Improving pairwise comparison models using Empirical Bayes shrinkage

Stephen Ragain  
Stanford University  
Stanford, CA  
sragain@stanford.edu

Alexander Peysakhovich  
Facebook Artificial Intelligence Research  
alex.peys@fb.com

Johan Ugander  
Stanford University  
Stanford, CA  
jugander@stanford.edu

ABSTRACT
Comparison data arises in many important contexts, e.g. shopping, web clicks, or sports competitions. Typically we are given a dataset of comparisons and wish to train a model to make predictions about the outcome of unseen comparisons. In many cases available datasets have relatively few comparisons (e.g. there are only so many NFL games per year) or efficiency is important (e.g. we want to quickly estimate the relative appeal of a product). In such settings it is well known that shrinkage estimators outperform maximum likelihood estimators. A complicating matter is that standard comparison models such as the conditional multinomial logit model are only models of conditional outcomes (who wins) and not of comparisons themselves (who competes). As such, different models of the comparison process lead to different shrinkage estimators. In this work we derive a collection of methods for estimating the pairwise uncertainty of pairwise predictions based on different assumptions about the comparison process. These uncertainty estimates allow us both to examine model uncertainty as well as perform Empirical Bayes shrinkage estimation of the model parameters. We demonstrate that our shrunk estimators outperform standard maximum likelihood methods on real comparison data from online comparison surveys as well as from several sports contexts.

1 INTRODUCTION
Comparison-based choice data, where one item is selected from a choice set, is ubiquitous in both online and offline applications. The use of past choices to predict future choices is at the heart of online commerce (e.g. recommender systems [44]), forecasting competitions (e.g. Elo ratings in games like chess [16]), and social science (e.g. modeling demand systems in economics [8]). A popular workhorse for this task is the conditional multinomial logit (MNL) model [29, 30], which is usually estimated using maximum likelihood methods [21, 25].

It is well known that maximum likelihood methods are typically consistent (that is, learn the true parameters as the amount of data goes to infinity), but can have high variance. This can be a problem in datasets where the effective number of data points per model parameter is small. For example, we may wish to learn a new chess competitor’s rating quickly, or our dataset may be restricted to be of a finite size (each consumer only makes a fixed amount of choices per week), or certain comparisons may be more likely than others (such as in sports leagues with conference structures).

In such “small data” situations, shrunken (regularized) estimators often perform much better than maximum likelihood estimators. In a classic example, Efron and Morris [14] showed that if we want to estimate a baseball batter’s skill from a sample of at-bats, we should shrink each batter’s estimated skill independently towards the global average by an amount that is proportional to the variance in our estimate. Procedures like this one are often called Empirical Bayes [7], and this shrinkage can substantially change estimates and also greatly improve out-of-sample prediction.

To apply a shrinkage estimator in practice one needs to know how much to shrink and what to shrink toward. When estimating multiple means simultaneously as Efron and Morris did, the choices are straight-forward. In comparison models, however, the direction and magnitude of shrinkage is intimately tied to model uncertainty, specifically uncertainty about pairwise comparisons and not absolute values of individual parameters. In this paper we propose a family of simple methods for estimating different notions of parameter uncertainty associated with any conditional multinomial logit model, which allows us both to examine the model’s weaknesses and to improve parameter estimates via shrinkage.

The MNL model, like many choice models, assumes that observed data results from a stochastic process where items have an underlying “score.” The MNL model further assumes that the probability that one item is chosen over another is proportional to these scores (Luce called this the ratio scale representation [29]). Given a comparison dataset, the task of the analyst is to estimate these latent scores from observed comparisons and then use them to make subsequent predictions in future comparisons.

Our contribution is to develop an Empirical Bayes-like procedure for shrinkage estimation with the MNL model. In the datasets we will consider each data point as a pair of items and a choice from that pair. This means that the uncertainty in individual parameters (e.g. scores for each object) is generally intertwined. Furthermore, this uncertainty can be greatly affected by the sampling process. Consider a season of American football (NFL) games, where games can be interpreted as comparisons between teams. In the NFL teams are split into conferences and divisions and play mostly other teams in the same conference. This split means that given a season of data we can be relatively sure how to rank two teams within one conference because we have seen many games within that conference. However, we may at the same time be quite uncertain about the outcome when the two teams are in different conferences because we have relatively few comparisons of teams across conferences.

We develop a family of procedures for shrinkage estimation with the MNL model. The key intuition behind all our approaches is that unlike in standard statistical models where we focus on the uncertainty of a parameter estimate, we need to focus on pairwise uncertainty among parameters. Towards this goal, we focus on the covariance matrix (or the related Fisher information matrix) of the estimated parameters, but can only do so under some assumptions about the distribution of how pairs are compared. How items are compared is generally not the domain of discrete choice modeling,
which models the probability of different outcome of a comparison conditional on the comparison being made, rather than the probability that the comparison is made. We discuss the importance of modeling this distribution in the shrinkage estimators we develop.

Given a set of assumptions about uncertainty, we adapt the classic James–Stein shrinkage estimator [22] to the choice problem as follows: first, we compute the maximum likelihood estimate of the model parameters and a covariance matrix of its parameters, based on a specific notion of uncertainty. We use this (generally non-diagonal) matrix to shrink our maximum likelihood estimates. We present four methods of computing a covariance matrix using bootstrap-based methods as well as two using Fisher information. We discuss which methods of covariance are most suited to different methods of data generation and apply our procedures to several real datasets. We find that the properly shrunk estimators have better predictive power than the standard MLE.

While we focus on the MNL model, being the most popular discrete choice model, similar procedures can be applied to choice models other than MNL such as Thurstonian models [46], random utility models (RUMs) [45], mixed logit [47], nested logit [20], elimination by aspects [48], Markov chain-based models [5, 40], the Blade-Chest model [10], and Elo ratings [16]. We leave these extensions to future work.

2 CHOICE MODELS

Our main object of interest will be a universe $U$ of items, where we denote the items by indexes $\{1, \ldots, n\}$. At training time we will be given a dataset $D$ consisting of ordered tuples $(i, j)$ representing that $i$ has been chosen over $j$, or equivalently $i$ has been chosen from the set $\{i, j\}$. We focus on shrinking choice models learned from binary choice sets (pairs), but the ideas are easily extensible to the general case of choices from sets of arbitrary and mixed size. At test time we will receive a new dataset of binary choice sets and our goal will be to predict which item will be chosen from each set. We will focus on random utility models.

