Parametric Models for Intransitivity in Pairwise Rankings

Rahul Makhijani
Stanford University
rahulmj@stanford.edu

Johan Ugander
Stanford University
jugander@stanford.edu

ABSTRACT

There is a growing need for discrete choice models that account for the complex nature of human choices, escaping traditional behavioral assumptions such as the transitivity of pairwise preferences. Recently, several parametric models of intransitive comparisons have been proposed, but in all cases the model log-likelihood is non-concave, making inference difficult. In this work we generalize this trend, showing that there cannot exist an parametric model that both (i) has a log-likelihood function that is concave in item-level parameters and (ii) can exhibit intransitive preferences. Given this observation, we also contribute a new simple model for analyzing intransitivity in pairwise comparisons, taking inspiration from the Condorcet method (majority vote) in social choice theory. The *majority vote model* we analyze is defined as a voting process over independent Random Utility Models (RUMs). We infer a multidimensional embedding of each object or player, in contrast to the traditional one-dimensional embedding used by models such as the Thurstone or Bradley-Terry-Luce (BTL) models. We show that a three-dimensional majority vote model is capable of modeling arbitrarily strong and long intransitive cycles, and can also represent arbitrary pairwise comparison probabilities on any triplet. We provide experimental results that substantiate our claims regarding the effectiveness of our model in capturing intransitivity for various pairwise choice tasks such as predicting choices in recommendation systems, winners in online video games, and elections.

ACM Reference Format:
Rahul Makhijani and Johan Ugander. 2019. Parametric Models for Intransitivity in Pairwise Rankings. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313717

1 INTRODUCTION

Modeling discrete choice is a key challenge at the heart of many ranking problems, with applications to recommendation systems [8], information retrieval [12], and predicting the outcomes of diverse pairwise competitions [13]. The "choice" being modeled is sometimes a choice made by an individual over alternative options, or sometimes a matter of a competition or game "choosing" a player to win that competition. Prominent examples of ranking systems based on discrete choice models are the the Elo system for Chess ranking and the TrueSkill system [9] for matching players in online games. Meanwhile pairwise comparisons are used in recommendation systems to infer global quality measures from pairwise preferences of users [20], an idea that builds on Thurstone's original observation from the 1920's that comparative judgments are often easier than absolute judgments in many contexts [21].

Two of the most widely used models of ranking from pairwise comparisons are the Thurstone [21] and the Bradley-Terry-Luce (BTL) [3] models. These two models both belong to a broader class of Random Utility Models (RUMs) [1, 14], which assume that each alternative i is completely described by a random utility U_i with a innate quality λ_i, also known as the nominal utility, and a zero-mean noise term ϵ_i such that $U_i = \lambda_i + \epsilon_i$. For a random utility model with n items and utilities $U_1, \ldots, U_n$, it is assumed that the probability that an item is chosen can be written as:

$$P(a \text{ chosen from } S) = Pr(U_a \geq U_c, \forall c \in S),$$

for all possible alternatives $a \in S$ and choice sets S. For pairwise choices, this structure reduces to assuming that choices

$$P_{AB} := \Pr[A \text{ chosen from } \{A, B\}] = \Pr[U_A - U_B > 0].$$

The Thurstone model assumes that each ϵ_i is i.i.d. Normal while the BTL model assumes that each ϵ_i is i.i.d. Gumbel (sometimes also called double-exponential). A model is an *independent* RUM if the random variables ϵ_i are all independent. Thurstone's model and the BTL model are therefore both independent RUMs. For the Thurstone and BTL models, the difference of two utilities $U_A - U_B$ obeys a Normal distribution and Logistic distribution, respectively.

An important general feature of RUMs with i.i.d. noise is that they implicitly assume stochastic transitivity. Informally, if A beats B more often than not and B beats C more often than not, then any such model will assume that A beats C more often than not. The stochastic transitivity of these RUMs follows very specifically from the requirement that the noise be both independent *and* identically distributed. A famous examples of stochastic *intransitivity* is that of *non-transitive dice* [16], which can be formulated as a RUM with independent but non-identical noise distributions. Thus, independent RUMs need not exhibit stochastically transitive choices.

Non-transitive dice raise an intriguing question, namely whether it is possible to efficiently learn the parameters of an independent RUM that models intransitive pairwise comparisons using distributions defined by a small set of parameters (such the moments of the noise distributions). Several models of intransitive comparisons have recently been proposed, notably the 2-dimensional BTL model [4] and the Blade-Chest model [6]. But these and other models suffer from non-concave log-likelihood functions, meaning that maximum likelihood estimation is challenging. In this work we therefore formally ask: is there some overlooked model with a tractable concave log-likelihood that can support intransitive pairwise preferences? Our main result answers this question with a general "no."

Given this result, we investigate interpretable models of intransitivity, drawing inspiration from social choice theory. Under the

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313717

Condorcet method of voting, individual rankings may be transitive but overall the collective social decision procedures are not. While the log-likelihood of this model is still non-concave, we find that standard routines for non-convex optimization work well. Furthermore, the relatively well-understood structure of majority vote allows us to derive a collection of expressivity results regarding what sorts of intransitivity this model can and can't represent.

