A New and Flexible Approach to the Analysis of Paired Comparison Data

Ivo F. D. Oliveira IVODAVID@GMAIL.COM
Department of Science, Engineering and Technology
UFVJM - Federal University of the Valleys of Jequitinhonha and Mucuri
Teofilo Otoni, Minas Gerais, Brazil

Nir Ailon NAILON@CS.TECHNION.AC.IL
Department of Computer Science
Technion - Israel Institute of Technionlogy
Haifa, Israel

Ori Davidov DAVIDOV@STAT.HAIFA.AC.IL
Department of Statistics
University of Haifa
Haifa, Israel

Editor: Nicolas Vayatis

Abstract

We consider the situation where I items are ranked by paired comparisons. It is usually assumed that the probability that item i is preferred over item j is $p_{ij} = F(\mu_i - \mu_j)$ where F is a symmetric distribution function, which we refer to as the comparison function, and μ_i and μ_j are the merits or scores of the compared items. This modelling framework, which is ubiquitous in the paired comparison literature, strongly depends on the assumption that the comparison function F is known. In practice, however, this assumption is often unrealistic and may result in poor fit and erroneous inferences. This limitation has motivated us to relax the assumption that F is fully known and simultaneously estimate the merits of the objects and the underlying comparison function. Our formulation yields a flexible semi-definite programming problem that we use as a refinement step for estimating the paired comparison probability matrix. We provide a detailed sensitivity analysis and, as a result, we establish the consistency of the resulting estimators and provide bounds on the estimation and approximation errors. Some statistical properties of the resulting estimators as well as model selection criteria are investigated. Finally, using a large data-set of computer chess matches, we estimate the comparison function and find that the model used by the International Chess Federation does not seem to apply to computer chess.

Keywords: linear stochastic transitivity, statistical ranking, semi-definite programming, model selection, sensitivity analysis, chess

1. Introduction

There are many situations in which a preference or a ranking among a set of items is desired. Ranking methods are widely used in settings such as product testing (Cremonesi et al., 2010), the evaluation of political candidates (Saari, 1995; Pacuit, 2012), psychometrics (Regenwetter et al., 2011), machine learning (Ailon, 2012; Shah et al., 2015a) and sports

(Herbrich et al., 2007; Govan, 2008). An ordering of a set of items can be inferred from different types of data including scores (Balinski and Laraki, 2010) and ranked lists (Marden, 1996). A ranked list may be complete, i.e., all items are compared and ranked, or partial or incomplete, when only a subset of items is compared and ranked. In particular, paired comparison data is obtained if all comparisons involve only two items (David, 1988). Here we focus on paired comparisons with a binary outcomes.

Given a set of I items, also called objects, or players, let Y_{ijk}, $1 \leq i, j \leq I$, be independent binary random variables where $Y_{ijk} = 1$ if item i is preferred over item j on their k^{th} comparison and $Y_{ijk} = 0$ otherwise. The probability of this event is denoted by p_{ij}, i.e., $p_{ij} = \mathbb{P}(Y_{ijk} = 1)$. We assume that m_{ij} comparisons were observed between each pair of items and we let $Y_{ij} = \sum_{k=1}^{m_{ij}} Y_{ijk}$ denote the number of times item i was preferred over item j. Further let $\boldsymbol{P} = [p_{ij}]$ denote the $I \times I$ underlying probability matrix.

Typically, it is assumed that

$$p_{ij} = F(\mu_i - \mu_j) \tag{1}$$

where $\mu_i, \mu_j \in \mathbb{R}$ are the merits (also called skills, scores or ratings) of items i and j respectively, and $F : \mathbb{R} \to [0, 1]$ is a known, strictly increasing, comparison function, i.e., a symmetric absolutely continuous distribution function. We assume that the merits are fixed and unknown. In some situations the merits may vary according to an "effort" (Jia et al., 2013), or depend on covariates (Herbrich et al., 2007; Allison and Christakis, 1994). Model (1) implies that $p_{ij} + p_{ji} = 1$ and imposes a form of stochastic transitivity (Morrison, 1963; Regenwetter et al., 2011) which is known as linear stochastic transitivity (LST). Models satisfying (1) will be referred to as LST models. Various LST models have been proposed, these differ in the choice of the comparison function F. In particular, two canonical LST models have been widely studied; the Thurstonian model (Thurstone, 1927) and the (Zarmelo) Bradley-Terry-Luce model (Zermelo, 1928; Bradley and Terry, 1952). The Bradley-Terry-Luce model (BTL, henceforth) assumes that F is a standard logistic distribution whereas Thurstone's model assumes F is a standard normal distribution. There are literally thousands of studies which employ these models and their variants. Other, albeit less popular, LST models are also studied in literature, e.g., the Threshold model which employs the Laplace distribution and is used for modelling animal behavior (Yellott, 1970), and the locally linear model (Batchelder et al., 1992) which employs a uniform on $[-1, 1]$ distribution.