Definition 1. A random utility model of choice consists of two components:

- A utility vector $\gamma$ of length $n$ where $\gamma_i > 0$ is a underlying numeric "quality" of item $i$.
- A choice rule, which we denote $p_{ij} = f(\gamma_i, \gamma_j)$, that determines the probability $i$ is chosen from $\{i, j\}$.

We will give specific emphasis to two such models in this work: the Bradley-Terry-Luce model (BTL) [6] for pairwise comparisons and the Rasch model [41] of item-response outcomes, both being special cases of the conditional Multinomial Logit (MNL) [34] model (sometimes also called the Plackett-Luce model [39]).

Definition 2. The Bradley-Terry-Luce (BTL) model uses the choice function

$$p_{ij} = f(\gamma_i, \gamma_j) = \frac{\gamma_i}{\gamma_i + \gamma_j}.$$

The BTL model is scale-invariant (multiplying $\gamma$ by a scalar factor results in the same probabilities). A common convention is to normalize the parameter vector $\gamma$ such that $||\gamma||_1 = 1$.

The Rasch model is a special case of the BTL model specifically targeted at comparisons between items from disjoint contexts such as (student, question) pairs in test-taking or (offense, defense) pairs in sports outcomes. The Rasch model divides the universe of alternatives $U$ into two disjoint subsets $U_1, U_2$ with comparisons only between items of different types. As with BTL, the Rasch model is scale invariant. To make it identified one can split the $\gamma$ vector into $\gamma^{U_1}$ and $\gamma^{U_2}$, and then require that $||\gamma^{U_1}||_1 + ||\gamma^{U_2}||_1 = 1$. Because the BTL model is a special case of the MNL and our work generalizes to the MNL model, we will refer to $\gamma$ as MNL parameters hereafter.

The MNL model can be thought of as a choice process where items have real underlying qualities but at choice time Gumbel noise is added to both items’ qualities, with choices made based on which item’s realized quality is higher [29, 31, 49]. The Gumbel noise can be changed to Gaussian noise to transform the BTL and Rasch models into Thurstone [46] and Lawley-Lord models [27], respectively. There is very little practical difference between the BTL/Rasch (logit) and Thurstone/Lawley-Lord (probit) models of discrete choice.

3 BIAS-VARIANCE TRADEOFFS IN MNL ESTIMATION

We now discuss the standard method for estimating the parameter vector $\gamma$, maximum likelihood estimation (MLE). Given the choice rule in the MNL model we can write the log-likelihood of $\gamma$ given choice dataset $D$ with $N$ datapoints and generic entry $i_k$ chosen from $\{i_k, j_k\}$ as:

$$\ell(\gamma; D) = \sum_{(i_k, j_k) \in D} \log(p_{i_k, j_k}) - \log(p_{i_k} + p_{j_k}).$$

It is well known that the MLE only exists under certain conditions on $D$. When $D$ contains only pairwise comparisons we can construct the auxiliary variables $M_{ij}$ as the number of times $i$ is chosen from $\{i, j\}$, and can consider these variables as a directed graph defined on the set of elements. We call this graph the comparison graph of $D$. There exists a unique maximizer to the MLE problem if and only if the comparison graph of $D$ is strongly connected [18]. There are many algorithms for computing this maximum [13, 32].

When the distinction is relevant we will let $\hat{\gamma}_{MLE}$ denote the maximum likelihood estimator while letting $\gamma_{MLE}(D)$ denote the maximum likelihood estimate for a given dataset $D$. The MLE is an unbiased estimator, that is, if the model is the true data generating process then $E[\gamma_{MLE}(D)] = \gamma$. We are however still faced with a bias-variance tradeoff.

A common way to make this bias-variance tradeoff comes from Empirical Bayes estimation. The simplest version is the James–Stein estimator (JS) which takes the MLE estimates and shrinks them toward a fixed vector. Though the JS is a biased estimator, it has lower variance than the MLE and gives better out-of-sample prediction in terms of mean-squared error. For multivariate Gaussian data, the JS estimator has a closed form solution [4] given by:

$$\hat{\gamma}_{JS} = (I - \Sigma(\Sigma + A)^{-1})\hat{\gamma}_{MLE} + \Sigma(\Sigma + A)^{-1}u,$$

where $\Sigma$ is the covariance matrix of $\gamma_{MLE}$ under the data $D$ and $u$ and $A$ are the respective mean and covariance of the true parameters $\gamma^*$. When the true parameters $\gamma$ are unknown (which is most of
the time) sample estimates of the parameters are used as plug-ins to the equation.

Applying the JS estimator to our problem is made difficult by two complications. First, the process we are modeling is non-Gaussian, as the quality parameters are constrained to live on a simplex, but this is a minor point in practice since we can proceed under the assumption that we are simply modeling the first two moments of the data. More importantly, the covariance matrix $\Sigma$ is tightly connected to a data-generating process that, for our setting, is unknown. Given a choice dataset $\mathcal{D}$ and the corresponding maximum likelihood estimator $\hat{\gamma}_{MLE}$, we should expect $\hat{\gamma}_{MLE}$ to have low variance when predicting matchups that occur often in $\mathcal{D}$, and high variance for pairs of items which do not appear often. Choice models such as MNL do not model which comparisons are being made, only the outcome; indeed, this is why MNL is called the conditional multinomial logit: it predicts outcomes conditional on what comparisons are made.

Though the estimator learns one parameter per item, the data-generating process can create high amounts of covariance between parameters – for example, if items can be split into two categories (e.g. two football conferences) then the MLE may deliver a low variance estimate of the probability of a team $x$ beating another team $y$ when both teams come from the same conference, but a high variance estimate of the same probability when teams are from different conferences.

Thus a main question will be how to estimate $\Sigma$, $A$ and $u$ for comparison data. We will refer to the estimator of $\gamma$ based on estimates of $\Sigma$, $A$ and $u$ as $\hat{\gamma}_{SHR}$, rather than $\hat{\gamma}_{JS}$, and reserve $\hat{\gamma}_{JS}$ for cases where those quantities are known.

### 3.1 What determines MLE uncertainty?

A central thesis of this work is that the MNL parameters $\gamma$ alone do not provide a generative model for data, because they do not give us information about which alternatives will be compared. The data arises both from randomness in the choice rule as well as randomness in what alternatives are compared.