2 MODELLING INTRANSITIVITY

There are three basic mathematical definitions of stochastic transitivity - weak, strong, and antitransitivity. *Weak stochastic intransitivity* is defined as the existence of three alternatives A, B, C in a set of alternatives such that:

$$P_{AB} > 0.5, \;\; P_{BC} > 0.5, \;\; P_{CA} > 0.5,$$

and we are specifically focused on weak stochastic intransitivity in this work. As we will now show, any RUM with i.i.d. uncertainties, including the Thurstone and BTL models, exhibit weak stochastic transitivity.

Proposition 2.1. *For any Random Utility Model with random utilities $U_i = \lambda_i + \epsilon_i$ with i.i.d. ϵ_i and $\lambda_i \neq \lambda_j, \forall i, j$, weak transitivity holds.*

Proof. For any three alternatives A, B, and C, we have that ϵ_A, ϵ_B, and ϵ_C are i.i.d., implying that $\epsilon_A - \epsilon_B$, $\epsilon_B - \epsilon_C$, and $\epsilon_A - \epsilon_C$ are all identically distributed and also symmetric. We see then that $\Pr[U_i - U_j] > 0.5$ iff $\lambda_i > \lambda_j$, and thus the three alternatives are totally ordered by their means. □

There are three traditional approaches to introducing pairwise intransitivity into choice models: RUMs with possibly dependent distributions, RUMs with non-identical distributions, or departures from the RUM framework. Our work here falls in the last category. The study of dependent RUMs has led to an expansive modeling literature, but inference in these settings is generally quite difficult [23]. For intransitivity from independent but non-identical RUMs, we refer the reader to the literature on non-transitive dice [16], which is bounded in the intransitivity it can achieve, a matter that we elaborate on in Section 5.1. It is also possible to look at non-parametric models where each pairwise probability is its own "parameter" p_{ij}, which can be estimated using maximum likelihood [10]. Chatterjee [5] introduced the *non-parametric Bradley-Terry model* with a strong stochastic transitivity assumption that $p_{ik} \geq p_{jk}$ for $k \neq i, j$ whenever i beats j. This assumption facilitates an efficient consistent estimator. Without this transitivity assumption, a simple empirical estimator would be consistent but not very efficient. This motivates the need for parametric models where the p_{ij} can be expressed as a function of item-level parameters for i and j that may be efficiently estimated.

2.1 Prior work on multidimensional models

The use of higher dimensional scores to model pairwise intransitivity is an idea with a notable history [4, 15] and also recent advancements [6, 7, 17]. A two-dimensional extension of BTL is due to Causeur and Husson [4], parameterized by two-dimensional scores $\lambda_i \in \mathbb{R}^2$, $i = 1, \ldots, n$, as well as a binary matrix $\Sigma \in \{-1, +1\}^{n \times n}$ of pairwise dominances. During estimation, each $\sigma_{ij} \in \Sigma$ is determined from the data, taking the value $+1$ if player i wins the majority of the matches between player i and j, and -1 otherwise. From these parameters, $P_{ij} = S(\sigma_{ij} || \lambda_i - \lambda_j ||_2)$, where $S(x) = 1/(1 + \exp(-x))$. Although the Casseur–Husson model can exhibit intransitivity, it is highly sensitive to changes in the empirical data since the dominance parameters Σ depend directly on the pairwise counts, and very little has been established regarding the transitive vs. intransitive properties of this model.

The Casseur–Husson model can be easily extended to obtain a d-dimensional BTL model where the score of each player $\lambda_i \in \mathbb{R}^d$. Note that such extensions require more constraints to impose uniqueness. As an example, for a 3-dimensional BTL of n players, the following constraints are sufficient for identifiability: $\sum_{i=1}^{3} \lambda_i = 0$ and $\lambda_i \cdot \lambda_j = 0 \; \forall i, j \; \in \{1, 2, 3\}, i \neq j$. The number of parameters for this *3D BTL model* are $3n - 6$ for λ (minus the six constraints) and $\binom{n}{2}$ for Σ. It is straight-forward to verify that the negative log-likelihood function of these extended BTL models are non-convex in general.

A recent model similarly based on a higher-dimensional embedding of alternatives is the blade-chest (BC) model [6, 7]. In this model, every player a is represented by two d-dimensional vectors, a "blade" vector $\mathbf{a}_{\text{blade}}$ and a "chest" vector $\mathbf{a}_{\text{chest}}$. There are two variations of the model, "BC distance" and "BC inner." In both models the probability of player a beating player b in the match is given by $S(Q(a, b))$ where $S(x) = 1/(1 + \exp(-x))$ is again the logistic function and $Q(a, b)$ is a choice of a so-called *match-up function* between players a and b. The match-up functions for the distance and inner variations of the Blade-Chest model are:

$$Q_{\text{dist}}(a, b) = |\mathbf{b}_{\text{blade}} - \mathbf{a}_{\text{chest}}|_2^2 - |\mathbf{b}_{\text{chest}} - \mathbf{a}_{\text{blade}}|_2^2 + \gamma_a - \gamma_b,$$
$$Q_{\text{inner}}(a, b) = \mathbf{b}_{\text{blade}} \cdot \mathbf{a}_{\text{chest}} \; - \mathbf{b}_{\text{chest}} \cdot \mathbf{a}_{\text{blade}} + \gamma_a - \gamma_b.$$

Here the γ parameters are additional parameters that assure that the BTL model is captured as a clean special case (when all blade and chest vectors are zero vectors). There are $2(d + 1)n$ parameters in such Blade-Chest models. BC models perform well when the dimension is large, specifically $O(n)$, leading to a total of $O(n^2)$ parameters. Performance improvements over BTL are modest when only a few dimensions are employed. High-dimensional BC models are computationally difficult to learn due to the necessary minimization of a non-convex non-separable negative log-likelihood function. These challenges of optimizing and interpreting high-dimensional models motivate the development of our low-dimensional model.