The assumption that F is known has been recognized as rather unrealistic (Morrison, 1963; David, 1988; Regenwetter and Davis-Stober, 2008; Hwang, 2009; Shah et al., 2015b; Heckel et al., 2016). This has motivated several authors to resort to the use of less restrictive transitivity relations. A variety of stochastic transitivity relations have been explored in the literature (Morrison, 1963; Regenwetter et al., 2011; Oliveira et al., 2018), the weakest of which is known as weak stochastic transitivity. Under weak stochastic transitivity if $p_{ij} \geq 1/2$ and $p_{jk} \geq 1/2$ then $p_{ik} \geq 1/2$. A stronger form of stochastic transitivity, referred to as strong stochastic transitivity (SST), postulates that if $p_{ij} \geq 1/2$ and $p_{jk} \geq 1/2$ then $p_{ik} \geq \max\{p_{ij}, p_{jk}\}$. Of course model (1) satisfies both of the relations. Various authors have developed methods for analyzing data under these less restrictive assumptions (deCani, 1969; Regenwetter and Davis-Stober, 2008; Chatterjee and Mukherjee, 2016; Shah et al., 2015b). It turns out that the estimation procedures associated with these less restrictive transitivity relations are in general, NP-hard. Since these models provide less structure, they may not be adequate when the comparison graph is sparse and most importantly may provide less predictive power. The SST model, for example, cannot guarantee that stronger players have higher chances than weaker players in knockout tournaments, whereas BTL can (Chung and Hwang, 1978; Israel, 1981; Adler et al., 2017; Baek et al., 2013). Thus we propose a different, potentially more powerful approach, within the LST framework, in which the assumption that the comparison function F is known is relaxed. The proposed methodology is flexible, tractable and retains the desirable computational and statistical characteristics of LST models.

A natural and widely used approach for estimating the model parameters in (1) is by least squares (LS). LS is often the method of choice due to its (relative) computational simplicity. Thus if (1) holds then the LS estimators solve the following optimization problem:

$$\hat{\boldsymbol{\mu}} \in \operatorname{argmin} \sum_{i \neq j} w_{ij}(\hat{\Delta}_{ij} - (\mu_i - \mu_j))^2, \tag{2}$$

where, typically, $\hat{\Delta}_{ij} \equiv F^{-1}(\hat{p}_{ij})$ and $\hat{p}_{ij}$ is an estimator of the probability p_{ij}. For now we assume that $\hat{p}_{ij}$ is bounded away from 0 and 1 and $\hat{p}_{ij} + \hat{p}_{ji} = 1$ so $\hat{\Delta}_{ij}$ is well defined. The weights w_{ij} are given and are either proportional to the variance of the estimated $\hat{\Delta}_{ij}$, or the number of comparisons between items i and j. Notice that (2) admits multiple solutions, for if $\boldsymbol{\mu}^*$ is a solution to problem (2) then so is $\boldsymbol{\mu}^* + v\mathbf{1}$ for any $v \in \mathbb{R}$. A unique solution exists if the comparison graph is connected and an additional linear constraint such as $\sum_i \mu_i = 0$ is enforced (Tsukida and Gupta, 2011).

When all w_{ij} in (2) are equal, then the solution is of the form $\mu_i^* = \kappa \sum_j \hat{\Delta}_{ij}$ for some $\kappa > 0$. For this reason the LS method is sometimes referred to as the row—sum procedure (Huber, 1963). There are other well known ranking methods which can be viewed as row—sum procedures with varying definitions of $\hat{\Delta}_{ij}$. For example, the Copland Method, popular within the voting literature (Levin and Nalebuff, 1995; Favardin et al., 2002), is a row sum procedure in which $\hat{p}_{ij}$ is defined as $\hat{p}_{ij} \equiv (\sum_k Y_{ijk})/I$ and $F(x) = 1/2 + x$ for $x \in [-1/2, 1/2]$ where Y_{ijk} equals one if the kth voter prefers candidate i above candidate j and zero otherwise. The Borda Count can be also shown to be a row—sum method. Another LS variant, known as Massey Ratings, is widely used in the rating of sport teams in college football, basketball, hockey, and baseball, see Chapter 4 of Massey (2017). For more on the LS literature, refer to Hodge rank in Jiang et al. (2010).