Consider the symmetric matrix $B_D$ defined as counting the number of times each pair is compared: $B_D^{ij}$ is the number of times $(i, j)$ or $(j, i)$ are in $\mathcal{D}$. In this work we assume that the winners of matchups in $\mathcal{D}$ are independent of which other choice sets appear in $\mathcal{D}$. This means that the probability of observed the data $\mathcal{D}$ can be written as:

$$Pr(\mathcal{D}) = Pr(B_D) \left( \prod_{(i,j) \in \mathcal{D}} p_{ij} \right).$$

The parameters of the choice model only affect the latter product of choice probabilities, but the variance of the MLE choice parameters for a given $\mathcal{D}$ is also affected by probability of $B_D$, the matchup structure observed in $\mathcal{D}$.

In this work, we only consider two very simple models for $Pr(B_D)$, one of which fixes $B_D$, fixing which matchups are seen based on the observe data, and the other of which samples matchups with replacement from $\mathcal{D}$. We introduce this notation to highlight that a formalization of the distribution of $\hat{\gamma}_{MLE}$ requires us to make such an assumption. Whether we can e.g. leverage domain knowledge to model $Pr(B_D)$ in some principled way to further improve out-of-sample prediction is an interesting direction for future work, but the best approach is likely domain-dependent.

### 4 ESTIMATING $\Sigma$, $A$, AND $u$

The covariance matrix $\Sigma$ can be estimated using either analytic methods based on asymptotic theory or using bootstrap-based methods. These methods each have different costs and benefits. We will introduce the methods in this section and compare them in experiments with real data.

#### 4.1 Fisher Information-based methods

The Fisher Information of the data $\mathcal{D}$ can be used to estimate $\Sigma$. There are two ways to express the Fisher information:

**Definition 3.** The observed Fisher information is a function of the dataset and $\hat{\gamma}_{MLE}$ and we refer to it as $\mathcal{I}(\hat{\gamma}_{MLE}, \mathcal{D})$. It is computed by taking the sample mean of the log-likelihood function evaluated at $\hat{\gamma}_{MLE}$ with the data $\mathcal{D}$.

**Definition 4.** The expected Fisher information is a function of the dataset and $\hat{\gamma}_{MLE}$ and we refer to it as $\mathcal{I}(\hat{\gamma}_{MLE}, \mathcal{D})$. It is calculated by taking the dataset $\mathcal{D}$, keeping the distribution of choice sets but replacing the choices with draws from the MNL model implied by $\hat{\gamma}_{MLE}$.

We can think of observed Fisher Information as “non-parametric,” in the sense that it depends on the data and the likelihood alone, whereas the expected Fisher information is computed assuming that $\hat{\gamma}_{MLE}$ is the true model behind the choices in the data. Typically the expected Fisher information for a parametric model is a function of the model parameters alone, e.g. only $\gamma$, and not the observed data. However, we again note that the choice model parameters $\gamma$ do not control the distribution of which choice sets are observed, so $\mathcal{D}$ implicitly serves as a parameter of the information matrices because we use the distribution of choice sets in $\mathcal{D}$ to estimate the true distribution of choice sets. Here we give the derivations when using the empirical distribution of choices given by $\mathcal{D}$, with a discussion of more general choice distributions given later.

To get an estimate of the covariance matrix $\Sigma$ we simply invert $\mathcal{I}(\hat{\gamma}_{MLE}, \mathcal{D})$ or $\mathcal{I}(\hat{\gamma}_{MLE}, \mathcal{D})$ and divide by the number of observed matchup sets $N = |\mathcal{D}|$. We refer to these estimators as $\hat{\Sigma}_J$ and $\hat{\Sigma}_F$, respectively. While some work exists on the asymptotic properties of these estimators [1], we are mostly interested in the finite-sample case, which has its own issues. In particular, there are issues of numerical stability with the matrix inverse and the potential singularity of the matrix. We discuss ways circumvent inverting the information in the following section. Note that the sparsity pattern of the non-zero entries off the diagonal of both $I$ and $\mathcal{I}$ are the same as the sparsity pattern of the comparison graph for $\mathcal{D}$, giving us a concrete connection between shrinkage using these matrices and the observed data.

**Implicit shrinkage estimation using Fisher Information.** For larger $n$, especially for comparisons arising from the Rasch model, the Fisher information may not be invertible. However we do not need to invert an estimator $\hat{\Sigma}$ for $\Sigma^{-1}$ to estimate the shrinkage when $\Sigma$ is invertible. We only need to estimate the matrix

$$R = \hat{\Sigma}(\Sigma + A)^{-1}$$

3
whether or not to use parametric methods as part of the resampling

4.2 Bootstrap-based methods

liers. The canonical estimator to first consider for
estimator in finite samples, the sample covariance matrix generated
estimates of \( \Sigma \) rather than using \( S^{-1} \) as an
so if we use \( S \) is an estimator of \( \Sigma^{-1} \) rather than using \( S^{-1} \) as an estimator of \( \Sigma \), we have that
\[
R^{-1} = (\Sigma + A)\Sigma^{-1} = I + A\Sigma^{-1}
\]
Examples of estimators \( S \) for the inverse of the covariance include the number of observed matchups \( N \) times the observed Fisher information \( J(\hat{\gamma}_{MLE}, D) \) or expected Fisher information \( I(\hat{\gamma}_{MLE}, D) \).

4.3 Which bootstrap? Which information

Another way to estimate \( \Sigma \) is to employ a bootstrapping method to
generate \( K \) replicates of \( D \), obtain estimates \( \hat{\gamma}_{MLE}^{(1)}, \ldots, \hat{\gamma}_{MLE}^{(K)} \) and plug these sample estimates into standard estimators for covariance matrices. The simplest way is to construct each replicate is by sampling items from \( D \) with replacement. This procedure may seem attractive, but will often fail because even though \( D \) may induce a strongly connected graph of comparisons – a requirement for maximum likelihood estimation – a replicate of \( D \) may not. Consider our running example of two football conferences: if each conference is strongly connected but there are only two games between conferences, one won by a team from the first conference and one by a team from the other, then the graph of \( D \) is strongly connected but if either of those two cross-conference games is not sampled in a particular replicate, the MLE for that replicate will be undefined.

