.d. standard normal distribution: For $k = 1, ..., 8$,

$$\boldsymbol{\beta}_k \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mathbf{0}_{12}, \mathbf{I}_{12}). \tag{4.1}$$

In the *low rank* setting we first simulate two intermediary matrices $\mathbf{A} \in \mathbb{R}^{12\times 2}$ and $\mathbf{C} \in \mathbb{R}^{8\times 2}$ with i.i.d. standard normal entries, and we then define $\mathbf{B} = \mathbf{A}\mathbf{C}^T$ so that the rank of $\mathbf{B}$ is 2. In each simulation we fit ridge regression and NPMR to the training sample of size n and estimate the out-of-sample error by simulating 10,000 test observations, comparing the model's predictions on those test observations with the simulated response. The results of 3500 simulations in each setting, for each training sample size n, are presented in Figure 3.

In these simulations and throughout this manuscript we evaluate the methods using



(a) *Full-rank coefficient matrix*

(b) *Low-rank coefficient matrix*

Figure 3: *Simulation results. We plot the percentage of deviance explained in a test set against training sample size. The oracle prediction is based on the known class probabilities from which the test class was drawn. In (a), the full rank setting, ridge regression out-performs NPMR by a slim margin. In (b), the low rank setting, NPMR wins, especially for smaller sample sizes.*

percentage of deviance explained in the test set. Multinomial deviance is twice the negative log of the probability predicted for the class observed. The null deviance corresponds to using the overall frequency of each class in the training set as the predicted probability for that class. The percentage of deviance explained is the difference between the null deviance and the deviance of the method's predictions, divided by the null deviance.

In the full rank setting we expect ridge regression to out-perform NPMR because ridge regression shrinks all coefficient estimates toward zero, which is the mean of the generating distribution for the coefficients in the simulation. If this were a Gaussian regression problem instead of a multinomial regression problem, then the ridge re-

gression coefficient estimates would correspond (Hastie et al., 2009) to the posterior mean estimate under a Bayesian prior of (4.1). In fact, ridge regression does beat NPMR in this simulation (for all training sample sizes n), but NPMR's performance is surprisingly close to that of ridge regression.

The low rank setting is one in which NPMR should have better test performance than does ridge regression. NPMR bets on sparsity in the singular values of the coefficient matrix, and in this setting the bet pays off. The simulation results verify that this intuition is correct. NPMR beats ridge regression for all training sample sizes n but especially for smaller sample sizes. By betting (correctly in this case) on the coefficient matrix having less than full rank, NPMR learns more accurate estimates of the coefficient matrix. As the training sample size increases, learning the coefficient matrix becomes easier, and the performance gap between the two methods shrinks but remains evident.

In summary, this simulation demonstrates that each of NPMR and ridge regression is superior in a simulation tailored to its strengths, confirming our intuition. Furthermore, in a simulation constructed in favor of ridge regression, NPMR performs nearly as well. Meanwhile, NPMR leads to more significant gains over ridge regression in the low rank setting. In this simulation and in the applications to follow, the number of response categories is more than just a few. This is intentional; for a small number of classes, e.g. $K = 3$ or 4, then ridge regression would estimate a low-rank regression coefficient matrix itself, as the rank can be no larger than the number of columns.

5 Results

5.1 Implementation details

The 2015 MLB play-by-play dataset from Retrosheet includes an entry for every plate appearance during the six-month regular season. For the purposes of fitting NPMR to predict the outcomes of PAs, the following relevant variables are recorded for the i^{th} PA: the identity (B_i) of the batter; the identity (P_i) of the pitcher; the identity (S_i) of the stadium where the PA took place; an indicator (H_i) of whether the batter's team is the home team; and finally an indicator (O_i) of whether the batter's handedness (left or right) is opposite that of the pitcher.