Our Contribution. We will weaken the assumption that the comparison function F is known and assume only that it belongs to, a new, large family of parametric functions. Our parametric set, can be understood as an interior approximation, with arbitrary precision, to the full set of comparison functions. We then simultaneously estimate the merits as well as the comparison function F by generalizing the LS approach. Estimating the probabilities p_{ij} is now an easy consequence. We show that this can be done efficiently both computationally and statistically. In particular, we develop a procedure that takes as input an estimate of the probability matrix and returns an estimate of the comparison function F, the merit vector $\boldsymbol{\mu}$ and a refinement of the original estimate of the probability matrix. Estimation reduces to a semi-definite programming problem with a tractable solution. We provide a thorough sensitivity analysis and derive statistical properties such as convergence and concentration bounds on the refined estimator and the estimated function. By applying our methodology to a large data-set of computer chess matches, we verify that the ubiquitous (Zarmelo) Bradley-Terry-Luce model may be inappropriate for computer chess.

2. Formulation and Estimation

Formulation: Least Squares Estimation Over Polynomial Families. First, we may generalize problem (2) by rewriting it in matrix notation in the following way

$$\hat{\boldsymbol{\mu}} \in \text{argmin}_{\boldsymbol{\mu}} ||F^{-1}(\hat{\boldsymbol{P}}) - \Delta\boldsymbol{\mu}||_{\boldsymbol{W}}. \tag{3}$$

Here $\Delta\boldsymbol{\mu}$ is an $I \times I$ matrix whose ij^{th} element is $\mu_i - \mu_j$ and $F^{-1}(\hat{\boldsymbol{P}})$ is a matrix with the same dimensions whose ij^{th} element is $F^{-1}(\hat{p}_{ij})$, if $w_{ij} > 0$ and 0 otherwise. Unless specified otherwise, in this paper the norm $|| \cdot ||_{\boldsymbol{W}}$ will be the weighted Frobenius (semi-)norm, with pre-specified weights and $|| \cdot ||$ will be its unweighted counterpart. With a slight abuse of notation we will refer to the frobenius norm as the $\mathcal{L}_2$ norm. The minimization in (3) can also be formulated with respect to the sum of the absolute values of the elements, which we refer to as the $\mathcal{L}_1$ norm, or maximum value of the elements of a matrix, which we refer to as the $\mathcal{L}_\infty$ norm. The mechanics involved in solving (3) are norm dependent. If one views Δ as an operator from $\mathbb{R}^I \rightarrow \mathbb{R}^{I \times I}$ then $\Delta\hat{\boldsymbol{\mu}}$ is the projection of $F^{-1}(\hat{\boldsymbol{P}})$ on the image set of the operator Δ. Finally note that the least squares procedure takes an estimator $\hat{\boldsymbol{P}}$ and produces a refined estimator denoted by $\boldsymbol{P}^* = F\left(\Delta\hat{\boldsymbol{\mu}}\right)$.

The assumption that the comparison function F is known is relaxed and instead it is assumed that $F \in \mathcal{F}$ where $\mathcal{F}$, where:

A$_1$ *(Parametric Assumption): The family $\mathcal{F}$ indexed by $\boldsymbol{\beta} \in \mathbb{R}^{D+1}$ consists of all distribution functions whose inverse, i.e., its quantile function, may be written as a polynomial, of the form*

$$F_{\boldsymbol{\beta}}^{-1}(p) = \beta_0 + \beta_1 p + ... + \beta_D p^D \text{ where } p \in [0, 1/2]. \tag{4}$$

Equation (4) defines a quantile regression model (Takeuchi et al., 2006; Su, 2015). By the LST condition $F_{\boldsymbol{\beta}}$ is symmetric, i.e., $F_{\boldsymbol{\beta}}(x) + F_{\boldsymbol{\beta}}(-x) = 1$ so $F_{\boldsymbol{\beta}}^{-1}(p) + F_{\boldsymbol{\beta}}^{-1}(1-p) = 0$. It immediately follows that $F_{\boldsymbol{\beta}}^{-1}(p)$ is also a polynomial when $p \in [1/2, 1]$. Also, (4) implies that the support of $F_{\boldsymbol{\beta}}$ is the finite interval $[\beta_0, -\beta_0]$, where $\beta_0 < 0$. It is further assumed that:

A$_2$ *(Lipschitz Assumption): For all $F_{\boldsymbol{\beta}} \in \mathcal{F}, F_{\boldsymbol{\beta}}$ is L-Lipschitz and $||\boldsymbol{\beta}||_\infty \leq U$* for some constants L and U.

We note that each fixed value of (D, U, L) generates a parametric family of distributions $\mathcal{F}(D, U, L)$; which we denote for convenience by $\mathcal{F}$. This is a new, non-standard, rich family of distributions, in which the quantile function, not the density, is parametrized. Figure 1 shows that the Bradley-Terry-Luce model can be can be approximated by a low degree polynomial over the interval $p \in [0.01, 0.99]$. Furthernote, that by increasing D, U and L we can approximate any quantile function with arbitrary precision.