3 NON-CONCAVITY OF INTRANSITIVITY

We now give our main theoretical result, that intransitivity necessarily implies non-concavity for the log-likelihood of any item-parametric choice model. The extended BTL model and both Blade-Chest models are merely different choices of nonlinear pairwise probability functions that fall into this more general analysis. We begin with a useful definition.

Definition 3.1. A *pairwise probability function* $f : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ is a function mapping the d-dimensional parametric representations of two alternatives to a choice probability such that $f(i, j) + f(j, i) = 1, \forall i, j$.

For the BTL and Thurstone models, $f(a,b) = S(a-b)$ and $f(a,b) = \Phi(a-b)$, respectively, for scalars a and b and where S is the logit function and Φ is the cumulative distribution function of the standard Normal. For the extended BTL model we have that $f(i,j) = S(\sigma_{ij}||\lambda_i - \lambda_j||_2)$. For the Blade-Chest model we have e.g. $f_{\text{BC-dist}}(i,j) = S(Q_{\text{dist}}(i,j))$. For the Blade-Chest models we note that the dimension d of the score representation is different from the dimension D of the embedding: $d = 2(D+1)n$.

While high-dimensional BTL and Blade-Chest models are capable of exhibiting intransitivity, their negative log-likelihood functions are in general not convex in the parameters of the alternatives, which typically makes global minimization challenging. This difficulty raises a natural question: are there models of intransitive pairwise choice where the log-likelihood function is concave? Notice that for a generic pairwise probability function $f(i,j)$, the log-likelihood objective of the model given data becomes:

$$\ell(\lambda; \{n_{ij}\}) = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} n_{ij} \log(f(i,j)),$$

where n_{ij} denotes the number of times i is chosen over j.

In order to furnish a concave log-likehood function, we therefore seek pairwise probability functions $f(a,b)$ that are log-concave in the score parameters. Can any such f exhibit intransitivity? We obtain the following negative result.

Theorem 3.2. *For any pairwise probability function $f(\lambda_i, \lambda_j)$ with $\lambda_i, \lambda_j \in \mathbb{R}^d$, f can exhibit an intransitive cycle only if f is not log-concave.*

Proof. Let A, B and C be three objects with score parameters $\lambda_i \in \mathbb{R}^d$ and let them form an intransitive cycle, meaning that

$$\begin{aligned} P_{AB} &= f(\lambda_A, \lambda_B) > 0.5, \\ P_{BC} &= f(\lambda_B, \lambda_C) > 0.5, \\ P_{CA} &= f(\lambda_C, \lambda_A) > 0.5. \end{aligned} \tag{1}$$

Further assume that f is log-concave. We will show that this implies a contradiction.

Notice first that the function f is bounded, $0 \leq f(i,j) \leq 1$, $\forall i,j$. As f is log-concave we also have the following two log-concavity properties for pairs $f(p,q)$ and $f(x,y)$ and $0 \leq \alpha \leq 1$:

$$f(\alpha p + (1-\alpha)x, \alpha q + (1-\alpha)y) \geq \min\{f(p,q), f(x,y)\},$$

$$f(\alpha p + (1-\alpha)x, \alpha q + (1-\alpha)y) \geq f(p,q)^{\alpha} f(x,y)^{1-\alpha}.$$

Using this second log-concavity property for $\alpha = 1/2$ we have

$$f((\lambda_A + \lambda_B)/2, (\lambda_B + \lambda_C)/2) \geq \sqrt{f(\lambda_A, \lambda_B) f(\lambda_B, \lambda_C)}.$$

Further employing $f(\lambda_A, \lambda_B) > 0.5$ and $f(\lambda_B, \lambda_C) > 0.5$, from Equation (1) and the log-concavity, we have

$$f((\lambda_A + \lambda_B)/2, (\lambda_B + \lambda_C)/2) > 0.5. \tag{2}$$

Clearly $f(-\frac{\lambda_B}{2}, -\frac{\lambda_B}{2}) = 0.5$ because it is a pairwise probability function. By convexity between $f(-\frac{\lambda_B}{2}, -\frac{\lambda_B}{2})$ and Equation (2) with $\alpha = 1/2$, we obtain $f(\frac{\lambda_A}{4}, \frac{\lambda_C}{4}) > 0.5$ and hence $f(\frac{\lambda_C}{4}, \frac{\lambda_A}{4}) < 0.5$.

