# Preference Modeling with Context-Dependent Salient Features

**Amanda Bower** [1] **Laura Balzano** [2]

## Abstract

We consider the problem of estimating a ranking on a set of items from noisy pairwise comparisons given item features. We address the fact that pairwise comparison data often reflects irrational choice, e.g. intransitivity. Our key observation is that two items compared in isolation from other items may be compared based on only a salient subset of features. Formalizing this framework, we propose the *salient feature preference model* and prove a finite sample complexity result for learning the parameters of our model and the underlying ranking with maximum likelihood estimation. We also provide empirical results that support our theoretical bounds and illustrate how our model explains systematic intransitivity. Finally we demonstrate strong performance of maximum likelihood estimation of our model on both synthetic data and two real data sets: the UT Zappos50K data set and comparison data about the compactness of legislative districts in the US.

## 1. Introduction

The problem of estimating a ranking is ubiquitous and has applications in a wide variety of areas such as recommender systems, review of scientific articles or proposals, search results, sports tournaments, and understanding human perception. Collecting full rankings of $n$ items from human users is infeasible if the number of items $n$ is large. Therefore, $k$-wise comparisons, $k < n$, are typically collected and aggregated instead. Pairwise comparisons ($k = 2$) are popular since it is believed that humans can easily and quickly answer these types of comparisons. However, it has been observed that data from $k$-wise comparisons for small $k$ often exhibit what looks like irrational choice, such as systematic intransitivity among comparisons. Common models address this issue with modeling noise, ignoring its systematic nature. We observe, as others have before us (Seshadri et al., 2019; Rosenfeld et al., 2020; Pfannschmidt et al., 2019; Kleinberg et al., 2017; Benson et al., 2016; Chen and Joachims, 2016b;a), that these systematic irrational behaviors can likely be better modeled as *rational behaviors made in context*, meaning that the particular $k$ items used in a $k$-wise comparison will affect the comparison outcome.

Consider the most common model for learning a single ranking from pairwise comparisons, the Bradley-Terry-Luce (BTL) model. In this model, there exists a judgment vector $w^* \in \mathbb{R}^d$ that indicates the favorability of each of the $d$ features of an item (e.g. for shoes: cost, width, material quality, etc), and each item has an embedding $U_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, indicating the value of each feature for that given item. Subsequently, the outcome of a comparison is made with probability related to the inner product $\langle U_i, w^* \rangle$; the larger this inner product, the more likely item $i$ will be ranked above other items to which it is compared. A key implicit assumption is that the features used to rank all $n$ items are the same features used to rank just $k$ items in the absence of the other $n - k$ items. However, we argue that the context of that particular pairwise comparison is also relevant; it is likely that when a pairwise comparison is collected, if there are a small number of features that "stand out," a person will use only these features and ignore the rest when he or she makes a comparison judgment. Otherwise, if there are no salient features between a pair of items, a person will take all features into consideration. This theory has been hypothesized by the social science community to explain violations of rational choice (Tversky, 1972; Tversky and Simonson, 1993; Rieskamp et al., 2006; Brown and Peterson, 2009; Shepard, 1964; Torgerson, 1965; Tversky, 1977; Bordalo et al., 2013). For example, (Kaufman et al., 2017) collected preference data to understand human perception of the compactness of legislative districts. They hypothesized that the features respondents use in a pairwise comparison task to judge district compactness vary from pair to pair, which explains why their data are more reliable for larger $k$. To illustrate this point, we highlight a concrete example from their experiments. Given two images of districts, they asked respondents to pick which district is more compact. When comparing district $A$ with district $B$ or district $C$ in Figure 1, one of the most salient features

[1]Department of Mathematics, University of Michigan, Ann Arbor, MI [2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI. Correspondence to: Amanda Bower <amandarg@umich.edu>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Online, PMLR 119, 2020. 

is the degree of nonconvexity. However, when comparing district $B$ and district $C$, the degree of nonconvexity is no longer a salient feature. These districts look similar on many dimensions, forcing a person to really think and consider all the features before making a judgment. Let $P_{ij}$ be the empirical probability that district $i$ beats district $j$ with respect to compactness. Then, from the experiments of (Kaufman et al., 2017), we have $P_{AB} = 100\%$, $P_{BC} = 67\%$, and $P_{AC} = 70\%$. These three districts violate strong stochastic transitivity, the requirement that if $P_{AB} \geq 50\%$ and $P_{BC} \geq 50\%$, then $P_{AC} \geq \max\{P_{AB}, P_{BC}\}$.

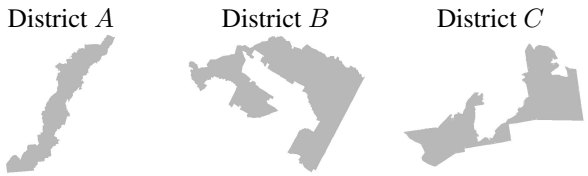


*Figure 1.* Three districts used in pairwise comparison tasks in (Kaufman et al., 2017)

We propose a novel probabilistic model called the *salient feature preference model* for pairwise comparisons such that the features used to compare two items are dependent on the context in which two items are being compared. The salient feature preference model is a variation of the standard Bradley-Terry-Luce model. At a high level, given a pair of items in $\mathbb{R}^d$, we posit that humans perform the pairwise comparison in a coordinate subspace of $\mathbb{R}^d$. The particular subspace depends on the salience of each feature of the pairs being compared. Crucially, if any human were able to rank all the items at once, he or she would compare the items in the ambient space without projection onto a smaller subspace. This single ranking in the ambient space is the ranking that we would like to estimate. Our contributions are threefold. First, we precisely formulate this model and derive the associated maximum likelihood estimator (MLE) where the log-likelihood is convex. Our model can result in intransitive preferences, despite the fact that comparisons are based off a single universal ranking. In addition, our model generalizes to unseen items and unseen pairs. Second, we then prove a necessary and sufficient identifiability condition for our model and finite sample complexity bounds for the MLE. Our result specializes to the sample complexity of the MLE for the BTL model with features, which to the best of our knowledge has not been provided in the literature. Third, we provide synthetic experiments that support our theoretical results and also illustrate scenarios where our salient feature preference model results in systematic intransitives. We also demonstrate the efficacy of our model and maximum likelihood estimation on real preference data about legislative district compactness and the `UT Zappos50K` data set.