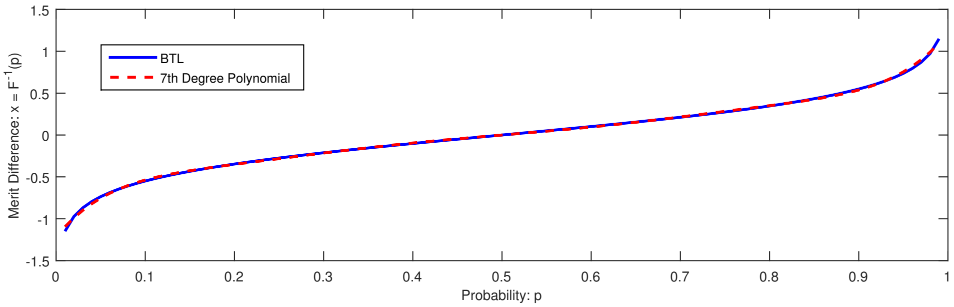

Figure 1: An approximation of the Bradley-Terry-Luce quantile function over the interval $p \in [0.01, 0.99]$ by a function $F \in \mathcal{F}$ for $D = 7$.

It is known that F and G are equivalent comparison functions iff $F(x) = G(\kappa x)$ for some positive κ, see Yellott (1977), and therefore F^{-1} is equivalent to G^{-1} iff $\kappa F^{-1}(p) = G^{-1}(p)$. A valid linear constraint on the coefficient vector $\boldsymbol{\beta}$ is thus imposed to ensure identifiability. For example, fixing the support of F, which amounts to fixing $\beta_0 < 0$, is sufficient. Another natural choice is to fix the derivative of F at 0, which amounts to fixing $\beta_1 > 0$. A rescaling argument shows that the resulting inferences do not depend on the chosen constraint. As noted earlier, identifiablity requires that the merits satisfy a constraint. Henceforth it will be assumed that:

A$_3$ *(Scaling Assumption): The merits and the comparison function are scaled to satisfy*

$$\sum_i \mu_i = 0 \quad \textit{and} \quad \beta_0 = -1. \tag{5}$$

Thus, if F belongs to $\mathcal{F}$ we may estimate $(\boldsymbol{\mu}, \boldsymbol{\beta})$ by solving the following optimization problem:

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}) \in \text{argmin}_{\boldsymbol{\mu} \in \mathbb{R}, F_{\boldsymbol{\beta}} \in \mathcal{F}} ||F_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{P}}) - \Delta\boldsymbol{\mu}||_{\boldsymbol{W}}. \tag{6}$$

Although (6) is a least squares problem it is non-standard as "both sides", i.e., the "predictor" and the "response" in the regression equation, are associated with unknown parameters which are estimated simultaneously. In the following subsection we will study problem (6) under assumptions A$_1$ to A$_3$.

Solution via Semidefinite Programming. Our first concern is to characterize the set of feasible solutions for $(\boldsymbol{\mu}, \boldsymbol{\beta})$. As noted earlier the quantile function (4) satisfies $F^{-1}(p) = -F^{-1}(1-p)$ so we need only consider constraints generated by values $p \in [0, 1/2]$. In this interval $F_{\boldsymbol{\beta}}^{-1}(p)$ is increasing hence its derivative, which is a polynomial of degree $D-1$, is non-negative. Furthermore, the Lipschitz continuity constraint on F implies that F^{-1} is strictly monotone with derivative greater or equal to $1/L$. Thus,

$$\beta_1 + \ldots + D\beta_D p^{(D-1)} - \frac{1}{L} \geq 0 \text{ for all } p \in [0, 1/2] \tag{7}$$

In addition $F(0) = 1/2$, so $F_{\boldsymbol{\beta}}^{-1}(1/2) = 0$ and $||\boldsymbol{\beta}||_\infty \leq U$ therefore

$$\sum_{i=0}^{D} \left(\frac{1}{2}\right)^i \beta_i = 0 \text{ and } -U \leq \beta_i \leq U \text{ for all } i. \tag{8}$$

Minimizing $||F^{-1}(\hat{\boldsymbol{P}}) - \Delta\boldsymbol{\mu}||_{\boldsymbol{W}}$ subject to (7) and (8) yields a semi-infinite programming problem (SIP) (Mutapcic and Boyd, 2009; Stein, 2012), i.e., an optimization problem with an infinite number of, in this case linear, constraints.