For each outcome $k \in \mathcal{K} \equiv \{\text{K, G, F, BB, HBP, 1B, 2B, 3B, HR}\}$, the multinomial model fit by both NPMR and ridge regression is specified by

$$\mathbb{P}(Y_i = k) = \frac{e^{\eta_{ik}}}{\sum_{\ell \in \mathcal{K}} e^{\eta_{i\ell}}}, \text{ where}$$

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

The parameters introduced have the following interpretation: α_k is an intercept corresponding to the league-wide frequency of outcome k; $\beta_{B_i k}$ corresponds to the tendency of batter B_i to produce outcome k; $\gamma_{P_i k}$ corresponds to the tendency of pitcher P_i to produce outcome k; $\delta_{S_i k}$ corresponds to the tendency of stadium S_i to produce outcome k; ζ_k corresponds to the increase in likelihood of outcome k due to home field advantage; and θ_k corresponds to the increase in likelihood of outcome k due to the batter having the opposite handedness of the pitcher's.

NPMR and ridge regression fit the same multinomial regression model and differ only

in the regularizations used in their objective functions, yielding different results. See Section 3 for details. However, there is a minor tweak to NPMR for application to these data. Instead of adding to the objective a penalty on the nuclear norm of the whole coefficient matrix, we add penalties on the nuclear norms of the three coefficient sub-matrices corresponding to batters, pitchers and stadiums. The coefficients for home-field advantage and opposite handedness remain unpenalized. The result is that NPMR learns different latent variables for batters than it does for pitchers, instead of learning one set of latent variables for both groups.

We process the PA data before applying NPMR and ridge regression. First, we define a minimum PA threshold separately for batters and pitchers. For batters the threshold is the 390^{th}-largest number of PAs among all batters. This corresponds roughly to the number of rostered batters at any given time during the MLB regular season. Batters who fall below the PA threshold are labelled "replacement level" and within each defensive position are grouped together into a single identity. For example, "replacement-level catcher" is a batter in the dataset just like Mike Trout is, and the former label includes all PAs by a catcher who does not meet the PA threshold. Similarly, we define the PA threshold for pitchers to be the 360^{th}-largest number of PAs among all pitchers, and we group all pitchers who fall below that threshold under the "replacement-level pitcher" label. Additionally, we discard all PAs in which a pitcher is batting, and we discard PAs which result in a catcher's interference or an intentional walk. The result is a set of 176,559 PAs featuring 400 unique batters and 362 unique pitchers in 30 unique stadiums. Note that we have more than 390 batter and 360 pitcher identities because of ties at the PA threshold and because of the replacement-level identities we have introduced.

5.2 Validation

We fit NPMR and ridge regression to the baseball data, using a training sample that varied from 5% (roughly 9,000 PAs) to 75% (roughly 135,000 PAs) of the data. We used the remaining data to test the models, reporting the percentage of deviance explained in the test set, as described in Section 4. Figure 4 gives the results.

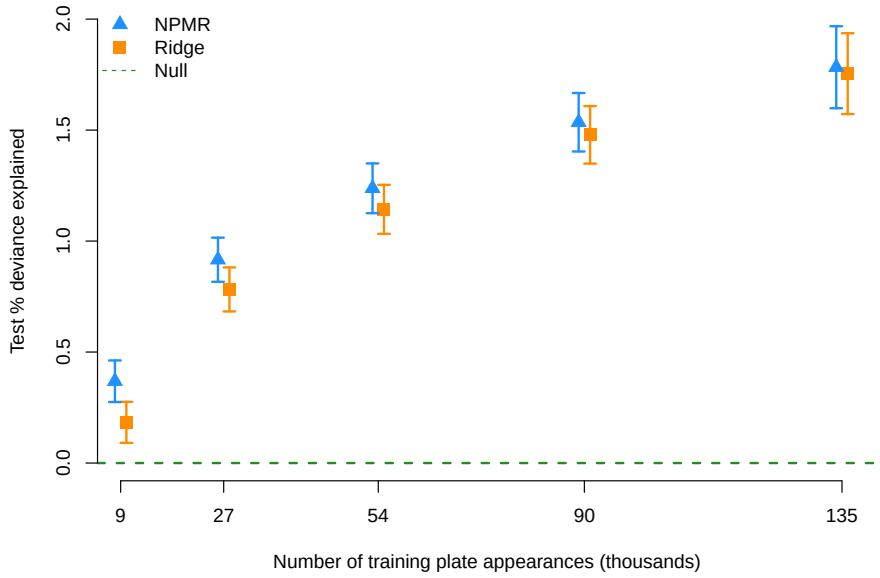


Figure 4: *Out-of-sample test performance of NPMR, ridge and null estimators on baseball plate appearance result prediction. Each estimator was trained on a fraction of the 2015 regular season data (varying from 5 to 75 percent) and tested on the remaining data. The error bars correspond to one standard error. The standard error of the mean test deviance is its standard deviation across test samples, divided by the square root of the number of test samples.*