### 1.1. Related Work

**The Bradley-Terry-Luce Model** One popular probabilistic model for pairwise comparisons is the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959). In this model, there are $n$ items each with an unknown utility $u_i$ for $i \in [n]$, and the items are ranked by sorting the utilities. The BTL model defines

$$\mathbb{P}(\text{item } i \text{ beats item } j) = \frac{e^{u_i}}{e^{u_i} + e^{u_j}}. \tag{1}$$

Although the BTL model makes strong parametric assumptions, it has been analyzed extensively by both the machine learning and social science community and has been applied in practice. For instance, the World Chess Federation has used a variation of the BTL model in the past for ranking chess players (Menke and Martinez, 2008). The sample complexity of learning the utilities or the ranking of the items with maximum likelihood estimation (MLE) has been studied recently in (Rajkumar and Agarwal, 2014; Negahban et al., 2016). Moreover, there is a recent line of work that analyzes the sample complexity of learning the utilities with MLE and other algorithms under several variations of the BTL model, including when the items have features that may or may not be known (Li et al., 2018; Oh et al., 2015; Lu and Negahban, 2015; Park et al., 2015; Saha and Rajkumar, 2018; Niranjan and Rajkumar, 2017). Our model is also a variation of the BTL model where the utility of each item is dependent on the items it is being compared to.

**Violations of Rational Choice** The social science community has long recognized and hypothesized about irrational choice (Shepard, 1964; Torgerson, 1965; Tversky, 1977; 1972; Bordalo et al., 2013). See (Rieskamp et al., 2006) for an excellent survey of this area including references to social science experiments that demonstrate scenarios where humans make choices that can violate a variety of rational choice axioms such as transitivity. There has been recent progress in modeling and providing evidence for violations of rational choice axioms in the machine learning community (Seshadri et al., 2019; Rosenfeld et al., 2020; Heckel et al., 2019; Pfannschmidt et al., 2019; Kleinberg et al., 2017; Shah and Wainwright, 2017; Ragain and Ugander, 2016; Niranjan and Rajkumar, 2017; Benson et al., 2016; Chen and Joachims, 2016b;a; Rajkumar et al., 2015; Yang and B. Wakin, 2015; Agresti, 2012). In contrast to our work, none of these works model preference data that both violates rational choice and admits a universal ranking of the items with the exception of (Shah and Wainwright, 2017; Heckel et al., 2019). Assuming there is a true ranking of the items, our model makes a direct connection between pairwise comparison data that violates rational choice and the underlying ranking. Violations of rational choice, including intransitivty, occur in our model because of contextual ef-

fects due to which pairs of items are being compared. These contextual effects distort the true ranking, whereas in the work of (Shah and Wainwright, 2017; Heckel et al., 2019) the intransitive choices define the ranking. Specifically, the items are ranked by sorting the items by the probability that an item beats any other item.

We now focus on the works most similar to ours. The work in (Seshadri et al., 2019), which generalizes (Chen and Joachims, 2016b;a) from pairwise comparisons to $k$-wise comparisons, considers a model for context dependent comparisons. However, because they do not assume access to features, their model cannot predict choices based on new items, which is a key task for very large modern data sets. In contrast, our model can predict pairwise outcomes and rankings of new items. Both (Rosenfeld et al., 2020) and (Pfannschmidt et al., 2019) assume access to features of items and propose learning contextual utilities with neural networks. In contrast, we propose a linear approach with typically far fewer parameters to estimate. Furthermore, the latter work does not contain any theory, whereas we prove a sample complexity result on estimating the parameters of our model. In all of the aforementioned works in this paragraph, the resulting optimization problems are non-convex with the exception of a special case in (Seshadri et al., 2019) that requires sampling every pairwise comparison. In contrast, the negative log likelihood of our model is convex. Interestingly, the work in (Makhijani and Ugander, 2019) shows that for a class of parametric models for pairwise preference probabilities, if intransitives exist, then the negative log likelihood cannot be convex. Our model does not belong to the class of parametric models they consider.

**Notation** For an integer $d > 0$, $[d] := \{1, \dots, d\}$. For $x, y \in \mathbb{R}^d$, $\langle x, y \rangle := \sum_{i=1}^d x_i y_i$. For $x \in \mathbb{R}^d$ and $\Omega \subset [d]$, let $x^\Omega \in \mathbb{R}^d$ where $(x^\Omega)_i = x_i$ if $i \in \Omega$ and 0 otherwise. For $i, j \in [n]$, "$i >_B j$" means "item $i$ beats item $j$." Let $\mathcal{P}(X)$ be the power set of a set $X$. Given a set of vectors $S = \{x_i \in \mathbb{R}^d\}_{i=1}^q$, $\text{span}(S) = \{\sum_{i=1}^q \alpha_i x_i : \alpha_i \in \mathbb{R}\}$.

## 2. Model and Algorithm

**Salient Feature Preference Model** Suppose there are $n$ items, and each item $j \in [n]$ has a known feature vector $U_j \in \mathbb{R}^d$. Let $U := \begin{bmatrix} U_1 U_2 \cdots U_n \end{bmatrix} \in \mathbb{R}^{d \times n}$. Let $w^* \in \mathbb{R}^d$ be the unknown *judgment weights*, which signify the importance of each feature when comparing items. Let $\tau : [n] \times [n] \to \mathcal{P}([d])$ be the known *selection function* that determines which features are used in each pairwise comparison. Let $P := \{(i, j) \in [n] \times [n] : i < j\}$ be the set of all pairs of items. Let $S_m = \{(i_\ell, j_\ell, y_\ell)\}_{\ell=1}^m$ be a set of $m$ independent pairwise comparison samples where $(i_\ell, j_\ell) \in P$ are chosen uniformly at random from $P$ with replacement, and $y_\ell \in \{0, 1\}$ indicates the outcome of the pairwise comparison where 1 indicates item $i_\ell$ beat item $j_\ell$ and 0 indicates item $j_\ell$ beat item $i_\ell$. We model $y_\ell \sim \text{Bern}(\mathbb{P}(i_\ell >_B j_\ell))$ where

$$\mathbb{P}(i_\ell >_B j_\ell) = \frac{\exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w^* \rangle\right)}{1 + \exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w^* \rangle\right)}. \tag{2}$$

To understand the probability model given by Equation (2), note that $\langle U_i^{\tau(i,j)}, w^* \rangle$ is the inner product of $U_i$ and $w^*$ after $U_i$ is projected to the coordinate subspace given by $\tau(i, j)$. Therefore, Equation (2) is simply the utility model of Equation (1) where the utilities are inner products computed in the subspace defined by the selection function $\tau$. If the selection function returns all the coordinates, i.e. $\tau(i, j) = [d]$, then Equation (2) becomes the standard BTL model where the utility of item $i$ is $\langle U_i, w^* \rangle$ and fixed regardless of context, i.e., regardless of which pair is being compared. This model is typically called "BTL with features," and we will refer to it as FBTL. See Section 6 in the Supplement for a natural extension of Equation (2) to $k$-wise comparisons for $k > 2$. Furthermore, we assume that the true ranking of all the items depends on all the features and is given by sorting the items by $\langle U_i, w^* \rangle$ for $i \in [n]$.