There are a number of methods for solving SIPs. One natural approach is discretization, which in our case means replacing (7) by N constraints of the form $\beta_1 + \beta_2 p_j ... + D\beta_D p_j^{(D-1)} - 1/L \geq 0$ where $0 < p_1 < p_2 < ... < p_N < 1/2$ for some finite N. This yields a simple quadratic programming problem. From a practitioner's point of view, discretization may be a method of choice due to its simplicity. This is specifically true when $|| \cdot ||_{\boldsymbol{W}}$ is either the weighted or unweighted $\mathcal{L}_1$ or $\mathcal{L}_\infty$ norms, since in these cases discretization results in a simple linear program. However, the resulting estimate of F is not guaranteed to be strictly monotone (although this can be overcome, see section 3.2 of Mutapcic and Boyd 2009) and more importantly in the worst case the solution may not be polynomially computable. These issues may be overcome by noting that the constraints in (7)-(8) are equivalent to a combination of linear constraints and positive semi-definite cone constraints, see Parrilo (2016) for further details. Thus our optimization problem is (also) a semi-definite optimization problem (SDP). There is a large literature on SDPs (Nemirovski and Todd, 2009) and in particular it is known that SDPs can be solved by interior point methods (Vandenberghe and Boyd, 1996) in polynomial time. Therefore we may formally rewrite (6) as:

Theorem 1 *Given $U, L \geq 0$, $D = 2d+1$ with $d \in \mathbb{N}$, a symmetric weight matrix $\boldsymbol{W}$ and an estimator $\hat{\boldsymbol{P}}$, then problem (6) is equivalent to:*

$$\begin{array}{ll} \textit{minimize} & \sum_{(i,j)\in\mathcal{S}} w_{ij}(\beta_0 + ... + \beta_D \hat{p}_{ij}^D + \mu_j - \mu_i)^2 \\ \textit{subject to} & \\ & \beta_i = \frac{1}{i}(\frac{1}{2}t_{i-2} - t_{i-3} + s_{i-1}) \textit{ for } i = 2, ..., D, \\ & \beta_1 = s_0 + \frac{1}{L}, \qquad \sum_{k=0}^{D} \left(\frac{1}{2}\right)^k \beta_k = 0, \qquad ||\boldsymbol{\beta}||_\infty \leq U, \\ & \\ & s_i = \sum_{j+k=i} Q^0_{jk} \textit{ for } i = 0, ..., D-1, \qquad \boldsymbol{Q^0} \in \mathbb{S}_+^{d+1}, \\ & t_i = \sum_{j+k=i} Q^1_{jk} \textit{ for } i = 0, ..., D-3, \qquad \boldsymbol{Q^1} \in \mathbb{S}_+^{d}. \end{array} \tag{9}$$

where $\mathcal{S} = \{(i,j) \mid p_{ij} \leq .5 \text{ or } p_{ij} = .5 \text{ and } i < j\}$, and $t_{D-1} = t_{D-2} = t_{-1} = t_{-2} = 0$, $\mathbb{S}_+^k$ is the set of $k \times k$ symmetric positive semi-definite matrices, and the rows and columns of $\boldsymbol{Q^0}$ and $\boldsymbol{Q^1}$ are indexed by 0 to d and 0 to $d-1$ respectively.

For brevity we present here only the case when D is odd, a similar characterization holds for D even. We note that analogues of Theorem 1 could also be formulated for the $\mathcal{L}_1$ and $\mathcal{L}_\infty$ norms and their weighted versions in which case the constraints in (9) would remain unaltered whereas the objective function would be as defined by the corresponding norm.

Note that (9) admits a unique solution when the objective function is positive definite. Lemma 2 below provides an example in a simple but important case. In large samples the solution to (9) is uniquely determined when the system of equations $\beta_1 p_{ij} + ... + \beta_D p_{ij}^D - (\mu_i - \mu_j) = -\beta_0$ for $(i, j) \in \mathcal{S}$ is of full rank. This condition is met when (i) there are at least $I + D - 1$ connected pairs $(i, j) \in \mathcal{S}$; which (ii) the coefficients appearing in the linear equations, which are derived from the comparison probabilities p_{ij}, are sufficiently diverse, otherwise the resulting equations would not be linearly independent. Thus we assume that:

A$_4$ *(Connectivity & Diversity Assumption): The comparison graph has at least $I + D - 1$ edges. These edges correspond to a set of linearly independent equations of the form $\beta_1 p_{ij} + ... + \beta_D p_{ij}^D - (\mu_i - \mu_j) = -\beta_0$.*

If we label these equations $(ij)_1, ..., (ij)_{D+I-1}$ then together with the constraint $\sum \mu_i = 0$ we may write the resulting system of equations (with a slight abuse of notation) as:

$$\boldsymbol{A} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} -\boldsymbol{\beta_0} \\ 0 \end{pmatrix} \tag{10}$$

where the k'th row of $\boldsymbol{A}$ is $\boldsymbol{A}_k = (p_{(ij)_k} \;\; ... \;\; p_{(ij)_k}^D \quad -\boldsymbol{e}_{(ij)_k})$ for $k = 1, ..., I + D - 1$ and $\boldsymbol{A}_{I+D} = (\boldsymbol{0}^T \;\; ... \;\; \boldsymbol{1}^T)$ where $\boldsymbol{e}_{ij} \in \mathbb{R}^I$ is defined by $\boldsymbol{e}_{ij} \equiv \boldsymbol{e}_i - \boldsymbol{e}_j$, where $\boldsymbol{e}_i$ is the standard basis. The condition number of $\boldsymbol{A}$ plays a role in the quality of our estimators.