For each training sample size, NPMR outperforms ridge regression though the difference is not statistically significant. Note that the standard errors reflect random

variation in the test set, for a single training set. At the smallest sample size NPMR, unlike ridge regression, explains significantly more test deviance than does the null estimator. There is value in improved estimation of players' skills in small sample sizes because this can inform early-season decision-making. For all other sample sizes, both NPMR and ridge regression achieves performances which are statistically significantly better than the null. The primary benefit of NPMR relative to ridge regression is the interpretation, as described in the next section.

5.3 Interpretation

We focus on the results of fitting NPMR on 5 percent of the training data because there the difference between NPMR and ridge regression is greatest (Figure 4). As the sample size increases, the need for a low-rank regression coefficient matrix is reduced, and the NPMR solution becomes more similar to the ridge solution. Table 2 visualizes the singular value decomposition of the fitted regression coefficient submatrices corresponding to batters and pitchers. Unlike in Table 1, the diagonal entries shown in the bottom row are shrunk toward zero and cannot be interpreted in the context of percent variance explained. Note that for both batters and pitchers, six of the nine diagonal entries are exactly zero, illustrating the ability of NPMR to perform selection on the latent skills.

We observe that for both batters and pitchers, NPMR identifies three latent variables which differentiate players from one another. By construction, these latent variables are measuring separate aspects of players' skills; across players, expression in each latent skill is uncorrelated with expression in each other latent skill. In that sense, we have identified three separate skills which characterize hitters and three separate

skills which characterize pitchers. In baseball scouting parlance, these skills are called "tools", but unlike the five traditional baseball tools (hitting for power, hitting for contact, running, fielding and throwing), the tools we identify are uncorrelated with one another.

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

(a) *Latent variables for batter regression coefficient matrix*

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

(b) *Latent variables for pitcher regression coefficient matrix*

Table 2: *Visualization of fitted regression coefficient matrices from NPMR on 5% of the baseball data. The matrix displayed is* $\mathbf{V}$ *in the* $\mathbf{U\Sigma V}^T$ *decomposition of* $\mathbf{B}$ *from (3.2), with columns corresponding to latent variables and rows corresponding to outcomes. The bottom row gives the entry in the diagonal matrix* $\mathbf{\Sigma}$ *corresponding to the latent variable.*

The interpretation of Table 2 is very attractive in the context of domain knowledge. In reading the columns of $\mathbf{V}$, note that they are unique only up to a change in sign, so we can interpret positive expression in each skill as positive or negative values of the corresponding latent variable. We suggest the following interpretation of the first three latent skills for batters:

- *Skill 1: Patience.* The loadings of the first latent variable discriminate perfectly between the TTO outcomes and the BIP outcomes described in Section 1. We label this skill as "patience" because when a batter swings at fewer pitches, he is less likely to hit the ball in play.

- *Skill 2: Trajectory.* The second latent variable distinguishes primarily between F and G, corresponding to the vertical angle of the ball off the bat.

- *Skill 3: Speed.* The third latent variable distinguishes primarily between 1B and G. Examining the players with strong positive expression of this skill, we find fast players who are more difficult to throw out at first base on a ground ball.

From this interpretation we learn that the primary skill which distinguishes betwen batters is how often they hit the ball into the field of play. One outcome over which batters have relatively large control is how often they swing at pitches. Among balls that are put into play, batters have less but still subtantial control over whether those are ground balls or fly balls. It is the vertical angle of the batter's swing plane, along with whether he tends to contact the top half or the bottom half of the ball, that determines his trajectory tendency. Finally, given the trajectory of the ball off the bat, the batter has relatively little control over the outcome of the PA. But to the extent that he can influence this outcome, fast runners tend to hit more singles and fewer groundouts.