When bootstrapping often leads to a comparison graph that is not strongly connected, we can get around the above issue by using a block-bootstrap [11]. In the block-bootstrap procedure we take every choice set \( \{i,j\} \) that is represented in \( D \) and construct a replicate dataset by resampling with replacement, for each pair, among the items in \( D \) that compare that pair.

Note that this procedure has the weakness that if \( D \) contains only a single instance of a choice from \( \{i,j\} \), or if all of the choices between the pair are the same, those same choices appear in all blocked non-parametric bootstraps. Because we expect Empirical Bayes methods to be applied specifically in domains with data constraints, we will consider a parametric bootstrap as a way around these issues.

Given a model \( \hat{\gamma}_{MLE} \) we construct a parametric bootstrap replicate by taking the dataset \( D \) and replacing the actual choices observed for each entry with a sample from the MNL model with parameters \( \hat{\gamma}_{MLE} \). Here we can choose to either fix the pairs that are compared – fixing \( B^D \) from Section 3.1 – or sample with replacement from the pairs compared in \( D \).

Thus deciding whether or not to block the data on the pairs and whether or not to use parametric methods as part of the resampling scheme gives us four possible approaches to bootstrapping choice data: blocked and parametric (bp), blocked and non-parametric (bnp), non-blocked and parametric (nb,p), and non-blocked and non-parametric (nbnp).

Shrinking the variance estimate. While \( \hat{\gamma}_{MLE} \) is a consistent estimator in finite samples, the sample covariance matrix generated by the bootstrapping procedures above can be sensitive to outliers. The canonical estimator to first consider for \( \Sigma \) is the sample covariance matrix \( \hat{\Sigma} \) is
\[
\hat{\Sigma}_S = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\gamma}_{MLE}^{(k)} - \bar{\hat{\gamma}}_{MLE}) (\hat{\gamma}_{MLE}^{(k)} - \bar{\hat{\gamma}}_{MLE})^T
\]
where \( \bar{\hat{\gamma}}_{MLE} \) is simply the mean \( K^{-1} \sum_k \hat{\gamma}_{MLE}^{(k)} \) of the sample MLE vectors.

In these cases we can produce more effective estimators of \( \Sigma \) by shrinking the sample covariance as well, toward a diagonal matrix of the mean variance. We will use the estimator introduced by [28]:
\[
\hat{\Sigma}_{SHR} = (1 - \nu)\hat{\Sigma}_S + \nu \hat{\sigma} \hat{\Sigma}
\]
where \( \hat{\sigma} \) is the (scalar) mean sample variance \( \hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}_i \). The shrinkage factor \( \nu \) can be tuned by cross validation or one can methods such as those proposed by Ledoit and Wolf [28]. The Ledoit-Wolf method chooses the shrinkage factor to minimize the mean squared error in the covariance matrix, but note that we are here not interested in accurately estimating the covariance matrix other than in the service of then performing shrinkage on the quality parameters \( \gamma \). The problem of shrinking estimates of both the location and covariance simultaneously with the goal of minimize the mean squared error in the location parameters is a difficult problem known as double shrinkage [50]. Optimal double shrinkage estimators are only known for problems with diagonal covariance matrices, whereas non-diagonal covariance is fundamental to our approach. As a result, we employ and recommend selecting this shrinkage factor using cross-validation.

We find that using this shrunk variance estimate \( \hat{\Sigma}_{SHR} \) improves performance over the sample covariance in every application we consider. As a result, in our empirical results we always apply this Ledoit-Wolf shrinkage when estimating \( \Sigma \) through bootstrapping.

4.3 Which bootstrap? Which information

matrix?

The bootstrap procedures are simpler to understand and the best bootstrap methods outperform Fisher information based methods in our experiments, but bootstrapping is more computationally costly and in certain datasets finding bootstraps which are strongly connected can be intractable, and rejecting such samples may inject bias into the estimation of \( \Sigma \). When bootstrapping methods are onerous for these reasons, we recommend using Fisher information based methods for shrinkage.

Another important practical question is whether the bootstraps should resample matchups according to the parametric distribution given by the MLE rather than the empirical distribution of that matchup. We find that this distinction is especially important in practice. Consider the NFL2016 dataset, where each team plays 10 of its 16 games against opponents that it only faces once, so the blocked non-parametric bootstrap will fix this game. Further, the other 6 games come from playing each of 3 divisional opponents twice, and can only change when those games are split. As a result, the non-parametric bootstrap is unable to capture the variance in nearly all of the played games, and as a result, gives a poor estimation of \( \hat{\Sigma} \).

Choosing the correct bootstrap for a given dataset requires careful consideration of the problem at hand. One cue that the blocked
bootstrap may be appropriate is that for some domains the sampling structure of pairs is static or roughly static. For example, we find the block bootstrap is appropriate for the dataset of NFL2016 matchups we study in which the regular season schedule features similar structure year after year. Meanwhile for the MLB dataset we study comparing baseball batters and pitchers, there is less regularity as the rotation of starting pitching and batting orders are decided independently from the underlying team schedules and feature significant changes as the season unfolds based on injuries, performances, trades, etc. For that setting we find a bootstrap (non-blocked) is more appropriate.

**Using priors to ensure strongly connected data.** A common solution to MNL inference when the comparison graph is not strongly connected such is to employ a Gamma(ε, 1) prior on ̂γMLE (ignoring normalization, which does not change the model) [19, 33], which leads to a Dirichlet(ε, ε, . . . , ε) prior for the normalization of γ [9].

Noting that the conjugate of the aforementioned Dirichlet prior is the multinomial distribution and that the MNL model extends to choice sets of arbitrary size, we can smooth our data to give a well defined y by adding ε "choices" of each x ∈ U from the full set of alternatives. Although we have focused on pairwise comparisons, these larger set comparisons can also be modeled by MNL and the inference algorithms we employ uses a Dirichlet prior with ε = 10^−6. Further details appear in Appendix A.

### 4.4 Estimation of A and u

In order to estimate the covariance of the true γ∗ in our estimator ̂γMLE we use a Dirichlet prior for γ∗ centered at n · ̂γMLE where we multiply with n so that the mean is 1, giving a typically unimodal prior which becomes uniform for ̂γMLE near the uniform vector u where u_i = 1/n. The resulting covariance A has

\[
A_{ii} = \frac{\hat{γ}_{MLE,i}(1 - \hat{γ}_{MLE,i})}{n(n+1)}, \quad A_{ij} = \frac{\hat{γ}_{MLE,i} \hat{γ}_{MLE,j}}{n+1}.
\]

Following this procedure A is not strictly diagonal, which is related to the constraint ||γ||_1 = 1 coupling entries in γ. As a result this estimator does not shrink estimates of entries γ_i independently, especially if n is small. In general we choose to set u_i = 1/n for all i ∈ U, though given some domain knowledge we may choose a different vector towards which to contract ̂γMLE.