**Selection Function** We propose a selection function $\tau$ inspired by the social science literature, which posits that violations of rational choice axioms arise in certain scenarios because people make comparison judgments on a set of items based on the features that differentiate them the most (Rieskamp et al., 2006; Brown and Peterson, 2009; Bordalo et al., 2013).

For two variables $w, z \in \mathbb{R}$, let $\mu := (w + z)/2$ be their mean and $\bar{s} := ((w - \mu)^2 + (z - \mu)^2)/2$ be their sample variance. Given $t \in [d]$ and items $i, j \in [n]$, the *top-$t$ selection function* selects the $t$ coordinates with the $t$ largest sample variances in the entries of the feature vectors $U_i, U_j$.

**Algorithm: Maximum Likelihood Estimation** Given observations $S_m = \{(i_\ell, j_\ell, y_\ell)\}_{\ell=1}^m$, item features $U \in \mathbb{R}^{d \times n}$, and a selection function $\tau$, the negative log-likelihood of $w \in \mathbb{R}^d$ is

$$\mathcal{L}_m(w; U, S_m, \tau) = \sum_{\ell=1}^m \log\left(1 + \exp\left(u_{i_\ell, j_\ell}\right)\right) - y_\ell u_{i_\ell, j_\ell}, \tag{3}$$

where $u_{i_\ell, j_\ell} = \left\langle w, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \right\rangle$.

Equation 3 is equivalent to logistic regression with features $x_\ell = U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}$. See Section 7 of the Supplement

for the derivation. We estimate $w^*$ with the maximum likelihood estimator $\hat{w}$, which requires minimizing a convex function: $\hat{w} := \text{argmin}_w \mathcal{L}_m(w; U, S_m, \tau)$.

## 3. Theory

In this section, we analyze the sample complexity of estimating the judgment weights with the MLE given by minimizing $\mathcal{L}_m$ of Equation (3). We first consider the sample complexity under an arbitrary selection function, and then specialize to two concrete selection functions: one that selects all features per pair and another that selects just one feature per pair. Throughout this section, we assume the set-up and notation presented in the beginning of Section 2.

First, the following proposition completely characterizes the identifiability of $w^*$. Identifiability means that with infinite samples, it is possible to learn $w^*$. Precisely, the salient feature preference model is identifiable if for all $(i,j) \in P$ and for $w_1, w_2 \in \mathbb{R}^d$, if $\mathbb{P}(i >_B j; w_1) = \mathbb{P}(i >_B j; w_2)$, then $w_1 = w_2$ where $\mathbb{P}(i >_B j; w)$ refers to Equation (2) where $w$ is the judgement vector. The proof is in Section 8 of the Supplement.

**Proposition 1** (Identifiability)**.** *Given item features $U \in \mathbb{R}^{n \times d}$, the salient feature preference model with selection function $\tau$ is identifiable if and only if span$\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\} = \mathbb{R}^d$.*

Now we present our main theorem on the sample complexity of estimating $w^*$. Let

$$b^* := \max_{(i,j) \in P} |\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle|,$$

which is the maximum absolute difference between two items' utilities when comparing them in context, i.e. based on the features given by the selection function $\tau$. Let

$$\mathcal{W}(b^*) := \{w \in \mathbb{R}^d : \max_{(i,j) \in P} |\langle w, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle| \leq b^*\}.$$

We constrain the MLE to $\mathcal{W}(b^*)$ so that we can bound the entries of the Hessian of $\mathcal{L}_m$ in our theoretical analysis. We do not enforce this constraint in our synthetic experiments.

**Theorem 1** (Sample complexity of learning $w^*$)**.** *Let $U \in \mathbb{R}^{d \times n}$, $w^* \in \mathbb{R}^d$, $\tau$, and $S_m$ be defined as in the beginning of Section 2. Let $\hat{w}$ be the maximum likelihood estimator, i.e. the minimum of $\mathcal{L}_m$ in Equation (3), restricted to the set $\mathcal{W}(b^*)$. The following expectations are taken with respect to a uniformly chosen random pair of items from $P$. For $(i,j) \in P$, let*

$$\begin{aligned} Z_{(i,j)} &:= (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T \\ \lambda &:= \lambda_{\min}(\mathbb{E} Z_{(i,j)}), \\ \eta &:= \sigma_{\max}(\mathbb{E}((Z_{(i,j)} - \mathbb{E} Z_{(i,j)})^2)), \\ \zeta &:= \max_{(k,\ell) \in P} \lambda_{\max}(\mathbb{E} Z_{(i,j)} - Z_{(k,\ell)}), \end{aligned}$$

*where for a positive semidefinite matrix $X$, $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ are the smallest/largest eigenvalues of $X$, and where for any matrix $X$, $\sigma_{\max}(X)$ is the largest singular value of $X$. Let*

$$\beta := \max_{(i,j) \in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_\infty. \tag{4}$$

*Let $\delta > 0$. If $\lambda > 0$ and*

$$m \geq \max \left\{ C_1(\beta^2 d + \beta\sqrt{d}) \log(4d/\delta) \, , \; C_2(\eta + \lambda\zeta) \frac{\log(2d/\delta)}{\lambda^2} \right\},$$

*then with probability at least $1 - \delta$,*

$$\|w^* - \hat{w}\|_2 = O\left( \frac{\exp(b^*)}{\lambda} \sqrt{\frac{(\beta^2 d + \beta\sqrt{d}) \log(4d/\delta)}{m}} \right)$$

*where $C_1, C_2$ are constants given in the proof and the randomness is from the randomly chosen pairs and the outcomes of the pairwise comparisons.*

We utilize the proof technique of Theorem 4 in (Negahban et al., 2016), which proves a similar result for the standard BTL model of Equation (1), i.e. when $U = I_{n \times n}$, the $n \times n$ identity matrix, $d = n$, and $\tau(i,j) = [d]$ for all $(i,j) \in P$. We modify the proofs for arbitrary $U$ and $d$. See Section 10 in the Supplement for the proof.