To obtain the contradiction we simply employ log-concavity for the pairs $\{\lambda_C, \lambda_A\}$ and $\{0, 0\}$ to obtain $f(\frac{\lambda_C}{4}, \frac{\lambda_A}{4}) > 0.5$, contradicting the earlier conclusion. □

Among the many possible choices of intransitive pairwise probability functions f, the very general family of functions that define a model, we might have suspected that there were some function with favorable inferential properties, i.e. one that yielded an overall log-likelihood function that was concave. The above result tells us that no such function exists, at least for item-level parameterizations. We note that this result covers independent RUMs (non-transitive dice) where one might think it possible to achieve intransitivity using multiple "score parameters" to separately represent the location, variance, and/or skewness of the random utilities. Such a model does have the capacity to model intransitive choice, but the above result applies even to such models, meaning they can not have a concave log-likelihood function.

In the absence of a convexity-based optimization argument for one pairwise probability function over another—motivating one choice of model over another—we now turn to a particularly intuitive and low-dimensional pairwise probability function based on the majority vote model from social choice theory.

4 MAJORITY VOTE MODEL

In this section we present and investigate a model of pairwise comparisons based on the basic concept of majority vote. In the majority vote model, each object (player) has an attribute representation $\mu \in \mathbb{R}^k$, $k \geq 1$ odd, and the pairwise probabilities of choosing an object (player) is modeled as a simple dimension-wise majority vote function over the different attribute dimensions.

Consider a contest where there are n players competing and each player i has a vector of attributes μ_i. A game between players i and j is won by player i if $\mu_i - \mu_j + \xi$ is positive in $(k+1)/2$ or more coordinates, and player j otherwise. Here ξ is a noise vector defined in terms of a difference distribution, a definition due to Yellott [25].

Definition 4.1. If f and F are the probability density function (p.d.f.) and cumulative distribution function (c.d.f.) of two i.i.d. random variables X_1 and X_2, then the *difference density function* and *difference distribution function* d_f and D_f are the p.d.f. and c.d.f. of $X_1 - X_2$ respectively.

The noise vector ξ in the majority vote model is drawn independently for each match-up from a multivariate distribution G with independence between the dimensions such that $G = \prod_{i=1}^{k} g_i$, where g_i is a distribution function such that the difference distribution function of two g_i's is continuous and strictly increasing over $\mathbb{R}$. Two useful examples of g_i are the Normal distribution and the Gumbel distribution, leading to d_{g_i} being Normal and Logistic, respectively. We refer to the model with Gaussian noise as the Gaussian majority vote model. The Gaussian majority vote model reduces to the Thurstone model if the "voting" is conducting using only one dimension.

Although this majority vote model can be extended to arbitrarily large dimensions, we focus our analysis on the simple case of $k = 3$ dimensions and furnish a complete derivation of the maximum likelihood estimate of the parameter representations in this case. This choice of k limits the number of parameters to $3n$, where n is the number of alternatives (players). We show in Section 5 of this work that three dimensions is sufficient to (1) model strong and long intransitive cycles and (2) completely express the pairwise choice probabilities of any triplet. Based on these arguments we argue that three dimensions are sufficient for many applications. That said, it is straight-forward to generalize the MLE derivation below to arbitrarily large dimensions.

4.1 MLE derivation for three dimensions

For the three-dimensional model we let $M \in \mathbb{R}^{3n}$ denote the matrix of attributes μ of all the n players. The i^{th} row of matrix M is then $\mu_i = (\mu_i^1, \mu_i^2, \mu_i^3)$, the attributes of the i^{th} player. Let $p_1(i,j), p_2(i,j)$, and $p_3(i,j)$ be the probabilities that $\mu_i - \mu_j + \xi$ are positive in the first, second, and third dimensions, respectively. We write $p_k(i,j)$ as p_k for notational simplicity. Hence

$$p_k = D_g(\mu_i^k - \mu_j^k), \;\; \forall k = 1, 2, 3.$$

For additional notational simplicity we define the win probability w_{ij} to be the pairwise probability function f evaluated for two players i and j, were $x \geq y$ indicates that x is greater than y in at least two coordinates:

$$\begin{aligned} w_{ij} \quad &= \quad \Pr[\mu_i - \mu_j + \xi \geq 0] \\ &= \quad p_1p_2p_3 + (1 - p_1)p_2p_3 + p_1(1 - p_2)p_3 + p_1p_2(1 - p_3). \end{aligned}$$

The log-likelihood $\ell(M; N)$ is, dropping additive constants and focusing on only pairs $i < j$:

$$\ell(M; N) = \sum_{i,j:i<j} n_{ij} \log(w_{ij}) + n_{ji} \log(w_{ji}).$$

Here N denotes the pairwise comparison matrix and n_{ij} be the number of times i has been chosen over j. The gradients for the majority vote model can be calculated easily for gradient-based optimization methods [2]. As an example, for the Gaussian majority vote model we obtain the following gradient for parameter μ_{i1}:

$$\frac{\partial \ell}{\partial \mu_{i1}} = \sum_{i<j} \left[\left(\frac{n_{ij}}{w_{ij}} - \frac{n_{ji}}{1 - w_{ij}} \right) \phi(\mu_{i1} - \mu_{j1}) \Big(p_2 + p_3 - 2p_2p_3 \Big) \right].$$

5 CHARACTERISTICS OF THE MAJORITY VOTE MODEL

A possibly desirable property of a model capturing intransitivity is the ability to represent all pairwise win probabilities between players, but the ability to represent any $n \times n$ matrix of pairwise probabilities would clearly require $\binom{n}{2}$ parameters. In this work we are interested in models with few parameters, at most $O(n)$ and specifically $3n$ for our 3D majority vote model, and seek focused results that establish the expressiveness of the majority vote model despite only having $3n$ parameters. Specifically, we show that it is capable of representing arbitrarily long intransitive cycles and that it is capable of representing any submatrix of probabilities between a triplet of players (though this is not the same as simultaneously being able to represent any submatrix on all triplets).