Based on Table 2, we interpret the pitchers' skills as follows:

- *Skill 1: Power.* The first latent variable distinguishes primarily between K and F (and G), thus identifying how the pitcher gets outs. Pitchers who tend to get their outs via the strikeout are known in baseball as "power pitchers".

- *Skill 2: Trajectory.* As with batters, the second latent variable distinguishes primarily between F and G, corresponding to the batted ball's angle.

- *Skill 3: Command.* The third latent variable distinguishes primarily between positive outcomes for the pitcher (F, G and K) and negative outcomes for the pitcher (BB and 1B), reflecting how well he is able to control his pitches.

The interpretation of the first two skills for pitchers is very similar to the interpretation of the first two skills for batters. Primarily, pitchers can influence how often balls are hit into play against them, but they exhibit less control over this than batters do. Secondarily, as with hitters, pitchers exhibit some control over the vertical launch angle of the ball off the bat. This is based on the location and movement of their pitches. The third skill, distinguishing between positive and negative outcomes, has a relatively small magnitude.

Table 3: *Top 5 and bottom 5 batters in the three latent skills identified by NPMR.*

Skill	**Patience**	**Trajectory**	**Speed**
	More K, BB	More F	More 1B
Top 5	Peter Bourjos	Ian Kinsler	Yoenis Cespedes
	Eddie Rosario	Freddie Freeman	Lorenzo Cain
	Carlos Santana	Omar Infante	José Iglesias
	George Springer	Kolten Wong	Kevin Kiermaier
	Mike Napoli	José Altuve	Delino DeShields Jr
Bottom 5	Josh Reddick	Dee Gordon	Evan Longoria
	JT Realmuto	Alex Rodriguez	Ryan Howard
	AJ Pollock	Cameron Maybin	Odubel Herrera
	Kevin Pillar	Shin-Soo Choo	Seth Smith
	Eric Aybar	Francisco Cervelli	Jake Lamb
	More F, G, 1B	More G, 1B	More G

Table 3 lists the top five and bottom five players in each of the three latent batting skills learned by NPMR. These results largely match intuition for the players listed, and to the extent that they do not, it is worth a reminder that they are based on 5%

of the full season's data. That is roughly equivalent to nine days' worth of data from the six-month season. The median number of PAs per batter in the training set is 21.

Table 4: *Top 5 and bottom 5 pitchers in the three latent skills identified by NPMR.*

Tool	Power	Trajectory	Command
	More K	More F	More F, G, K
Top 5	José Quintana	Jesse Chavez	Max Scherzer
	Corey Kluber	Justin Verlander	Masahiro Tanaka
	Madison Bumgarner	Jake Peavy	Jacob deGrom
	Max Scherzer	Johnny Cueto	Rubby de la Rosa
	Clayton Kershaw	Chris Young	Matt Harvey
Bottom 5	John Danks	Dallas Keuchel	Mike Pelfrey
	Dan Haren	Garrett Richards	Chris Tillman
	Cole Hamels	Sam Dyson	Eddie Butler
	Alfredo Simón	Brett Anderson	Gio Gonzalez
	RA Dickey	Michael Pineda	Jeff Samardzija
	More F, G	More G	More BB, 1B

The results in Table 4, listing the top and bottom players in each of the three latent pitching skills, are more interesting. The top five power pitchers are all among the top starting pitchers in the game. All the way on the other side of the spectrum is knuckleball pitcher RA Dickey. The knuckleball is a unique pitch in baseball thrown relatively softly with as little spin as possible to create unpredictable movement. Its goal is not to overpower the opposing batter but to induce weak contact. Another interesting pitcher low on power is Cole Hamels. Two of the leading sabermetric websites, Baseball Prospectus and FanGraphs, disagree greatly on Hamels' value. The discrepancy stems from Baseball Prospectus giving full weight to BIP outcomes while FanGraphs ignores them. Because Hamels tends to get outs via fly balls and ground balls rather than strikeouts, FanGraphs estimates a much lower value for Hamels than Baseball Prospectus does.