We now discuss the terms that appear in Theorem 1. First, the $d \log(d/\delta)$ terms are natural since we are estimating $d$ parameters. Second, estimating $w^*$ well essentially requires inverting the logistic function. When $b^*$ is large, we need to invert the logistic function for pairwise probabilities that are close to 0 and 1. This is precisely the challenging regime, since a small change in probabilities results in a large change in the estimate of $w^*$, and thus we expect to require many samples to estimate $w^*$ when $b^*$ is large. The exponential dependence on $b^*$ is standard for this type of analysis and arises from the Hessian of $\mathcal{L}_m$. Third, $\eta$ and $\zeta$ arise from a matrix concentration bound applied to the Hessian of $\mathcal{L}_m$. Fourth, $\lambda$ arises from the minimum eigenvalue of the Hessian of $\mathcal{L}_m$ in a neighborhood of $w^*$, which controls the convexity of $\mathcal{L}_m$. This type of dependence also appears in other state of the art finite sample complexity analyses (Negahban et al., 2012). In addition, to better understand the role of $\lambda$, we present the following proposition whose proof is in Section 9 in the Supplement. Proposition 2 shows that the requirement $\lambda > 0$ in Theorem 1 is fundamental, because we would otherwise be unable to bound the estimation error for the non-identifiable part of $w^*$, i.e., the projection of $w^*$ onto the orthogonal complement of span$\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\} = \mathbb{R}^d$.

**Proposition 2.** *$\lambda > 0$ if and only if the salient feature preference model is identifiable.*

Finally, if one assumes $\lambda, \eta, \zeta, \beta, \exp(b^*)$ are $O(1)$, then $\Omega(d \log(d/\delta))$ samples are enough to guarantee the error is $O(1)$. However, as we will show in the corollaries, these parameters are not always $O(1)$, increasing the complexity. We point out that the combination of the features $U$ and the selection function $\tau$ is what dictates the parameters of Theorem 1. For the top-$t$ selection function in particular, we plot $\lambda, \zeta, \eta, b^*, \beta$, the number of samples required by Theorem 1, and the bound on the estimation error as a function of intransitivity rates in the Supplement in Section 13.1, to provide further insight into these parameters. Since we envision practical selection functions will be dependent on the features themselves, further analysis is a challenging but exciting subject of future work.

For deterministic $U$, we now specialize our results to FBTL as well as to the case where a single feature is used in each comparison. The following corollaries provide insight into how a particular selection function $\tau$ impacts $\lambda$, $\eta$, and $\zeta$ and thus the sample complexity.

First, we consider FBTL. In this case, the selection function selects all the features in each pairwise comparison, so there cannot be intransitivities in the preference data. The following Corollary of Theorem 1 gives a simplified form for $\lambda$ and upper bounds $\zeta$ and $\eta$. The terms involving the conditioning of $UU^T$ are natural; since we make no assumption on $w^*$, if the feature vectors are concentrated in a lower dimensional subspace, estimation of $w^*$ will be more difficult. See Section 11 of the Supplement for the proof.

**Corollary 1.1** (Sample complexity for FBTL)**.** *For the selection function $\tau$, suppose $|\tau(i,j)| = d$ for any $(i,j) \in P$. In other words, all the features are used in each pairwise comparison. Let $\nu := \max\{\max_{(i,j)\in P} \|U_i - U_j\|_2^2, 1\}$. Assume $n > d$. Without loss of generality, assume the columns of $U$ sum to zero: $\sum_{i=1}^n U_i = 0$. Let $\delta > 0$. Then,*

$$\lambda = \frac{n\lambda_{\min}(UU^T)}{\binom{n}{2}},$$
$$\zeta \le \nu + \frac{n\lambda_{\max}(UU^T)}{\binom{n}{2}}, \text{ and}$$
$$\eta \le \frac{\nu n \lambda_{\max}(UU^T)}{\binom{n}{2}} + \frac{n^2 \lambda_{\max}(UU^T)^2}{\binom{n}{2}^2}.$$

*Hence, if*

$$m \ge \max\Big\{C_1(\beta^2 d + \beta\sqrt{d})\log(4d/\delta), C_3 \log(2d/\delta)\nu n \bar{\lambda}\Big\}$$

*where*

$$\bar{\lambda} = \left(\frac{\lambda_{\max}(UU^T) + \lambda_{\max}(UU^T)^2 + \lambda_{\min}(UU^T)}{\lambda_{\min}(UU^T)^2}\right)$$

*then with probability at least $1 - \delta$,*

$$\|w^* - \hat{w}\|_2 = O\left(\frac{\exp(b^*)n}{\lambda_{\min}(UU^T)}\sqrt{\frac{(\beta^2 d + \beta\sqrt{d})\log(\frac{4d}{\delta})}{m}}\right)$$

*where $C_1$ and $C_3$ are constants given in the proof.*

To the best of our knowledge, this is the first analysis of the sample complexity for the MLE of FBTL parameters. There are related results in (Saha and Rajkumar, 2018; Negahban et al., 2012; Heckel et al., 2019; Shah and Wainwright, 2017) to which our bound compares favorably, and we discuss this in Section 11.2 of the Supplement.

Second, suppose the selection function is very aggressive and selects only one coordinate for each pair, i.e. $|\tau(i,j)| = 1$ for all $(i,j) \in P$. For instance, the top-1 selection function has this property. This type of selection function can cause intransitivities in the preference data as we show in the synthetic experiments of Section 4.1.

**Corollary 1.2.** *Assume that for any $(i,j) \in P$, $|\tau(i,j)| = 1$. Partition $P = \sqcup_{k=1}^d P_k$ into $d$ sets where $(i,j) \in P_k$ if $\tau(i,j) = \{k\}$ for $k \in [d]$. Let $\beta$ be defined as in Theorem 1 and*

$$\epsilon := \min_{(i,j)\in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_\infty.$$

*Let $\delta > 0$. Then*

$$\lambda \ge \frac{\epsilon^2}{\binom{n}{2}} \min_{k\in[d]} |P_k|,$$
$$\zeta \le \beta^2 + \frac{\beta^2}{\binom{n}{2}} \max_{k\in[d]} |P_k|, \text{ and}$$
$$\eta \le \frac{\beta^4}{\binom{n}{2}} \max_{k\in[d]} \left(|P_k| + \frac{|P_k|^2}{\binom{n}{2}}\right).$$