### 5 EXPERIMENTS

We now demonstrate in several datasets that analysts can benefit greatly from capturing the pairwise uncertainty of model parameters in choice models. We consider competition datasets from a variety of sports (baseball, basketball, and American football), as well as a large-scale survey of civic priorities taken as comparisons through the wikisurvey platform AllOurIdeas [43]. We begin by evaluating inference of MNL models from MNL data, where improvements from the James–Stein estimator are guaranteed. We then focus on out-of-sample prediction on real data, predicting the percent of time an alternative is chosen (e.g. win percentage of a sports team) in the dataset.

#### 5.1 Semi-synthetic data

We wish to measure the improvement in parameter estimation given by shrinkage. However, ground truth parameters are never known for any real-world dataset. Thus, we will begin with a semi-synthetic data. To construct our dataset, we begin with a real match-up structure from our NFL2016 dataset (discussed further in the next section) that contains all of the games in the 2016 NFL regular season. This season consists of N = 256 games played between n = 32 teams (16 games per team). There is a multi-year rotation of NFL schedules, all of which provide poor connectivity between the two conferences (the NFC and the AFC) [35].

Next, we randomly generate “skill” parameters for each team from the uniform distribution on the simplex and then construct win/loss records by sampling from the implied model using the real game schedule. We then fit our models using this synthetic data and see how well we recover parameters as well as predict unseen matchups.

We consider two metrics. The first is the relative improvement of the MSE with respect to γ∗ when using the inferred parameters ̂γSHR compared to ̂γMLE:

\[
α = \mathbb{E}\left[\frac{||γ − ̂γ_{MLE}||_2^2 − ||γ − ̂γ_{SHR}||_2^2}{||γ − ̂γ_{MLE}||_2^2}\right].
\]

We are also concerned with direct improvement of estimation of the pairwise probabilities themselves. For γ, γ’ let

\[
||γ − γ'||_P = \frac{1}{n^2} \sum_{i,j} \frac{γ_i γ_j - γ'_i γ'_j}{γ_i + γ_j - γ'_i - γ'_j}
\]

denote the mean difference in pairwise probabilities between an MNL model with parameters γ and with parameters γ’. Then let

\[
β = \mathbb{E}\left[\frac{||γ − ̂γ_{MLE}||_P - ||γ − ̂γ_{SHR}||_P}{||γ − ̂γ_{MLE}||_P}\right].
\]

When the model is specified correctly, shrinkage can give large increases in accuracy measured both in terms of parameters and in terms of pairwise probabilities.

Across 1000 random γ∗ and resamples of the season, we see the average MSE in recovering γ’ reduced by an average of 51% (α = 0.51) with a shrunk estimator using the expected Fisher information ̂Fγ. Likewise we observe a relative average MSE improvement of 12% (β = 0.12) on the pairwise probabilities.

#### 5.2 Out-of-sample NFL/NBA predictions

Having confirmed that shrinkage gives better parameter estimates than MLE when the underlying behavior is generated by an MNL model, we now turn to real world data. We use the real win/loss outcomes from the 2016 NFL season.

To highlight the differences in improvements between datasets of different sizes, we also introduce the NBA2016 dataset, which contains all of the games played in the 2016 NBA season. NBA2016 consists of N = 1260 games played among n = 30 teams (82 games per team). Although each team plays the majority of its games within its conference, each team plays two games against each teams in the other conference, providing less sparsity (more connectivity) relative to the NFL2016 dataset.
Because we cannot measure MSE with respect to unknown ground truth parameters, we instead note that under the MNL model, teams facing the same schedule of opponents have a win percentage that is simply a scaling of their quality parameter. The schedules are not identical for e.g. the NFL, but we accept predicting win percentage it as a suitable proxy for accuracy.

We use 2-fold cross-validation to evaluate our shrinkage estimator, averaged across 10 runs, which evaluates how well we can predict one half of the games from another half. We use only 2 folds both because this gives us smaller training sets, the data realm we predict one half of the games from another half. We use only 2 folds.

We fit \( \hat{\gamma}_{MLE} \) for an MNL model and then compute \( \hat{\gamma}_{SHR} \) with various choices of the covariance matrix estimator: using observed Fisher Information (we refer to this as \( \hat{\Sigma}_I \)), expected Fisher Information \( \hat{\Sigma}_F \), parametric and non-parametric blocked bootstrap \( \hat{\Sigma}_{b,p} \) and \( \hat{\Sigma}_{b,np} \) respectively, and the parametric and non-parametric (non-blocked) bootstrap \( \hat{\Sigma}_{nb,p} \) and \( \hat{\Sigma}_{nb,np} \) respectively.

In Figure 1 we observe that for the blocked bootstraps \( \hat{\gamma}_{SHR} \) does significantly better in out-of-sample win percentage prediction than \( \hat{\gamma}_{MLE} \) on both the NFL2016 and NBA2016 dataset. We further see that while the Fisher information estimations for \( \Sigma \) give strong gains, the best improvements come from bootstrap based estimators. We note that while the decision to use the blocked or non-blocked bootstrap has little impact on the effectiveness of the shrinkage, the parametric bootstrap is significantly more effective than the non-parametric bootstraps, particularly on the NFL2016 data. This is because the non-parametric bootstrap is unable to find the variance when the data is as sparse as the NFL2016 dataset, where many pairs of teams have played only one game that is then fixed by all non-parametric bootstraps.

The difference in improvement between NFL and NBA is due to the NFL teams playing relatively few games and the conference structure being less well connected. This is reflected in the Fisher Information matrix (Figure 1 middle) and can be seen starkly when we plot the changes in estimated skill levels for each team between \( \hat{\gamma}_{MLE} \) and \( \hat{\gamma}_{SHR} \) (Figure 1 bottom). In the NFL we see that the shrinkage changes the ordering of some teams while in the NBA data shrinkage does little beyond shrink the estimate of one or two extreme teams slightly.