5.4 Another application: Vowel data

In a dataset collected by Deterding (1990) and popularized in part by Robinson (1989), the author recorded samples of 11 vowels spoken by 15 unique speakers. Each audio clip is split into 6 frames during a duration of steady audio, yielding 6 pseudo-replicates. Hence, the dataset consists of $n = 11 \times 15 \times 6 = 990$ observations. Each of these observations is represented by $p = 10$ features extracted from an audio file and is labelled as one of the vowels outlined in Table 5. Robinson (1989) split the dataset into a training sample of 528 observations and a test sample of 462 observations, stratified by speaker so that the 8 speakers in the training set are distinct from the 7 speakers in the test set.

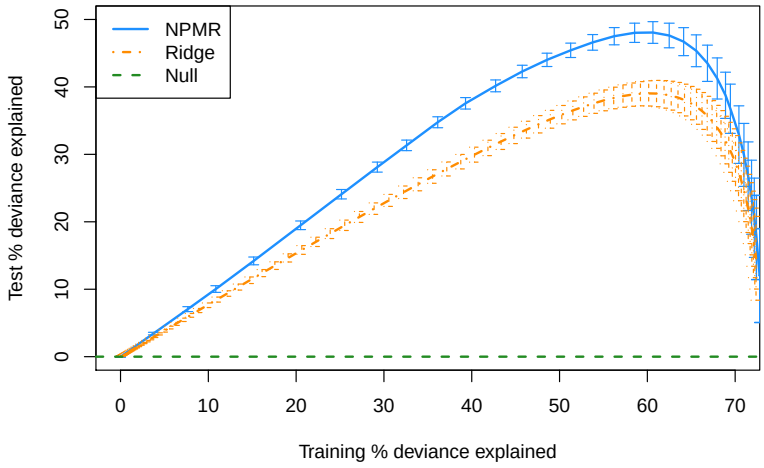
Table 5: *Vowels and corresponding words from Deterding (1990). Each word is provided to illustrate the vowel sound represented by the corresponding symbol.*

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

We fit NPMR and ridge regression to the training data over a wide range of regularization parameters, with the results reported in Figure 5. As the regularization parameter decreases for each method, the training performance improves. The test performance initially improves and then worsens as the model overfits to the training data. We observe that over the whole solution path, for the same training deviance NPMR consistently explains more of the test deviance than ridge regression.

Table 6 reveals a possible explanation why NPMR outperforms ridge regression on

Figure 5: *Results of fitting NPMR and ridge regression on vowel data. Test deviance explained is plotted against training deviance explained. Training performance serves as a surrogate for degrees of freedom in the model fit. The null prediction assigns equal probability to all categories. Error bars represent one standard error in estimation of the test deviance explained.*



the vowel data. For example the results show that when the vowel i is a likely label, the vowel I is also a likely label. The first two latent variables explain a significant portion of the variance in the regression coefficients for the vowels. The first latent variable distinguishes between two groups of vowels, with C:, U and u: having the most negative values and E, A, a: and Y having the most positive values. NPMR beats ridge regression here by leveraging a hidden structure among response classes.

Table 6: *Visualization of fitted regression coefficient matrices from NPMR applied to the vowel data. The matrix displayed is* $\mathbf{V}$ *in the* $\mathbf{U\Sigma V}^T$ *decomposition of the regression coefficient matrix* $\mathbf{B}$*, with columns corresponding to latent variables and rows corresponding to outcomes. The bottom row gives the entry in the diagonal matrix* $\mathbf{\Sigma}$ *corresponding to the latent variable.*

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

6 Discussion

The potential for reduced-rank multinomial regression to leverage the underlying structure among response categories has been recognized in the past. But the computational cost for the state-of-the-art algorithm for fitting such a model is so great as to make it infeasible to apply to a dataset as large as the baseball play-by-play data in the present work. Using a convex relaxation of the problem, by penalizing the nuclear norm of the coefficient matrix instead of its rank, leads to better results.