*Hence, if*

$$m \ge \max\Big\{C_1(\beta^2 d + \beta\sqrt{d})\log(4d/\delta), C_4(Q_1 + Q_2)\Big\},$$

*where*

$$Q_1 = \left(\frac{\beta^4}{\epsilon^4}\right)\frac{\binom{n}{2}\max_{k\in[d]}|P_k| + \max_{k\in[d]}|P_k|^2}{\min_{k\in[d]}|P_k|^2},$$
$$Q_2 = \left(\frac{\beta^2}{\epsilon^2}\right)\frac{\binom{n}{2} + \max_{k\in[d]}|P_k|}{\min_{k\in[d]}|P_k|},$$

*then with probability at least* $1-\delta$,

$$\|w^* - \hat{w}\|_2 = O\left(\frac{\exp(b^*)\binom{n}{2}}{\epsilon^2 \min_{k\in[d]}|P_k|}\sqrt{\frac{(\beta^2 d + \beta\sqrt{d})\log(\frac{4d}{\delta})}{m}}\right)$$

*where* $C_1$ *and* $C_4$ *are constants given in the proof.*

There are two main implications of Corollary 1.2 if we consider $\beta$ and $\epsilon$ constant. First, suppose there is a coordinate $k \in [d]$ such that $|P_k| := |\{(i,j) \in P : \tau(i,j) = k\}|$ is small. Intuitively it will take many samples to estimate $w^*$ well, since the chance of sampling a pairwise comparison that uses the $k$-th coordinate of $w^*$ is $|P_k|/\binom{n}{2}$. Corollary 1.2 formalizes this intuition. In particular, $\lambda = O(|P_k|/\binom{n}{2})$, and since $\lambda$ comes into the bounds of Theorem 1 in the denominator of both the lower bound on samples and the upper bound on error, a small $\lambda$ makes estimation more difficult.

Second, on the other hand, if $\epsilon$ is fixed, the maximum lower bound on $\lambda$ given by Corollary 1.2 is $\max\min_{i\in[d]}|P_i| = \binom{n}{2}/d$ where the maximum is with respect to any partition of $P$. In this case, $|P_i| \approx |P_j|$ for all $i,j \in [d]$, so the chance of sampling a pairwise comparison that uses any coordinate is approximately equal. Therefore, $\lambda, \eta, \zeta = O(1/d)$, and by tightening a bound used in the proof of Theorem 1, $\Omega(d^2\log(d/\delta))$ samples ensures the estimation error is $O(1)$. See Section 11.4 in the Supplement for an explanation.

Ultimately, we seek to estimate the underlying ranking of the items. The following corollary of Theorem 1 says that by controlling the estimation error of $w^*$, the underlying ranking can be estimated approximately. The sample complexity depends inversely on the square of the differences of full feature item utilities. Intuitively, if the absolute difference between the utilities of two items is small, then many samples are required in order to rank these items correctly relative to each other. See Section 12 in the Supplement for the proof.

**Corollary 1.3** (Sample complexity of estimating the ranking)**.** *Assume the set-up of Theorem 1. Pick* $k \in [\binom{n}{2}]$. *Let* $\alpha_k$ *be the* $k$*-th smallest number in* $\{|\langle w^*, U_i - U_j\rangle| : (i,j) \in P\}$. *Let* $M := \max_{i\in[n]}\|U_i\|_2$. *Let* $\gamma^* : [n] \to [n]$ *be the ranking obtained from* $w^*$ *by sorting the items by their full-feature utilities* $\langle w^*, U_i\rangle$ *where* $\gamma^*(i)$ *is the position of item* $i$ *in the ranking. Define* $\hat{\gamma}$ *similarly but for the estimated ranking obtained from the MLE estimate* $\hat{w}$. *Let* $\delta > 0$. *If*

$$m \geq \max\Big\{C_1(\beta^2 d + \beta\sqrt{d})\log(4d/\delta),\; C_2(\eta + \lambda\zeta)\frac{\log(2d/\delta)}{\lambda^2},\; \frac{C_5 M^2 e^{2b^*}(\beta^2 d + \beta\sqrt{d})\log(4d/\delta)}{\alpha_k^2\lambda^2}\Big\},$$

*then with probability* $1-\delta$,

$$K(\gamma^*, \hat{\gamma}) \leq k - 1,$$

*where* $K(\gamma^*,\hat{\gamma}) = |\{(i,j) \in P : (\gamma^*(i) - \gamma^*(j))(\hat{\gamma}(i) - \hat{\gamma}(j)) < 0\}|$ *is the Kendall tau distance between two rankings and* $C_1$, $C_2$, *and* $C_5$ *are constants given in the proof.*

## 4. Experiments

See Sections 14.1, 14.2, and 14.8 of the Supplement for additional details about the algorithm implementation, data, preprocessing, hyperparameter selection, and training and validation error for both synthetic and real data experiments.

### 4.1. Synthetic Data

We investigate violations of rational choice arising from the salient feature preference model and illustrate Theorem 1 while highlighting the differences between the salient feature preference model and the FBTL model throughout. Given the very reasonable simulation setup we use, these experiments suggest that the salient feature preference model may sometimes be better suited to real data than FBTL.

For these experiments, the ambient dimension $d = 10$, the number of items $n = 100$, and comparisons are sampled from the salient feature preference model with top-$t$ selection function. The coordinates of $U$, respectively $w^*$, are drawn from $\mathcal{N}(0, 1/\sqrt{d})$, respectively $\mathcal{N}(0, 4/\sqrt{d})$, so that $\mathbb{P}(i >_B j)$ is bounded away from 0 and 1 for $i,j \in [n]$. This set-up ensures $b^*$ does not become too large.