Although we have focused on improvement in win rate prediction (Figure 1), we still observe improvements in mean squared error on individual matchups (equivalently, applying the Brier scoring rule to probabilities). At the level of matchups we observe MSE improvements for the NFL2016 dataset of 5.4% and 7.2% for the Fisher shrinkage with \( \hat{\Sigma}_I \) and \( \hat{\Sigma}_F \) respectively, and MSE improvements with the bootstrapped estimators of 9.2% for both \( \hat{\Sigma}_{b,p} \) and \( \hat{\Sigma}_{nb,p} \) and of 1% for \( \hat{\Sigma}_{b,np} \) and \( \hat{\Sigma}_{nb,np} \). These matchup results further highlight, in addition to the win percentage results in Figure 1, that the parametric bootstrap is important for capturing the variance in pairwise matchups that only occur once in the data, and that the Fisher shrinkage is more effective than the non-parametric bootstrap methods but not as effective as the parametric bootstrap methods. We observed similar patterns on the NBA2016 dataset, but because there are so many more matchups in that dataset (42 games per team in the training set), we see gains of less than 1% on matchup MSE.

### 5.3 Increasing survey power

We now consider a different application: increasing the power of comparison surveys. Here we use a survey dataset from the AllOurIdeas wikisurvey platform [43]. This data consists of a survey designed by the Washington Post and run on the platform, asking readers which political figure within a pair had “the worse year in Washington.” There are \( N = 143,704 \) comparisons made between \( n = 67 \) figures. The data is available on the AllOurIdeas website, http://www.allourideas.org/. Note that this particular Washington Post wikisurvey is relatively unique among wikisurveys because respondents were recruited through one of the largest news sites in the United States. Most wikisurveys operate in the small-data regime, hundreds or thousands of responses (rather than hundreds

|     | MLE  | \( \hat{\Sigma}_I \) | \( \hat{\Sigma}_F \) | \( \hat{\Sigma}_{b,p} \) | \( \hat{\Sigma}_{b,np} \) | \( \hat{\Sigma}_{nb,p} \) | \( \hat{\Sigma}_{nb,np} \) |
|-----|------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| NFL MSE | .0591 | .0525 | .0499 | .0491 | .0585 | .0491 | .0585 |
| % better | - | 11.1% | 15.5% | 16.8% | 0.9% | 16.8% | 0.9% |
| NBA MSE | .0104 | .0098 | .0098 | .0095 | .0099 | .0094 | .0099 |
| % better | - | 5.2% | 5.2% | 8.8% | 4.7% | 9.1% | 4.7% |

Figure 1: Results for the NFL2016 and NBA2016 datasets with conference structure. Top: MSEs and percentage improvement in MSE over the MLE for win percentages in the two datasets with different estimated covariance matrices. For shrinkage using the parametric block bootstrap \( \hat{\Sigma}_{b,p} \), most naturally suitable in this setting, the MSE improves by 16.8% on the NFL data. The NFL data consists of many fewer games per team and is much less strongly connected than the NBA data. This difference is reflected in the Fisher Information matrix (middle) as well as the pattern of shrinkage (bottom).
We now consider a dataset with more comparisons, many more.

covariance, giving faster estimation and avoiding concerns about across a season. Because the pitcher/batter data reflects a bipartite graph fitting the mold of a Rasch model, we should shrink batters to the mean quality of a batter and pitchers to the mean quality of a pitcher (rather than shrinking both pitchers and batters towards the same mean $\gamma = 1/n$). Letting $U_1$ be the set of pitchers and $U_2$ be the set of batters, and let $y_{U_1}^{\gamma_{ML}}, y_{U_2}^{\gamma_{ML}}$ be the restrictions of $y$ to $U_1$ and $U_2$ respectively. We thus have $u_i = \frac{1}{|U_1|} \sum_{j \in U_1} y_{U_1}^{\gamma_{ML}, j}$ for $i \in U_1$ and $u_j = \frac{1}{|U_2|} \sum_{i \in U_2} y_{U_2}^{\gamma_{ML}, i}$ for $i \in U_2$.

In the original study of shrinkage by Efron and Morris, batting averages for batters with exactly 45 at-bats were estimated using the James–Stein estimator. With the number of at-bats fixed, the shrinkage factor depends only on how much a batter’s batting average differs from the population average. The James–Stein estimator can be applied to datasets with multiple sample sizes, in which case the shrinkage factor for a batter depends on both the number of times they’ve batted as well as how their batting average differs from the population average [42]. But such a setting would not account for the differences in strength between the pitchers these batters face, as considered under a Rasch model, and is a method of shrinking averages but not parameters, meaning the output of the James–Stein shrinkage cannot be used directly for matchup prediction. Another key difference between our method and the Efron–Morris method is the consideration of covariance.

In Figure 3 we show the empirical batting averages for the 2016 MLB season, the multiple sample size (MSS) James–Stein shrunk estimates, and the batting averages when using the Rasch shrinkage developed in this work. We see that Rasch shrinkage is able to account for the strength of pitchers that the batters are facing — and shrinkage on those pitchers – while direct shrinkage of observed batting averages does not. We find through 20-fold cross validation where we train on 5% of the data and test on the remaining 95% that shrink estimators improve the prediction player win percentages (on-base percentage for batters, one minus this for pitchers) by 13.8% with the observed Fisher information and 17.2% with the expected Fisher information in terms mean squared error, relative to MLE estimates. We train on the smaller dataset in order to highlight the data realm where shrinkage is most effective. For this dataset we opt to use Fisher information because bootstrapped datasets is almost never strongly connected and resampling all 214,865 at-bats enough times to sufficiently estimate the covariance is intractable. Our inclusion of these estimates highlights the ability of Fisher information-based shrinkage to handle datasets that may be unreasonable or unruly to bootstrap.

6 CONCLUSION

In this work we examined how the covariance structure of parameter uncertainties for choice models can be used to derive shrunk estimates of the parameters. Estimating the covariance structure is itself a subtle task requiring consideration of the joint effects of randomness in comparison outcomes as well as randomness from the distribution of how comparisons are made. We developed four methods for bootstrapping comparison data based on whether we use bootstrapping to resample choices with or without blocking and non-parametrically or parametrically. We also discussed the use of observed and expected Fisher information matrices to estimate covariance, giving faster estimation and avoiding concerns about

Figure 2: The ratio of out-of-sample MSE using $\hat{\gamma}_{SHR}$ vs. $\hat{\gamma}_{MLE}$ as a function of the number of training samples for the AllOurIdeas dataset.
The top panel shows the shrinkage (left) of  MLE and  SHR, and the observed Fisher information matrix ( , right) for the matchup data. Bottom: the effect of shrinkage on estimating the average Fisher information matrix ( , left) for pitchers and batters (purple) and (teal), and the observed Fisher information matrix ( , right) for the matchup data. Note that we can see the block structure of pitchers and batters Fisher information and that the Rasch shrinkage contracts  MLE and  MLE to different baselines.