5.1 Cycle expressivity

The random utility model (RUM) framework is an important framework for the characterization of various ranking models [1, 14]. Random Utility Models in general do not necessarily assume stochastic transitivity, not even for independent RUMs: consider the example of non-transitive Efron dice [16] discussed in the introduction. There are, however, limits on how "strong" the intransitivity for independent RUMs can be: Trybula provides conditions for existence of independent random variables $X_1, X_2, ..., X_n$ such that $\min\{P(X_1 > X_2), P(X_2 > X_3), ..., P(X_n > X_1)\}$ is bounded by a constant $c_n < 1$ for all n [24]. The sequence c_n is increasing and converges to 3/4. It is easy to determine that $c_3 = (\sqrt{5} - 1)/2$ and $c_4 = 2/3$. In this subsection we show that majority vote models exhibit no such bound on the strength of intransitivity.

Our majority vote model can be understood as an aggregation of independent RUM models. There is no statistical dependence between the random variables used to represent players, yet we will show that the 3D majority vote model exceeds Trybula's bounds. Specifically, we will prove that the intransitivity of the majority vote model is not bounded away from 1 by any constant. As a result we are able to conclude that our model, despite not relying on any dependence between individuals or heterogeneous noise distributions, is qualitatively different from any univariate independent RUM.

Theorem 5.1. *Given $\epsilon > 0$ and n i.i.d. random variables with distribution g whose pairwise difference follow the difference distribution function D_g which is continuous, strictly increasing on $\mathbb{R}$ with finite variance c, there exists an attribute matrix M for n players such that*

$$\min\{P(X_1 > X_2), P(X_2 > X_3), ..., P(X_n > X_1)\} \geq 1 - \epsilon.$$

Proof. The proof is by construction. Let X_i and X_{i+1} denote cyclically consecutive players, where X_n is cyclically adjacent to X_1. For all i we have that:

$$\Pr(X_i > X_{i+1}) = p_2p_3 + p_1p_3 + p_1p_2 - 2p_1p_2p_3.$$

As before, we write $p_k(i,j)$ as p_k when the intended (i,j) pair is unambiguous. As defined before $p_k = D_g(\mu_i^k - \mu_{i+1}^k)$ where $k \in \{1, 2, 3\}$.

For our construction an appropriate μ is selected for the first three players, where we choose μ_1, μ_2, and μ_3 to be

$$\mu_1 = (a, 3a, -a), \quad \mu_2 = (-a, a, 3a), \quad \mu_3 = (3a, -a, a),$$

where $a \in \mathbb{R}^+$ is a constant to be determined. As D_g is a continuous cumulative distribution function there exists a constant $a_g \in \mathbb{R}^+$ s.t. $\forall a \geq a_g$ for the random variables X_1 and X_2 , $p_1 \geq 1 - \frac{\epsilon}{2}, p_2 \geq 1 - \frac{\epsilon}{2}$. As $p_3 \geq 0$ and $p_1 + p_2 - 2p_1p_2 \geq 0$, we observe that

$$\Pr(X_1 > X_2) \geq p_1p_2,$$

and hence $\Pr(X_1 > X_2) \geq 1 - \epsilon$. A similar argument holds for the pair $\{X_2, X_3\}$.

For players $i \geq 4$ the μ values are defined recursively:

$$\mu_i = (\mu_{i-1} + \mu_1)/2, \text{ for } i = 4, \ldots, n.$$

We aim to show the inequality $\Pr(X_i > X_{i+1}) \geq 1 - \epsilon$ for $i = 3$ and onward. As $p_2 \geq 0$ and $p_1 + p_3 - 2p_1p_3 \geq 0$, we observe that

$$\Pr(X_i > X_{i+1}) \geq p_1p_3.$$

By the given construction of the μ parameters, $\mu_i - \mu_{i+1}$ is decreasing in dimensions one and three for $i = 3, ..., n$. Hence the product p_1p_3 is decreasing for players X_i and X_{i+1} as i is increasing since g is a strictly increasing function.

Define $h(i,j) = p_1(i,j)p_3(i,j)$ for players i and j. By construction we automatically have $h(3,4) > h(4,5) > \ldots > h(n-1,n) = h(n,1)$. The last equality follows from the fact $\mu_{n-1} - \mu_n = \mu_n - \mu_1$. Hence we only need to prove the inequality $p_1(i,j)p_3(i,j) \geq 1 - \epsilon$ for players $i = n, j = 1$. For players X_n and X_1 we then have:

$$\Pr(X_n > X_1) \geq p_1 p_3, \text{ where } p_1 = p_3 = D_g\left(\frac{2a}{2^{n-3}}\right).$$