The interpretation of the results on the baseball data is promising in how it coalesces with modern baseball understanding. Specifically, the NPMR model has quantitative implications on leveraging the structure in PA outcomes to better jointly estimate outcome probabilities. Additional application to vowel recognition in speech shows improved out-of-sample predictive performance, relative to ridge regression. This

matches the intuition that NPMR is well-suited to multinomial regression in the presence of a generic structure among the response categories. We recommend the use of NPMR for any multinomial regression problem for which there is some nonordinal structure among the outcome categories.

Acknowledgments

The authors would like to thank Hristo Paskov, Reza Takapoui and Lucas Janson for helpful discussions, as well as Balasubramanian Narasimhan for computational assistance. We are grateful to Andreas Groll and an anonymous reviewer for a careful review and comments that led to improvements to this work.

References

Albert, J. (2016). Improved component predictions of batting and pitching measures. *Journal of Quantitative Analysis in Sports*, **12**(2), 73–85.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society B*, **46**(1), 1–30.

Baumer, B. and Zimbalist, A. (2014). *The Sabermetric Revolution*. University of Pennsylvania Press, Philadelphia.

Bhatia, R. (1997). *Matix Analysis*. Springer, New York.

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics*, **2**(1), 113–152.

Chen, K., Dong, H., and Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, **100**(4), 901–920.

Deterding, D. H. (1990). Speaker normalisation for automatic speech recognition. *PhD dissertation, University of Cambridge*.

Efron, B. and Morris, C. N. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, **70**(350), 311–319.

Friedman, J., Hastie, T. J., and Tibshirani, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22.

Grant, M., Boyd, S., and Ye, Y. (2008). *CVX: Matlab Software for Disciplined Convex Programming*. CVX Research. URL http://www.cvxr.com/.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine*, **13**(16), 1665–1677.

Hastie, T. J., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data mining, inference and prediction*. Springer Series in Statistics. Springer, 2nd edition.

Hastie, T. J., Tibshirani, R. J., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and its Generalizations*. Monographs on Statistics and Applied Probability. CRC Press, 1st edition.

Judge, J. and BP Stats Team (2015). DRA: An in-depth discussion. http://www.baseballprospectus.com/article.php?articleid=26196.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, **140**(22), 1–55.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, **78**(381), 47–55.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. Technical Report 2007076, Université Catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Null, B. (2009). Modeling baseball player ability with a nested Dirichlet distribution. *Journal of Quantitative Analysis in Sports*, **5**(2).

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Robinson, A. J. (1989). Dynamic error propagation networks. *PhD dissertation, University of Cambridge*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**(1), 267–288.

Tutz, G. and Gertheiss, J. (2016). Regularized regression for categorical data (with discussion and rejoinder). *Statistical Modelling*, **16**(3), 161–260.

Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, **32**(10), 1–34.

Yee, T. W. and Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling*, **3**(1), 15–41.

A Appendix

Identifiability of multinomial logistic regression model

We observe in Section 2 that the model (2.1) is not identifiable: For any $a \in \mathbb{R}$ and $\mathbf{c} \in \mathbb{R}^p$,

$$\frac{e^{\alpha_k - a + \mathbf{x}_i^T(\boldsymbol{\beta}_k - \mathbf{c})}}{\sum_{\ell=1}^K e^{\alpha_\ell - a + \mathbf{x}_i^T(\boldsymbol{\beta}_\ell - \mathbf{c})}} = \frac{e^{-a - \mathbf{x}_i^T\mathbf{c}} e^{\alpha_k + \mathbf{x}_i^T\boldsymbol{\beta}_k}}{e^{-a - \mathbf{x}_i^T\mathbf{c}} \sum_{\ell=1}^K e^{\alpha_\ell + \mathbf{x}_i^T\boldsymbol{\beta}_\ell}} = \frac{e^{\alpha_k + \mathbf{x}_i^T\boldsymbol{\beta}_k}}{\sum_{\ell=1}^K e^{\alpha_\ell + \mathbf{x}_i^T\boldsymbol{\beta}_\ell}}$$