Define $\boldsymbol{p} = (1,\ p,\ ...\ ,\ p^D)^T$ and note that if all the weights are equal then

$$\mu_i = \frac{1}{I} \sum_k F_{\boldsymbol{\beta}}^{-1}(p_{ik}) = \frac{1}{I} \sum_{(i,k) \in \mathcal{S}} F_{\boldsymbol{\beta}}^{-1}(p_{ik}) - \frac{1}{I} \sum_{(i,k) \notin \mathcal{S}} F_{\boldsymbol{\beta}}^{-1}(p_{ki}) \tag{11}$$

and thus we may eliminate $\boldsymbol{\mu}$ from (9) by means of equation (11). This considerably reduces the size of the SDP at hand when the number of items I is larger than D. Algorithm PolyRank (displayed below) takes advantage of this. Furthermore:

Lemma 2 *When all weights w_{ij} are equal, then, the objective function of problem (9) is equivalent to minimizing $\boldsymbol{\beta}^T \boldsymbol{M} \boldsymbol{\beta}$, where*

$$\boldsymbol{M} \equiv \sum_{(ij) \in \mathcal{S}} \boldsymbol{v}_{(ij)} \boldsymbol{v}_{(ij)}^T \tag{12}$$

and $\quad \boldsymbol{v}_{(ij)} \equiv -I\hat{\boldsymbol{p}}_{ij} + \sum_{(ik) \in \mathcal{S}} \hat{\boldsymbol{p}}_{ik} - \sum_{(ik) \notin \mathcal{S}} \hat{\boldsymbol{p}}_{ki} - \sum_{(jk) \in \mathcal{S}} \hat{\boldsymbol{p}}_{jk} + \sum_{(jk) \notin \mathcal{S}} \hat{\boldsymbol{p}}_{kj}.$

Hence the estimators can be efficiently calculated in three steps:

```
Algorithm: PolyRank

Input: $\hat{\boldsymbol{P}} \in [0,1]^{I \times I}$ $D \in \mathbb{N}$ an odd number and $U, L \geq 0$.

  1. Preprocessing: Calculate $\boldsymbol{M}$ as in equation (12);
  2. Functional Estimation: $\hat{\boldsymbol{\beta}} \equiv$ argmin $\{\boldsymbol{\beta}^T \boldsymbol{M} \boldsymbol{\beta}$ subject to $(9)$ and $(5)\}$;
  3. Merit Estimation: $\hat{\mu}_i = \frac{1}{I} \sum_{(i,k) \in \mathcal{S}} F_{\hat{\boldsymbol{\beta}}}^{-1}(\hat{p}_{ik}) - \frac{1}{I} \sum_{(i,k) \notin \mathcal{S}} F_{\hat{\boldsymbol{\beta}}}^{-1}(\hat{p}_{ki})$;

Output: $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{D+1}$, $\hat{\boldsymbol{\mu}} \in \mathbb{R}^n$ and $\boldsymbol{P}^* \equiv F_{\hat{\boldsymbol{\beta}}}(\Delta \hat{\boldsymbol{\mu}})$
```

Step 1 may be performed with no more than $O(I^3 D + I^2 D^2)$ operations, Step 2 with no more than $O(D^2 \sqrt{D})$ operations and Step 3 with no more than $O(I^2 D)$ operations. Thus, the overall computational complexity of solving problem (6) is no more than $O(I^3 D + I^2 D^2 + D^2 \sqrt{D})$. Notice also that Steps 1 and 3 can be done in a distributed fashion. When the weights are not all equal the merits cannot be written as in (11) and therefore Algorithm PolyRank as stated above cannot be used, in that case we solve (9) directly. Nevertheless we will refer to all versions of our estimation procedure as PolyRank. In our experience, problem (6) with any norm (weighted or unweighted) can be tackled successfully with a generic convex optimization solver on a desktop computer for problems of moderate size (e.g. with $D \leq 10$ and $I \leq 120$) in at most 2 or 3 seconds. Using the three step procedure (with the same generic solver) allows easy scaling up to problems where $D \leq 20$ and $I \leq 10000$. If (6) is treated as a SIP and solved via discretization, then significant reduction in computation time is observed at the cost of loosening the guarantee of optimality.