First, the salient feature preference model can produce preferences that systematically violate rational choice. In contrast, the FBTL model cannot. Let $P_{ij} = \mathbb{P}(i >_B j)$ and $T = \{(i,j,k) \in [n]^3 : P_{ij} > .5, P_{jk} > .5\}$. Then $(i,j,k) \in T$ satisfies strong stochastic transitivity if $P_{ik} \geq \max\{P_{ij}, P_{jk}\}$, moderate stochastic transitivity if $P_{ik} \geq \min\{P_{ij}, P_{jk}\}$, and weak stochastic transitivity if $P_{ik} \geq .5$ (Cattelan, 2012). We sample $U$ and $w^*$ 10 times as described in the beginning of the section and allow $t$ to vary in $[d]$. Figure 2 shows the average ratio of the number of weak, moderate, and strong stochastic transitivity violations to $|T|$ as a function of $t \in [d]$. There is very little deviation from the average. The standard error bars over the 10 experiments were plotted but they are so small that the

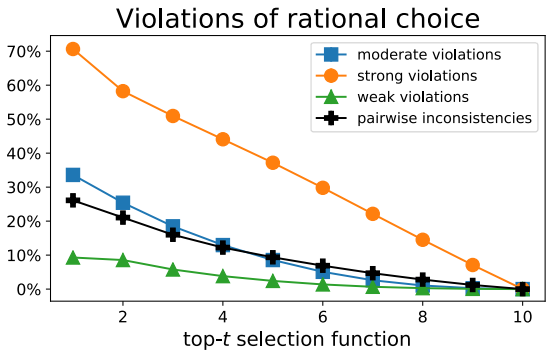


*Figure 2.* The salient feature preference model with the top-$t$ selection function produces systematic intransitives and pairwise comparisons that are inconsistent with the underlying ranking. When $t = 10$, the salient feature preference model with the top-$t$ selection function is the FBTL model, and hence there are no intransitives or pairwise inconsistencies.

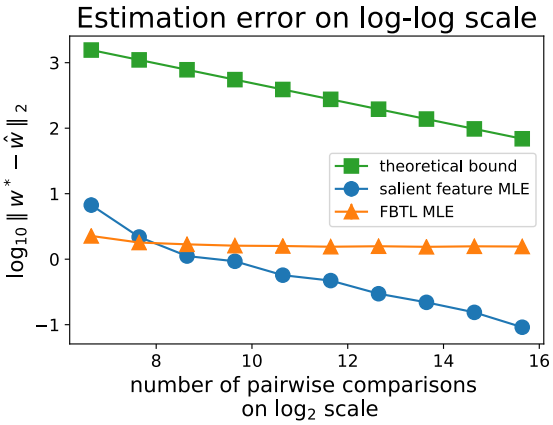


*Figure 3.* Illustration of Theorem 1 with the exact theoretical upper bound for the salient feature preference model with the top-1 selection function. Although there is a gap between the bound and the observed estimation error, they decrease at the same rate eventually. Excluding the first two points, the salient feature MLE error's slope on the log-log scale is -0.154, whereas the theoretical bound's slope is -0.151.

markers covered them. All $\binom{n}{2}$ probabilities given by Equation (2) are used to calculate the intransitivity rates. In the same figure we also show the percentage of pairwise comparisons that are inconsistent with the true ranking under the same experimental set-up. These are the pairs $i, j$ such that $\langle U_i - U_j, w^* \rangle < 0$, meaning item $i$ is ranked lower than item $j$ in the true ranking, but $\langle U_i^{\tau_t(i,j)} - U_j^{\tau_t(i,j)}, w^* \rangle > 0$ meaning item $i$ beats item $j$ by at least 50% when compared in isolation from the other items. Notice that when $t = 10$, the salient feature preference model is the FBTL model, so there are no pairwise inconsistencies or intransitives. Although this example is synthetic, real data exhibits intransitivity and even inconsistent pairs with the underlying ranking as discussed in the real data experiments in Section 4.2.

Second, we illustrate Theorem 1 with the top-1 selection function, and where $U$ and $w$ are sampled once as described in the beginning of this section. We sample $m$ pairwise comparisons for $m \in \{(100)2^{i-1} : i \in [10]\}$, fit the MLEs of both the salient preference model with the top-1 selection function and FBTL, and repeat 10 times. Figure 3 shows the average estimation error of $w^*$ on a logarithmic scale as a function of the number of pairwise comparison samples also on a logarithmic scale. Figure 3 also shows the exact theoretical upper bound where $\delta = \frac{1}{d} = \frac{1}{10}$ of Theorem 1 without constants $C_1$ and $C_2$ as stated in Section 10 of the Supplement. Again, there is very little deviation from the average. The standard error bars over the 10 experiments were plotted but they are so small that the markers covered them. There is a gap between the observed error and the theoretical bound, though the error decreases at the same rate. The error of the MLE of FBTL does not improve with more samples, since the pairwise comparisons are generated according to the salient feature preference model with the top-1 selection function. See Section 13.2 in the Supplement for investigating model misspecification, i.e. fitting the MLE of the top-$t$ selection function for $t \neq 1$ with the same experimental set-up.

By estimating $w^*$ well, we can estimate the underlying ranking well by Corollary 1.3. Under the same experimental set up, Figure 4 shows the Kendall tau correlation (definition given in Supplement 13.2) between the true ranking (obtained by ranking the items according to $\langle U_i, w^* \rangle$) and the estimated ranking (according to $\langle U_i, \hat{w} \rangle$) but on a new set of 100 items drawn from the same distribution. The maximum Kendall tau correlation between two rankings is 1 and occurs when both rankings are equal. Also, estimating $w^*$ well allows us to predict the outcome of unseen pairwise comparisons well, as shown in the Supplement in Section 13.2.

### 4.2. Real Data

For the following experiments, we use the top-$t$ selection function for the salient feature preference model, where $t$ is treated as a hyperparameter and tuned on a validation set. We compare to FBTL, RankNet (Burges et al., 2005) with one hidden layer, and Ranking SVM (Joachims, 2002). We append an $\ell_2$ penalty to $\mathcal{L}_m$ for the salient feature preference model and the FBTL model, that is, for regularization parameter $\mu$, we solve $\min_{w \in \mathbb{R}^d} \mathcal{L}_m(w) + \mu \|w\|_2^2$. For RankNet, we add to the objective function an $\ell_2$ penalty on the weights. As explained in more detail in subsections 14.6 and 14.11 in the Supplement, the hyperparameters for the salient feature

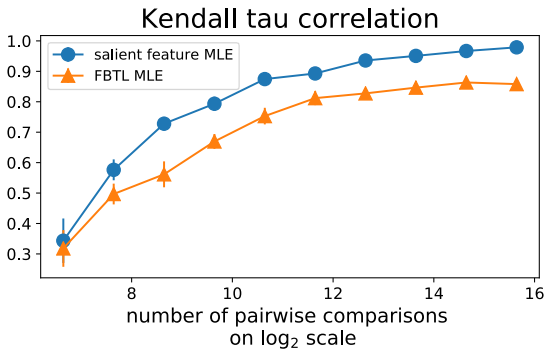


*Figure 4.* Kendall tau correlation between the true ranking and the estimated ranking where pairwise comparisons are sampled from the salient feature preference model with the top-1 selection function. Estimating $w^*$ well implies being able to estimate the underlying ranking well as stated in Corollary 1.3.

preference model are $t$ for the top-$t$ selection function and $\mu$, the hyperparameter for FBTL is $\mu$, the hyperparameter for Ranking SVM is the coefficient corresponding to the norm of the learned hyperplane, and the hyperparameters for RankNet are the number of nodes in the single hidden layer and the coefficient for the $\ell_2$ regularization of the weights.