Hence, $(\alpha, \mathbf{B})$ and $(\alpha - a1_K, \mathbf{B} - \mathbf{c}1_K^T)$ have the same likelihood. The ridge penalty in (3.7) provides a natural resolution. Any solution to this problem must satisfy

$$||\mathbf{B}||_F^2 = \min_{\mathbf{c} \in \mathbb{R}^p} ||\mathbf{B} - \mathbf{c}1_K^T||_F^2, \tag{A.1}$$

because otherwise $\mathbf{B}$ can be replaced by $\mathbf{B} - \mathbf{c}1_K^T$ with a smaller norm but the same likelihood and hence a lesser objective. Note that the optimization problem on the right-hand side of (A.1) is separable in the entries of $\mathbf{c}$ and has the unique solution $\mathbf{c}^* = \frac{1}{K}\mathbf{B}1_K$, meaning that the rows of $\mathbf{B}$ in the solution must have mean zero. The unpenalized intercept α still lacks identifiability, but we may take it to have mean zero as well.

Similarly, the NPMR solution must satisfy

$$||\mathbf{B}||_* = \min_{\mathbf{c} \in \mathbb{R}^p} ||\mathbf{B} - \mathbf{c}1_K^T||_*. \tag{A.2}$$

Whether this optimization problem always (for any $\mathbf{B} \in \mathbb{R}^{p \times K}$) has a unique solution is an open question. We speculate that it does and that the unique solution is $\mathbf{c}^* =$

$\frac{1}{K}\mathbf{B}1_K$. As evidence, each fit of NPMR in the present manuscript has a solution with zero-mean rows. As further evidence, we have used the MATLAB software CVX (Grant et al., 2008) to solve (A.2) for several randomly generated matrices $\mathbf{B}$, and each time the solution has been $\mathbf{c}^* = \frac{1}{K}\mathbf{B}1_K$.

Note that $\mathbf{c}^* = \frac{1}{K}\mathbf{B}1_K$ must always be *a* solution to (A.2). To see this, note that

$$\mathbf{B} - \mathbf{c}^* 1_K^T = \mathbf{B} - \frac{1}{K}\mathbf{B}1_K 1_K^T = \mathbf{B}\left(\mathbf{I} - \frac{1}{K}1_K 1_K^T\right) = \mathbf{B}(\mathbf{I} - \mathbf{H}),$$

where $\mathbf{H} = 1_K(1_K^T 1_K)^{-1}1_K^T$ is a projection matrix. Hence, $\mathbf{I} - \mathbf{H}$ is also a projection matrix and has spectral norm (maximum singular value) $||\mathbf{I} - \mathbf{H}||_\infty = 1$. By Hölder's inequality for Schatten p-norms (Bhatia, 1997),

$$||\mathbf{B}(\mathbf{I} - \mathbf{H})||_* \leq ||\mathbf{B}||_* ||\mathbf{I} - \mathbf{H}||_\infty = ||\mathbf{B}||_*,$$

so for any $\mathbf{B} \in \mathbb{R}^{p \times K}$,

$$||\mathbf{B} - \frac{1}{K}\mathbf{B}1_K 1_K^T||_* \leq ||\mathbf{B}||_*.$$

In order words, the nuclear norm can always be decreased, or at least not increased, by centering the rows to have mean zero.

The problem with a lack of identifiability in the multimonial regression model comes in the interpretation of the regression coefficients. When comparing coefficients across variables for the same outcome class, it is concerning that an arbitrary increase in either coefficient can corresond to the same fitted probabilities (if that same increase applys to all other coefficients for the same variable). This does not apply to any of the interpretation in Section 5.3, but in the absence of certainty that there is a unique solution to (A.2), we take the NPMR solution to be the one for which the mean of α and the row means of $\mathbf{B}$ are zero.