Now recall Chebyshev's inequality,

$$P(Z \geq z) \leq \frac{Var(Z)}{z^2}.$$

We apply the inequality with Z set to the first and the third dimensions of $X_n - X_1$. As the variance is bounded, $\mathrm{Var}(Z) \leq c$ where c is a positive constant and $\Pr(Z \geq z) = 1 - \Phi(z)$, we have that $D_g(z) \geq 1 - \frac{c}{z^2}$. Using this property with $z = \frac{2a}{2^{n-3}}$, we have

$$\begin{aligned}\Pr(X_n > X_1) \quad &\geq \quad p_1 p_3 \geq \left(1 - \frac{c}{z^2}\right)^2 \geq \left(1 - \frac{2^{2n-8}c}{a^2}\right)^2 \\ &\geq \quad 1 - \frac{2^{2n-7}c}{a^2},\end{aligned}$$

where for the last part we employ the fact that $(1-x)^2 \geq 1 - 2x$ when $x > 0$. We conclude that $1 - \frac{2^{2n-7}c}{a^2} \geq 1 - \epsilon$ when $\epsilon \geq \frac{2^{2n-7}c}{a^2}$. Combining this analysis with the earlier basic analysis for the initial $i = 1, 2, 3$, we can choose a such that $a \geq \max\{a_g, \frac{2^{n-3.5}\sqrt{c}}{\sqrt{\epsilon}}\}$ to complete the proof. □

The above theorem has two important implications. First, large intransitive cycles can be created using the majority vote model in merely three dimensions. Second, we can conclude that the majority vote model for a broad class of difference densities d_g is not equivalent to any independent RUM model.

5.2 Triplet expressivity

The lemma below shows how the three-dimensional majority vote model can represent all the pairwise win probability relations for a set of any three players.

Lemma 5.2. *Given $\epsilon > 0$ and a matrix W of pairwise win probabilities for any three players, there exists an attribute matrix M which can be used to construct a matrix $\tilde{W}$ of pairwise win probabilities from a Majority Vote model such that* $\max_{i,j} |W_{ij} - \tilde{W}_{ij}| \leq \epsilon$.

Proof. The proof is by construction. We are given a pairwise probability matrix W with elements w_{ij}. Let $c_1 = D_g^{-1}(w_{23})$. Define c_2 and c_3 similarly. Let $\alpha_1 = \frac{1-w_{12}}{w_{12}}$. Define α_2 and α_3 similarly in terms of w_{23} and w_{31} respectively. For the three players we can construct M as defined above to be

$$\begin{bmatrix} c_3 & 0 & 0 \\ 0 & -C + c_1 & C \\ C & -C & c_2 \end{bmatrix},$$

where D_g is continuous and strictly increasing and C is a positive constant such that $D_g(-C) \leq \frac{\epsilon}{2}$ and $D_g(C - c_i) \geq 1 - \alpha_i \epsilon$ for $i = 1, 2, 3$. The existence of C is guaranteed by the continuity of D_g.

Table 1: Pairwise comparison datasets.

Dataset	Players	Match-ups	Intransitive triplets
Election A5	16	44298	5
Election A9	12	95888	3
Election A17	13	21037	18
Election A48	10	25848	0
Election A81	11	20803	3
Jester	100	333956	90
Street Fighter (SF)	35	25000	476

We need to prove $|w_{i,j} - \tilde{w}_{i,j}| \leq \epsilon$ for $1 \leq i \neq j \leq 3$. Consider $i = 1$ and $j = 2$. We have $\tilde{w}_{12} = p_1p_2 + p_2p_3 + p_3p_1 - 2p_1p_2p_3$. With $p_1 = w_{12}, p_3 \leq \frac{\epsilon}{2}$, and $1 - \frac{\alpha_1 \epsilon}{2} \leq p_2 \leq 1$ we have

$$\tilde{w}_{12} \leq w_{12} + (\epsilon/2) + (\epsilon/2)w_{12} \leq w_{12} + \epsilon.$$

Next we need to prove the bound in the other direction. Since

$$\tilde{w}_{12} \geq p_1p_2 - 2p_1p_2p_3,$$

we have $\tilde{w}_{12} \geq w_{12} - \epsilon$. An identical argument can be applied to the other two pairs within the triplet, yielding the requisite bound on the maximum over pairs (i,j), as desired. □

It is not hard to verify that a three-dimensional Majority Vote model is not fully expressive for larger subsets of players (five players or more), but we consider the two expressivity results, for cycles and triplets, to be sufficiently flexible to model realistic competition data. Most significantly, we note that neither of these results can be achieved by any independent RUM.

6 EXPERIMENTAL RESULTS

We now study the predictive performance of the different parametric models of intransitive pairwise choice we have discussed: higher dimensional BTL models, the Blade Chest model, and our Gaussian Majority Vote model where each model is furnished with roughly the same number of free parameters ($\approx 3n$) for comparison. For the 3D BTL model, we do not count the elements of Σ as parameters for this comparison. We test the models on three datasets: a commonly studied set of election datasets (A5, A9, A17, A48) [22], the Jester joke rating dataset (underlying a recommendation system) [11], and the *Super Street Fighter IV* dataset (SF) (pairwise match-ups from competitive play between different video game characters) [6]. The election and Jester datasets are ranking datasets which were converted to pairwise comparisons datasets. Dataset details are given in Table 1, where the intransitive triplets count when the empirical pairwise probabilities violate weak stochastic transitivity.