**District Compactness** (Kaufman et al., 2017) collected preference data to understand human perception of compactness of legislative districts in the United States. Their data include both pairwise comparisons and $k$-wise ranking data for $k > 2$ as well as 27 continuous features for each district, including geometric features and compactness metrics. Although difficult to define precisely, the United States law suggests compactness is universally understood (Kaufman et al., 2017). In fact, the authors provide evidence that most people agree on a universal ranking, but they found the pairwise comparison data was extremely noisy. They hypothesize that pairwise comparisons may not directly capture the full ranking, since all features may not be used when comparing two districts in isolation from the other districts. Hence, this problem is applicable to our salient feature preference model and its motivation.

The goal as set forth by (Kaufman et al., 2017) is to learn a ranking of districts. We train on 5,150 pairwise comparisons collected from 94 unique pairs of districts to learn $\hat{w}$, an estimate of the judgment vector $w^*$, then estimate a ranking by sorting the districts by $\langle \hat{w}, U_i \rangle$. The $k$-wise ranking data sets are used for validation and testing. Since there is no ground truth for the universal ranking, we measure how close the estimated ranking is to each individual ranking. In this scenario, we care about the accuracy of the full ranking, and so we consider Kendall tau correlation. Given a $k$-wise comparison data set, Table 1 shows the average Kendall tau correlation between the estimated ranking and each individual ranking where the number in parenthesis is the standard deviation. The standard deviation on `shiny1` and `shiny2` is relatively high because the Kendall tau correlation between pairs of rankings in these data sets has high variability, shown in Figure 10 in the Supplement.

The MLE of the salient feature preference model under the top-$t$ selection function outperforms both the MLE of FBTL and Ranking SVM by a significant amount on 6 out 7 test sets, suggesting that pairwise comparison decisions may be better modeled by incorporating context. The MLE of the salient feature preference model, which is linear, is competitive with RankNet, which models pairwise comparisons as in Equation (1) except where the utility of each item uses a function $f$ defined by a neural network, i.e. $u_i = f(U_i)$.

The salient feature preference model may be outperforming FBTL and Ranking SVM since this data exhibits significant violations of rational choice. First, on the training set of pairwise comparisons, there are 48 triplets of districts $(i, j, k)$ where both (1) all three distinct pairwise comparisons were collected and (2) $P_{ij} > .5$ and $P_{jk} > .5$. Seventeen violate strong transitivity, 3 violate moderate transitivity, but none violate weak transitivity. Second, given a set of $k$-wise ranking data, let $\hat{P}_{ij}$ be the proportion of rankings in which item $i$ is ranked higher than item $j$. There are 20 pairs of districts that appear in both the $k$-wise ranking data and the pairwise comparison training data. Four of these pairs of items $i, j$ have the property that $(.5 - P_{ij})(.5 - \hat{P}_{ij}) < 0$, meaning item $i$ is typically ranked higher than item $j$ in the ranking data, but $j$ typically beats $i$ in the pairwise comparisons.

**`UT Zappos50k`** The `UT Zappos50K` data set consists of pairwise comparisons on images of shoes and 960 extracted vision features for each shoe (Yu and Grauman, 2014; 2017). Given images of two shoes and an attribute from {"open," "pointy," "sporty," "comfort"}, respondents picked which shoe exhibited the attribute more. The data consists of easier, coarse questions, i.e. based on comfort, pick between a slipper or high-heel, and harder, fine grained questions i.e. based on comfort, pick between two slippers.

We now consider predicting pairwise comparisons instead of estimating a ranking since there is no ranking data available. We train four models, one for each attribute. See Table 2 for the average pairwise comparison accuracy over ten train (70%), validation (15%), and test splits (15%) of the data. The pairwise comparison accuracy is defined as the percentage of items $(i, j)$ where $i$ beats $j$ a majority of the time and the model estimates the probability that $i$ beats $j$ exceeds 50%.

In this case, the MLE of the FBTL model and the salient feature preference model under the top $t$ selection function perform similarly. Nevertheless, while the FBTL model

*Table 1.* Average Kendall tau correlation over individual rankings on test sets for district compactness. The number in parenthesis is the standard deviation.

| Model: | Shiny1 | Shiny2 | UG1-j1 | UG1-j2 | UG1-j3 | UG1-j4 | UG1-j5 |
|---|---|---|---|---|---|---|---|
| Salient features | **0.14** (.26) | **0.26** (.2) | **0.48** (.21) | **0.41** (.09) | **0.6** (.1) | 0.14 (.14) | **0.42** (.09) |
| FBTL | 0.09 (.22) | 0.18 (.17) | 0.2 (.12) | 0.26 (.07) | 0.45 (.15) | 0.2 (.13) | 0.06 (.14) |
| Ranking SVM | 0.09 (.22) | 0.18 (.17) | 0.22 (.12) | 0.26 (.07) | 0.45 (.15) | 0.2 (.13) | 0.06 (.14) |
| RankNet | 0.12 (.24) | 0.24 (.18) | 0.28 (.14) | 0.37 (.08) | 0.53 (.11) | **0.28** (.08) | 0.15 (.15) |

*Table 2.* Average pairwise prediction accuracy over 10 train/validation/test splits on the test sets by attribute for UT Zappos50k. $C$ stands for coarse and $F$ stands for fine grained. The number in parenthesis is the standard deviation.

| Model: | open-$C$ | pointy-$C$ | sporty-$C$ | comfort-$C$ | open-$F$ | pointy-$F$ | sporty-$F$ | comfort-$F$ |
|---|---|---|---|---|---|---|---|---|
| Salient features | 0.73 (.02) | 0.78 (.02) | 0.78 (.03) | 0.77 (.03) | 0.6 (.04) | 0.59 (.04) | 0.59 (.03) | 0.56 (.03) |
| FBTL | 0.73 (.02) | 0.77 (.03) | 0.8 (.03) | 0.78 (.03) | 0.6 (.03) | 0.6 (.03) | 0.59 (.03) | 0.58 (.05) |
| Ranking SVM | 0.74 (.02) | 0.78 (.03) | 0.79 (.03) | 0.78 (.03) | 0.6 (.03) | 0.6 (.04) | 0.6 (.04) | 0.58 (.03) |
| RankNet | 0.73 (.01) | 0.79 (.01) | 0.78 (.03) | 0.8 (.02) | 0.61 (.02) | 0.59 (.02) | 0.59 (.04) | 0.59 (.05) |

utilizes all 990 features, the best $t$'s on each validation set and split of the data do not use all features, so our model is different from yet competitive to FBTL. See Table 3 in the Supplement. This suggests that the salient feature preference model under the top-$t$ selection function for relatively small $t$ is still a reasonable model for real data.