Table 1: Summary of MSE and % improvement.

|        | MLE |  | SHR |
|--------|-----|---|-----|
| MSE    | .0209 | .0180 | .0173 |
| % improvement | - | 13.8% | 17.2% |

Figure 3: Shrinkage for the Rasch model on the MLB dataset. The top panel shows the shrinkage (left) of  MLE for pitchers (purple) and  MLE, and batters (teal), and the observed Fisher information matrix ( , right) for the matchup data. Bottom: the effect of shrinkage on estimating the average skills of batters.

bootstrapped data being strongly connected, although theoretical guarantees for these estimates are typically asymptotic in the number of data, and shrinkage is most effective when there are fewer samples.

We showed how to turn both bootstrapped estimates of covariance as well as estimates based on the Fisher information matrix produce shrinkage estimators for the quality parameters of items. We showed through a variety of empirical and synthetic datasets that the estimation provides improved inference, especially on sparse and ill-connected data. We found that when feasible, shrinkage from bootstraps performed the best, though we still saw significant increases in performance using the more quickly estimated Fisher information matrices.

There are several interesting directions for future work based upon our findings here, including deriving the shrinkage for more complex choice models involving higher-dimensional embeddings of alternatives, using more complex models for the distribution of matchups to improve out-of-sample prediction when the comparisons in the choice data are from a different distribution than the test data, and developing a deeper theoretical understanding of the relationship between the distribution of the matchups and expected improvement in MSE provided by shrinkage.

Our research is an example of a larger trend of constructing specialized regularization procedures for important special cases where standard procedures are inappropriate (e.g. instrumental variable analysis [3, 36], causal inference [23], heterogeneous treatment effect estimation [2]). This trend is particularly pronounced in the social and behavioral sciences where analyses typically focus on MLE-based estimators but recent work has begun to show the promise of more modern statistical and machine learning techniques [12, 17, 24, 26, 37, 38]. Given the centrality of choice models in social science we hope our results contribute to this important endeavor.

A APPENDIX: EFFICIENTLY ESTIMATING WITH PRIORS

Bootstrapped datasets are not always strongly connected. In this appendix we discuss how to include a Dirichlet prior on the quality parameters γ of an MNL model so that the MLE exists. Adding a prior is equivalent to adding a small weighted "choice" to the data of each alternative from the full set.

The iterative Luce spectral ranking (I-LSR) algorithm introduced in [32] is both computationally and statistically efficient algorithm that we use to compute the MLE given the MNL model parameters given choice data D. It relies on iteratively estimating  MLE as the stationary distribution of a continuous time Markov chain (CTMC) whose rates are a function of the current estimate and the data D. Because this amounts to solving a linear system and the system is sparse for pairwise data but becomes dense when adding the prior "data," we show here how to efficiently solve for  MLE(ε), the MLE under the prior using a sparse linear system.

Recall that  M is a matrix where  is the number of times i beats j in D. Let ε be a length n vector with all entries ε. From the Markov chain interpretation we have for  =  MLE that

\[ \sum_{j \neq i} \left[ \frac{M_{ij}}{y_i + y_j} + \epsilon_i \right] y_i = \sum_{j \neq i} \left( \frac{C_{ij}}{y_i + y_j} + \epsilon_i \right) y_j, \]

which makes γ the solution to the balance equations of the underlying CTMC,  Q = 0 where

\[ Q_{ij} = \frac{M_{ij}}{y_i + y_j} + \epsilon_i, \quad Q_{ii} = -\sum_{j \neq i} Q_{ij} = -(n-1)\epsilon_i - \sum_{j \neq i} \frac{M_{ij}}{y_i + y_j}. \]

Let  be the rate matrix of the CTMC corresponding to the original matrix. Then  =  −diag(ε) + εT where 1 is a column of ones. So  =  −m(εy) + ||y||1ε where εy is a column vector of element wise multiplication and for ||y||1 = 1 (by assumption) we have  = ne * (γ - 1).

Let  be the rate matrix of the CTMC corresponding to the original matrix. Then  =  −diag(ε) + εT where 1 is a column of ones. So  =  −m(εy) + ||y||1ε where εy is a column vector of element wise multiplication and for ||y||1 = 1 (by assumption) we have  = ne * (γ - 1) −  = 0. None of the guarantees of the algorithm change because we have shown that this system is
equivalent to the system solved by I-LSR under Q. The advantage here is that Q is always dense, regardless of the sparsity of M, so when Q is sparse this equivalent system is still fast to solve.