Remark 1 Notice that if the machine precision is ϵ and if D is such that $\epsilon > (1/2)^D$ then the last terms of the polynomial F^{-1} are rounded to zero. Therefore for standard 32 bit floating point arithmetic one should choose D at most 22, similarly for a 64 bit arithmetic D should not surpass 44. $O(V^2 \sqrt{V})$

Remark 2 In theory F can be recovered from F^{-1} exaclty via Lagrange Inversion Theorem. Numerically though, calculating $F_{\boldsymbol{\beta}}(\mu_i - \mu_j)$ reduces to a polynomial root-finding problem. Although root-finding is an ill-conditioned problem for general polynomials (Trefethen 2011), it may be solved via binary search (with linear convergence in the worst case) or via Newton steps (with possible quadratic convergence).

Remark 3 When the weights w_{ij} are not all equal, or $|| \cdot ||_{\boldsymbol{W}}$ represents the $\mathcal{L}_1$ or $\mathcal{L}_\infty$ norms, then a full SDP with $V = I + D$ variables must be solved. In these cases the simplifying row-sum structure is absent and the worst case bounds are well known, and of the order $O(V^2 \sqrt{V})$, see the general SDP literature (Vandenberghe and Boyd, 1996).

3. Sensitivity Analysis

The goal of this section is to investigate the sensitivity of PolyRank with respect to the input matrix $\hat{\boldsymbol{P}}$. There are several reasons for developing thorough, non-stochastic, sensitivity bounds. Firstly, the analysis serves to clarify the mechanics of PolyRank providing bounds that apply to any choice of $\hat{\boldsymbol{P}}$. Secondly, using the sensitivity bounds statistical properties such as consistency of the refined estimator are easily derived. A third motivation is that different estimators $\hat{\boldsymbol{P}}$ have been investigated in the literature, e.g., Rajkumar and Agarwal (2014); Chatterjee and Mukherjee (2016); Shah et al. (2015b), and since PolyRank may be applied to any of them, the respective bounds on the refined estimator are universal and apply to any $\hat{\boldsymbol{P}}$. Sensitivity analysis is carried out under three increasingly general settings. First, we provide a benchmark by studying the LS method with known F. Then, we consider PolyRank in the case where the model is correctly specified, i.e., $F \in \mathcal{F}$. This is also called the realizable case. Finally, we consider agnostic cases, that is, situations where the model may be misspecified in some way. Three examples of misspecifications are analyzed.

For simplicity we first focus on the unweighted $\mathcal{L}_2$ norm, extensions to the respective weighted versions are similarly obtained.

3.1. Known Comparison Function

Here the function F is assumed to be a known L-Lipschitz continuous function with a U-Lipschitz inverse, i.e., it is bilipschitz. A common assumption in the literature, cf. Shah et al. (2015a,b), is that the probabilities in (1) are bounded away from 0 and 1, i.e., $p_{ij} \in [\epsilon, 1-\epsilon]$, for some $\epsilon > 0$. Over this domain the Bradley-Terry-Luce, Thurstone, Threshold and Locally Linear models are all bilipschitz.

Theorem 3 *Let F be a known L-Lipschitz continuous function with a $4U$-Lipschitz continuous inverse (over their respective domains). Let $\hat{\boldsymbol{\mu}}$ be as in (2) and $\boldsymbol{P^*} = F(\Delta\hat{\boldsymbol{\mu}})$. Then,*

$$||\boldsymbol{P^*} - \boldsymbol{P}|| \leq 4LU||\hat{\boldsymbol{P}} - \boldsymbol{P}||, \tag{13}$$

and

$$||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|| \leq \frac{4U}{\sqrt{2I}}||\hat{\boldsymbol{P}} - \boldsymbol{P}||. \tag{14}$$

If, additionally, it is assumed that $\hat{\boldsymbol{P}}$ obeys strong stochastic transitivity, then the estimators are order preserving, i.e.,

$$\hat{\mu}_i < \hat{\mu}_j \iff \hat{p}_{ij} < \hat{p}_{ji}. \tag{15}$$

By construction the constant $4LU \geq 1$ and so (13) guarantees that $||\boldsymbol{P^*} - \boldsymbol{P}||$ will be at most a constant times $||\hat{\boldsymbol{P}} - \boldsymbol{P}||$. Although it may be possible to improve the constant in (13), its value can never be less than 1, for if not, one could generate a converging sequence $\boldsymbol{P_1}^*, \boldsymbol{P_2}^*, ...$ by recursively applying the LS refinement to any initial (blind) guess of $\hat{\boldsymbol{P}}$. This argument holds for any refinement procedure, including PolyRank. Also, by construction the LS refinement defines $\boldsymbol{P}^* = F(\Delta\hat{\boldsymbol{\mu}})$ and thus $\min_{\boldsymbol{\mu}} ||F^{-1}(\boldsymbol{P}^*) - \Delta\boldsymbol{\mu}|| = 0$ for $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$ and so no improvement will be obtained via recursive LS type refinements. The bound in (13) is a "worst case" bound and on average we often observe that $||\boldsymbol{P^*} - \boldsymbol{P}||$ is indeed smaller that $||\hat{\boldsymbol{P}} - \boldsymbol{P}||$. For other norms refer to the appendix.