Each pairwise comparison dataset was randomly split into training and test sets using 5-fold cross-validation, splitting at the matchup level[1]. The main evaluation metric used are root mean square error (RMSE) between the predicted probabilities $\tilde{p}$ from the model and the empirical probabilities $\hat{p}$ from the test data scaled by the number

[1] For datasets derived from rankings, ideally the splitting should be done based on rankings, but original rankings were not always available. We compared the splitting based on rankings vs matches for the Jester dataset, where rankings were available, and did not observe any significant difference in the results or the inference

of players:

$$\text{RMSE} = \frac{1}{n}\sqrt{\sum_{i,j\in n, j\neq i} (\hat{p}(i,j) - \tilde{p}(i,j))^2}.$$

We also consider the log-likelihood of each dataset under the models. The Thurstone and 3D Majority Vote models are nested, allowing us to apply a likelihood ratio test between the model estimates.

The non-concave likelihoods of all models were maximized using the Nelder–Mead simplex algorithm. The optimization was repeated multiple times to help escape possible poor local optima.

The RMSE results are provided in Table 2. We see that across the election datasets there is relatively little difference between the intransitive models. For the Jester dataset our model reports an RMSE that is 32% lower than the best Blade-Chest model. For the Blade-Chest models we again use the best results between "inner" and "dist" versions. That said, we still see significant improvements over a one-dimensional Thurstone model with all our models of intransitivity. And of all the models, we see that our 3D Majority Vote model generally has the lowest RMSE and highest likelihood. This empirical performance is particularly impressive given the simplicity of the model. Meanwhile Table 3 provides a comparison for the average cross-validated log-likelihood for the same datasets. We also perform a chi-square test for the nested models, 3D Gaussian MV and 1D Thurstone [18]. The p-values suggest that the unrestricted model (3D Gaussian MV) better fit the datasets specifically when large number of intransitive triplets are present.

6.1 Spectral analysis of dimension

In order to understand the role of higher dimensions in modeling skill, we perform an analysis of the singular values of the attribute matrix $\hat{M}$ learned from data for our 3D Gaussian Majority Vote model. In general if the dimensions of $\hat{M}$ are uncorrelated then this indicates the presence of multiple uncorrelated skills. If the variance of $\hat{M}$ is concentrated on just one dimension, the competition data can be viewed as mostly transitive. We use the explained variance of the vectors of the singular value decomposition of $\hat{M}$ as measures of how correlated the skill dimensions are in the learned representation. In Table 4 we observe that most, but not all, of the variance is explained by the first singular vector of $\hat{M}$ in all the data sets we examine. The model that is "the least one-dimensional" is the model of the Street Fighter data, in the sense that the leading singular vector explains the last percentage of the variance. If of interest, the leading dimension across which the maximum variance is spread could be used as a method for producing ranked orderings of all the alternatives/players.

Table 2: RMSE results, lowest numbers in bold.

Dataset	Thurstone	3D MV	3D BTL	Blade-Chest
A5	0.193 ±0.014	**0.026 ±0.001**	0.033 ±0.009	0.029 ±0.004
A9	0.228 ±0.006	0.018 ±0.002	0.018 ±0.002	**0.017 ±0.002**
A17	0.305 ±0.013	0.032 ±0.003	0.046 ±0.022	**0.031 ±0.002**
A48	0.258 ±0.010	**0.025 ±0.002**	0.027 ±0.002	0.027 ±0.003
A81	0.203 ±0.007	**0.031 ±0.004**	0.035 ±0.003	0.032 ±0.004
Jester	0.187 ±0.017	**0.021 ±0.001**	0.031 ±0.005	0.031 ±0.002
SF	0.280 ±0.018	**0.06 ±0.003**	0.064 ±0.004	**0.06 ±0.002**

Table 3: Negative Log-likelihood values, highest in bold, and likelihood ratio test p-values between Thurstone and 3D Majority Vote models.

Dataset	Thurstone	3D MV (p-value)	3D BTL	Blade-Chest
A5	22136	**22368** (4.2e-5)	21786	22217
A9	49742	**49909** (<1e-8)	49877	49897
A17	11501	**11523** (8.2e-3)	11468	11508
A48	13692	13703 (0.23)	**13721**	13717
A81	10879	**10890** (0.34)	10852	10876
Jester	170021	**178180** (<1e-8)	171830	171510
SF	13256	**13753** (<1e-8)	13199	13680

Table 4: Fractional variance explained by the singular vectors of the Gaussian Majority Vote model matrix.

Dataset	σ_1	σ_2	σ_3
A5	0.62 ±0.08	0.29 ±0.04	0.09 ±0.04
A9	0.61 ±0.05	0.30 ±0.03	0.08 ±0.05
A17	0.65 ±0.10	0.31 ±0.10	0.03 ±0.01
A48	0.66 ±0.05	0.24 ±0.03	0.09 ±0.04
A81	0.68 ±0.05	0.25 ±0.03	0.07 ±0.02
Jester	0.72 ±0.01	0.18 ±0.01	0.08 ±0.01
SF	0.53 ±0.02	0.28 ±0.02	0.18 ±0.03

Table 5: Variance fraction explained by the singular vectors the Gaussian MV model attribute matrix for synthetic data.