REFERENCES

[1] Abt, M., and Wens, W. J. Fisher information and maximum-likelihood estimation of covariance parameters in gaussian stochastic processes. Canadian Journal of Statistics 26, 1 (1998), 127–137.
[2] Ayto, S., and Drens, G. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113, 27 (2016), 7353–7360.
[3] Belle, A., Chen, D., Chernozhukov, V., and Hansen, C. Sparse models and methods for optimal instruments with an application to epidemic domain. Econometrics 89, 6 (2012), 2569–2429.
[4] Berger, J. A robust generalized bayes estimator and confidence region for a multivariate normal mean. The Annals of Statistics (1980), 716–761.
[5] Blanchet, J., Gallego, G., and Goyal, V. A markov chain approximation to choice modeling. Operations Research 64, 4 (2016), 866–905.
[6] Bradley, R. A., and Terry, M. E. Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika 39, 3/4 (1952), 324–345.
[7] Carlin, B. P., and Louis, T. A. Bayes and empirical Bayes methods for data analysis, vol. 17. Chapman & Hall/CRC Boca Raton, FL, 2000.
[8] Clayton, D. W., and Perlofo, J. M. Modern industrial organization. Pearson Higher Ed. 2015.
[9] Caron, F., and Doucet, A. Efficient bayesian inference for generalized Bradley–Terry models. Journal of Computational and Graphical Statistics 21, 1 (2012), 174–196.
[10] Chen, S., and Joachims, T. Modeling intransitivity in matchup and comparison data. In Proceedings of the ninth acm international conference on web search and data mining (2016), ACM, pp. 227–236.
[11] Chernick, M. R. González-Manteiga, W., Crujeiras, R. M., and Barros, E. B. Bootstrap methods. In International Encyclopedia of Statistical Science. Springer, 2011, pp. 169–174.
[12] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. K. Double machine learning for treatment and causal parameters. Tech. rep., cemmap working paper. Centre for Microdata Methods and Practice, 2016.
[13] Dijkstra, O. A note on the rank analysis of incomplete block designs—applications beyond the scope of existing tables. Biometrics 12, 3 (1956), 301–306.
[14] Efron, B., and Morris, C. Stein’s estimation rule and its competitors—an empirical bayes approach. Journal of the American Statistical Association 68, 341 (1973), 117–130.
[15] Efron, B., and Morris, C. Stein’s paradox in statistics. WH Freeman, 1977.
[16] Ego, A. E. The rating of chessplayers, past and present. Arco Pub., 1978.
[17] Epstein, Z., Fysakhovich, A., and Rand, D. G. The good, the bad, and the unluckily selfish. Cooperative decision-making can be predicted with high accuracy when using only three behavioral types. In Proceedings of the 2016 ACM Conference on Economics and Computation (2016), ACM, pp. 547–559.
[18] Frazzini, L. Solution of a ranking problem from binary comparisons. The American Mathematical Monthly 64, 8 (1957), 28–33.
[19] Go, J., and Shelson, E. Bayesian inference for plackett-luce ranking models. In proceedings of the 26th annual international conference on machine learning (2009), ACM, pp. 377–384.
[20] Hausman, J., and McFadden, D. Specification tests for the multinomial logit model. Econometrica: Journal of the Econometric Society (1984), 1219–1240.
[21] Hunter, D. R. Min algorithms for generalized Bradley–Terry models. Annals of Statistics (2004), 384–406.
[22] James, W., and Stein, C. Estimation with quadratic loss. In Proceedings of the fourth Berkeley symposium on mathematical statistics and probability (1961), vol. 1, pp. 361–379.
[23] Johannsen, F., Shapit, U., and Sontag, D. Learning representations for counterfactual inference. In International Conference on Machine Learning (2016), pp. 3020–3029.
[24] Kleinerberg, J., Liang, A., and Mullainathan, S. The theory is predictive, but is it complete? An application to human perception of randomness. In Proceedings of the 2017 ACM Conference on Economics and Computation (2017), ACM, pp. 125–126.
[25] Krishnaperumal, B., Carin, L., Furerberico, M. A., and Hartemine, A. J. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. IEEE transactions on pattern analysis and machine intelligence 27, 6 (2005), 957–968.
[26] Krügel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Meta-learners for estimating heterogeneous treatment effects using machine learning. arXiv preprint arXiv:1706.05461 (2017).
[27] Lawley, D. N. On a class of problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh Section A: Mathematics 61, 3 (1943), 273–287.
[28] Ledoit, O., and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis 88, 2 (2004), 365–411.
[29] Luce, R. D. Individual Choice Behavior: A Theoretical Analysis. John Wiley and Sons, 1959.
[30] Luce, R. D. The choice axiom after twenty years. Journal of mathematical psychology 15, 3 (1977), 215–233.
[31] Manski, C. F. The structure of random utility models. Theory and decision 8, 3 (1977), 229–254.
[32] Maystre, L., and Grossejaquies, M. Fast and accurate inference of Plackett–Luce models. In Advances in Neural Information Processing Systems (2015), pp. 172–180.
[33] Maystre, L., and Grossejaquies, M. Choicerank: Identifying preferences from node traffic in networks. arXiv preprint arXiv:1610.05255 (2016).
[34] McFadden, D. Conditional logit analysis of qualitative choice behavior. In Frontiers in Econometrics. P. Zarembka, Ed. Academic Press, 1973, ch. 4, pp. 105–142.
[35] Miller, S. D. Cryptanalysis of the nfl schedule. http://sites.math.rutgers.edu/~sdmiller/Cryptanalysis-of-the-NFL-Schedule.html, 2016.
[36] Pysakhovich, A., and Eckles, D. Learning causal effects from many randomized experiments using regularized instrumental variables. In Proceedings of the The Web Conference 2018 (WWW2018) (2018).
[37] Pysakhovich, A., and Naucke, J. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. Journal of Economic Behavior & Organization 133 (2017), 573–584.
[38] Pysakhovich, A., and Rand, D. G. In-group favoritism caused by pokémon go and the use of machine learning for principled investigation of potential moderators.
[39] Plackett, R. L. The analysis of permutations. Applied Statistics (1975), 193–202.
[40] Ragan, S., and Ugander, J. Pairwise choice markov chains. In Advances in Neural Information Processing Systems (2016), pp. 3198–3206.
[41] Rask, G. Probabilistic Models for Some Intelligence and Attainment Tests. Studies in mathematical psychology. Danmarks Pedagogiske Institut, 1960.
[42] Said, C. Empirical bayes for multiple sample sizes. chris-said.io/2017/05/03/empirical-bayes-for-multiple-sample-sizes, 2017.
[43] Salganik, M. J., and Levy, K. E. Wiki surveys: Open and quantifiable social data collection. PloS one 10, 5 (2015), e0123483.
[44] Schafer, J. B., Konstan, J., and Riedl, J. Recommender systems in e-commerce. In Proceedings of the 1st ACM conference on Electronic commerce (1999), ACM, pp. 158–166.
[45] Soffian, H. A., Parkes, D. C., and Xia, L. Preference elicitation for general random utility models. arXiv preprint arXiv:1309.6864 (2013).
[46] Thruston, L. L. A law of comparative judgment. Psychological review 34, 4 (1927), 273.
[47] Train, K. E. Discrete choice methods with simulation. Cambridge university press, 2009.
[48] Tversky, A. Elimination by aspects: A theory of choice. Psychological review 79, 4 (1972), 281.
[49] Yellott, J. I. The relationship between luce’s choice axiom, thurstone’s theory of comparative judgment, and the double exponential distribution. Journal of Mathematical Psychology 15, 2 (1977), 109–144.
[50] Zhao, Z. Double shrinkage empirical bayesian estimation for unknown and unequal variances. Statistics and Its Interface 3, 4 (2010), 533–541.