## 5. Conclusion

We focused on the problem of learning a ranking from pairwise comparison data with irrational choice behaviors, and we formulated the salient feature preference model where one uses projections onto salient coordinates in order to perform comparisons. We proved sample complexity results for MLE on this model and demonstrated the efficacy of our model on both synthetic and real data. Going forward, we would like to develop techniques to learn both the selection function $\tau$ and feature embeddings simultaneously. Finally, it will be useful to consider how to incorporate context into models more sophisticated than BTL, and also consider contextual effects in other tasks that use human judgements such as ordinal embedding (Terada and Luxburg, 2014).

**Acknowledgements** L. Balzano was supported by NSF CAREER award CCF-1845076, NSF BIGDATA award IIS-1838179, ARO YIP award W911NF1910027, and the Institute for Advanced Study Charles Simonyi Endowment. A. Bower was also supported by ARO W911NF1910027 as well as University of Michigan's Rackham Merit Fellowship, University of Michigan's Mcubed grant, and NSF graduate research fellowship DGE 1256260.

## References

Alan Agresti. *Categorical data analysis.* John Wiley & Sons, 2012.

Austin R Benson, Ravi Kumar, and Andrew Tomkins. On the relevance of irrelevant alternatives. In *Proceedings of the 25th International Conference on World Wide Web*, pages 963–973. International World Wide Web Conferences Steering Committee, 2016.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843, 2013.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Thomas C Brown and George L Peterson. An enquiry into the method of paired comparison: reliability, scaling, and thurstone's law of comparative judgment. *Gen Tech. Rep. RMRS-GTR-216WWW. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station. 98 p.*, 216, 2009.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

Manuela Cattelan. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, pages 412–433, 2012.

Shuo Chen and Thorsten Joachims. Modeling intransitivity in matchup and comparison data. In *Proceedings of the*

*Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 227–236, New York, NY, USA, 2016a. ACM.

Shuo Chen and Thorsten Joachims. Predicting matchups and preferences in context. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 775–784. ACM, 2016b.

Reinhard Heckel, Nihar B Shah, Kannan Ramchandran, Martin J Wainwright, et al. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 47(6):3099–3126, 2019.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

Aaron Kaufman, Gary King, and Mayya Komisarchik. How to measure legislative district compactness if you only know it when you see it. *American Journal of Political Science*, 2017.

Jon Kleinberg, Sendhil Mullainathan, and Johan Ugander. Comparison-based choices. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 127–144. ACM, 2017.

Yao Li, Minhao Cheng, Kevin Fujii, Fushing Hsieh, and Cho-Jui Hsieh. Learning from group comparisons: Exploiting higher order interactions. In *Advances in Neural Information Processing Systems 31*, pages 4981–4990. Curran Associates, Inc., 2018.

Yu Lu and Sahand N Negahban. Individualized rank aggregation using nuclear norm regularization. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1473–1479. IEEE, 2015.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 1959.

Rahul Makhijani and Johan Ugander. Parametric models for intransitivity in pairwise rankings. In *The World Wide Web Conference*, pages 3056–3062, 2019.

Joshua E Menke and Tony R Martinez. A bradley–terry artificial neural network model for individual ratings in group competitions. *Neural computing and Applications*, 17(2):175–186, 2008.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2016.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

UN Niranjan and Arun Rajkumar. Inductive pairwise ranking: going beyond the n log (n) barrier. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Sewoong Oh, Kiran K Thekumparampil, and Jiaming Xu. Collaboratively learning preferences from ordinal data. In *Advances in Neural Information Processing Systems*, pages 1909–1917, 2015.

Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 1907–1916, 2015.

Karlson Pfannschmidt, Pritha Gupta, and Eyke Hüllermeier. Learning choice functions. *preprint*, abs/1901.10860, 2019. URL http://arxiv.org/abs/1901.10860.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

Stephen Ragain and Johan Ugander. Pairwise choice markov chains. In *Advances in Neural Information Processing Systems*, pages 3198–3206, 2016.

Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 118–126, Bejing, China, 22–24 Jun 2014. PMLR.

Arun Rajkumar, Suprovat Ghoshal, Lek-Heng Lim, and Shivani Agarwal. Ranking from stochastic pairwise preferences: Recovering condorcet winners and tournament solution sets at the top. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 665–673. JMLR.org, 2015.

Jörg Rieskamp, Jerome R Busemeyer, and Barbara A Mellers. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, 44(3):631–661, 2006.

Nir Rosenfeld, Kojin Oshiba, and Yaron Singer. Predicting choice with set-dependent aggregation. In *Proceedings of the 37th International Conference on Machine learning*, 2020.

Aadirupa Saha and Arun Rajkumar. Ranking with features: Algorithm and a graph theoretic analysis. In *preprint*, 2018. URL https://arxiv.org/pdf/1808.03857.pdf.

Arjun Seshadri, Alexander Peysakhovich, and Johan Ugander. Discovering context effects from raw choice data. *International Conference on Machine Learning*, 2019.

Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

Roger N Shepard. Attention and the metric structure of the stimulus space. *Journal of mathematical psychology*, 1 (1):54–87, 1964.

Yoshikazu Terada and Ulrike Luxburg. Local ordinal embedding. In *International Conference on Machine Learning*, pages 847–855, 2014.

Warren S Torgerson. Multidimensional scaling of similarity. *Psychometrika*, 30(4):379–393, 1965.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Amos Tversky. Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281, 1972.

Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Amos Tversky and Itamar Simonson. Context-dependent preferences. *Management science*, 39(10):1179–1189, 1993.

Dehui Yang and Michael B. Wakin. Modeling and recovering non-transitive pairwise comparison matrices. *2015 International Conference on Sampling Theory and Applications, SampTA 2015*, pages 39–43, 07 2015.

A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition (CVPR)*, Jun 2014.

A. Yu and K. Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *International Conference on Computer Vision (ICCV)*, Oct 2017.