Dataset	σ_1	σ_2	σ_3
Synthetic 1D	0.81 ±0.08	0.13 ±0.05	0.06 ±0.05
Synthetic 2D	0.58 ±0.09	0.26 ±0.05	0.16 ±0.05
Synthetic 3D	0.51 ±0.04	0.29 ±0.05	0.20 ±0.03

To complement the empirical analysis, we also simulate data which is inherently one, two, and three dimensional, respectively. This is done by simulating a 3D Majority Vote model with a μ vector that is drawn from a uniform distribution for each player which is 1, 2, or 3-dimensional. For the 3-dimensional μ, the Gaussian Majority Vote model is then simulated. For the 1-dimensional and 2-dimensional μ, a version of the Gaussian Majority Vote model is then simulated where the excess dimensions are unbiased coin-flips. We use this process to produce a pairwise comparison matrices for 50 players with 10,000 games for each pair. We train the majority vote model based on the the simulated data, repeating the training five times, and calculate the fraction of variance explained by each singular vector of the learned $\hat{\mu}$. We see that even when the data is truly three-dimensional, there is still a "dominant" vector that explains most of the variation. When the data is two or one dimensional, the majority of the variation can be explained by fewer singular vectors, as expected.

7 CONCLUSIONS

We establish that intransitivity implies non-convexity in the log-likelihood objective of item-parametric choice models. This is a broad statement about many models of choice, though we note that convexity is possible for parameterizations that are not item-level [19]. We then introduce the Majority Vote model that can represent intransitive relations richer than what are possible with independent RUM models. The model exhibits empirical performance that is comparable to earlier multidimensional models, with the added benefit of simplicity and interpretability. Another significant benefit is the ability to interpret the set of three- (or more) dimensional representations as a point cloud in a latent space, and we give a straight-forward spectral approach to measuring the "dimensionality" of intransitive data using singular values.

Acknowledgements. This work was supported in part by an ARO Young Investigator Award.

REFERENCES

[1] Marschak Jacob Block, Henry David. 1960. Random orderings and stochastic theories of responses. *Contributions to probability and statistics* (1960), 97 −132.

[2] Olivier Bottou, Léon; Bousquet. 2008. The Tradeoffs of Large Scale Learning. *Advances in Neural Information Processing Systems* (2008), 161–168.

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

[4] D. Causeur and F. Husson. 1 December 2005. A 2-dimensional extension of the Bradley–Terry model for paired comparisons. *Journal of Statistical Planning and Inference, Volume 135, Issue 2* (1 December 2005), 245–259.

[5] Sourav Chatterjee. 2015. Matrix estimation by Universal Singular Value Thresholding. *The Annals of Statistics, 43(1)* (2015), 177 − 214.

[6] Shuo Chen and Thorsten Joachims. 2016. Modeling Intransitivity in Matchup and Comparison Data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. ACM, New York, NY, USA, 227–236.

[7] Shuo Chen and Thorsten Joachims. 2016. Predicting Matchups and Preferences in Context. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 775–784.

[8] David F. Gleich and Lek-heng Lim. 2011. Rank Aggregation via Nuclear Norm Minimization. (2011), 60–68.

[9] Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™: A Bayesian Skill Rating System. In *Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS'06)*. MIT Press, Cambridge, MA, USA, 569–576.

[10] R. Hunter. 2004. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics* 32 (2004), 2004.

[11] D. Gupta K. Goldberg, T. Roeder and C. Perkins. 2001. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval, 4(2):* (2001), 133–151.

[12] Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval. *Found. Trends Inf. Retr.* 3, 3 (March 2009), 225–331.

[13] C. Varin M. Cattelan and D. Firth. 2013. Dynamic Bradley-Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* (2013).

[14] Charles F. Manski. 1977. The structure of random utility models. *Theory and decision* (1977), 8(3):229–254.

[15] K. O. May. 1954. Intransitivity, utility, and the aggregation of preference patterns. *Econometrica: Journal of the Econometric Society* (1954), 1–13.

[16] Richard P.Savage. 1960. The Paradox of Nontransitive Dice. *The American Mathematical Monthly Vol. 101, No. 5* (1960), 97 −132.

[17] Stephen Ragain and Johan Ugander. 2016. Pairwise Choice Markov Chains. *Advances in Neural Information Processing Systems* (2016), 3198–3206.

[18] Karin Schermelleh-Engel, Helfried Moosbrugger, and Hans Müller. 2003. Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online* 8, 2 (2003), 23–74.

[19] Arjun Seshadri, Alexander Peysakhovich, and Johan Ugander. 2019. Discovering Context Effects from Raw Choice Data. *arXiv preprint arXiv:1902.03266* (2019).

[20] Christoph Freudenthaler Steffen Rendle. 2014. Improving pairwise learning for item recommendation from implicit feedback. *Proceedings of the 7th ACM international conference on Web search and data mining* (2014).

[21] Louis L. Thurstone. 1927. A law of comparative judgment. *Psychological review, 34(4):273* (1927).

[22] N. Tideman. 2006. Collective decisions and voting. *Ashgate Burlington* (2006).

[23] Kenneth E Train. 2009. *Discrete Choice Methods with Simulation.* Cambridge University Press.

[24] S Trybula. 1965. On the paradox of n random variables. *Zastos Math 8.* (1965), 143− 154.

[25] J Yellott. 1977. Relationship between Luces choice axiom, thurstones theory of Comparative judgement, and double exponential distribution. *Journal of Mathematical Psychology Vol. 15* (1977), 109–144.