Proof of Lipschitz condition for multinomial log likelihood

We prove that the multinomial log-likelihood $\ell(\alpha, \mathbf{B}; \mathbf{X}, Y)$ from (3.2) has Lipschitz gradient with constant $L = \sqrt{K}||\mathbf{X}||_F^2$. Assume (without loss of generality) that the covariate matrix $\mathbf{X}$ has a column of 1s encoding the intercept, so $\alpha = 0$. The gradient of $\ell(\mathbf{B}; \mathbf{X}, Y)$ with respect to $\mathbf{B}$ is given by $\mathbf{X}^T(\mathbf{Y} - \mathbf{P})$, where $\mathbf{Y}$ and $\mathbf{P}$ are defined as in (3.4). What we must show is that, for any $\mathbf{B}, \mathbf{B}' \in \mathbb{R}^{p \times K}$:

$$||\mathbf{X}^T(\mathbf{Y} - \mathbf{P}) - \mathbf{X}^T(\mathbf{Y} - \mathbf{P}')||_F \leq \sqrt{K}||\mathbf{X}||_F^2||\mathbf{B} - \mathbf{B}'||_F. \tag{A.3}$$

Recall that $\mathbf{P}$ is a function of $\mathbf{B}$, so $\mathbf{P}'$ corresponds to $\mathbf{B}'$.

Consider a single entry $\mathbf{P}_{ik}$ of $\mathbf{P}$. Note that the gradient of $\mathbf{P}_{ik}$ with respect to $\mathbf{B}$ is given by $\mathbf{x}_i\mathbf{w}_{ik}^T$, where $\mathbf{w}_{ik} \in \mathbb{R}^p$ and

$$(\mathbf{w}_{ik})_j = \begin{cases} -\mathbf{P}_{ik}\mathbf{P}_{ij} & j \neq k \\ \mathbf{P}_{ik}(1 - \mathbf{P}_{ik}) & j = k \end{cases}.$$

For any $\mathbf{P} \in (0, 1)^{n \times K}$,

$$||\mathbf{w}_{ik}||_2 \leq ||\mathbf{w}_{ik}||_1 = \mathbf{P}_{ik}(1 - \mathbf{P}_{ik}) + \mathbf{P}_{ik}\sum_{j \neq k}\mathbf{P}_{jk} = 2\mathbf{P}_{ik}(1 - \mathbf{P}_{ik}) \leq \frac{1}{2}.$$

This implies that the norm of the gradient of $\mathbf{P}_{ik}$ is bounded above by the inequality $||\mathbf{x}_i\mathbf{w}_{ik}^T||_F \leq ||\mathbf{x}_i||_2||\mathbf{w}_{ik}^T||_F \leq ||\mathbf{x}_i||_2$. So for any $\mathbf{B}, \mathbf{B}' \in \mathbb{R}^{p \times K}$:

$$|\mathbf{P}_{ik} - \mathbf{P}'_{ik}| \leq ||\mathbf{x}_i||_2||\mathbf{B} - \mathbf{B}'||_F. \tag{A.4}$$

Now we are ready to prove (A.3).

$$
\begin{aligned}
||\mathbf{X}^T(\mathbf{Y}-\mathbf{P})-\mathbf{X}^T(\mathbf{Y}-\mathbf{P}')||_F &= ||\mathbf{X}^T(\mathbf{P}-\mathbf{P}')||_F \\
&\leq ||\mathbf{X}||_F||\mathbf{P}-\mathbf{P}'||_F \\
&= ||\mathbf{X}||_F\sqrt{\sum_{i=1}^{n}\sum_{k=1}^{K}(\mathbf{P}_{ik}-\mathbf{P}'_{ik})^2} \\
&\leq ||\mathbf{X}||_F\sqrt{\sum_{i=1}^{n}\sum_{k=1}^{K}||\mathbf{x}_i||_2^2||\mathbf{B}-\mathbf{B}'||_F^2} \qquad \text{from (A.4)} \\
&= ||\mathbf{X}||_F\sqrt{K||\mathbf{B}-\mathbf{B}'||_F^2\sum_{i=1}^{n}||\mathbf{x}_i||_2^2} \\
&= ||\mathbf{X}||_F\sqrt{K||\mathbf{B}-\mathbf{B}'||_F^2||\mathbf{X}||_F^2} \\
&= \sqrt{K}||\mathbf{X}||_F^2||\mathbf{B}-\mathbf{B}'||_F \qquad \blacksquare
\end{aligned}
$$