The benchmarks provided by Theorem 3 will be used for comparison with the more general cases tackled by PolyRank. As will be shown, inequality (13) also holds when F is unknown (with different constant factors); similarly, the order preservation is maintained in all the settings considered.

3.2. Realizable Case

Under the hypothesis of realizability, i.e., when the model is correctly specified, we have:

Theorem 4 *Let $\boldsymbol{P}^{*} = F_{\hat{\boldsymbol{\beta}}}(\Delta\hat{\boldsymbol{\mu}})$ where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ are estimated using* PolyRank. *Then,*

$$||\boldsymbol{P}^{*} - \boldsymbol{P}|| \leq (1 + 4LU)||\hat{\boldsymbol{P}} - \boldsymbol{P}||, \tag{16}$$

and

$$\left|\left|\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \end{pmatrix}\right|\right| \leq K_1||\hat{\boldsymbol{P}} - \boldsymbol{P}||, \tag{17}$$

as well as,

$$\max_{x\in[-1,1]} |F_{\hat{\boldsymbol{\beta}}}(x) - F_{\boldsymbol{\beta}}(x)| \leq K_2||\hat{\boldsymbol{P}} - \boldsymbol{P}||_{\infty}, \tag{18}$$

where $K_1 \leq U(1 + 4LU)\sqrt{D(I + D)}||\boldsymbol{A}^{-1}||$ and $K_2 \leq 16LDU^2 \max_i \sum_j |A^{-1}_{ij}|$. If, additionally, it is assumed that $\hat{\boldsymbol{P}}$ obeys strong stochastic transitivity, then the estimators are order preserving.

Notice that the constants in (13) and (16) depend solely on the product of the Lipschitz constants of F and F^{-1}. Moreover the constant in (16) does not depend on D nor on the condition number of A . In contrast, the constants in inequalities (14) and (17) do depend on the dimensions of the problem. Equation (18) guarantees the convergence of $F_{\hat{\boldsymbol{\beta}}}$ to the true function F with respect to the Chebyshev distance, thus, one can eventually recover F with arbitrary precision.

3.3. Agnostic Cases

We will now investigate the properties of PolyRank under several types of misspecification. First, we investigate the effect of misspecifying the degree of the polynomial (4). Then, we provide results analogous to those provided by Theorem 4 by replacing the assumption that $F \in \mathcal{F}$ by the assumption that the true F is an analytic function. Finally, we drop the assumption that $\boldsymbol{P}$ satisfies the LST hypothesis all together and verify that we can still derive, albeit, weaker sensitivity bounds and rank consistency properties if strong stochastic transitivity is assumed.

Theorem 5 *Assume the true model satisfies (4), however the fitted model was of degree $D' \leq D-1$. Then for any $\boldsymbol{\beta'}$ of dimension $D' \leq D-1$, a lower bound on the approximation error, in the Chebyshev norm, is:*

$$\frac{1}{2U}\frac{|\beta_D|}{8^D} \leq \max_{\alpha} |F_{\boldsymbol{\beta'}}(\alpha) - F_{\boldsymbol{\beta}}(\alpha)| \tag{19}$$

The lower bound (19) shows that the Chebyshev distance between the true function and the estimated function cannot be arbitrarily minimized when the degree of the fitted polynomial is under-specified. The lower-bound, though, decreases with the value of D at an exponential rate.

Theorem 6 *Let $\boldsymbol{P^*} = F_{\hat{\boldsymbol{\beta}}}(\Delta\hat{\boldsymbol{\mu}})$ where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ are estimated with* PolyRank. *Assume that the true probability matrix $\boldsymbol{P} = F(\Delta\boldsymbol{\mu})$ for some $\boldsymbol{\mu}$ and some unknown L-Lipschitz continuous function F with an analytic inverse function F^{-1} whose coefficients are upper-bounded by U. Then for the estimated probability matrix we have:*

$$||\boldsymbol{P}^* - \boldsymbol{P}|| \leq (1 + 4LU)||\hat{\boldsymbol{P}} - \boldsymbol{P}|| + \frac{1}{2^D} LUI \tag{20}$$

If additionally it is assumed that $\hat{\boldsymbol{P}}$ obeys strong stochastic transitivity, then the estimators are order preserving.

Equation (20) shows that the error of $\boldsymbol{P}^*$ can be controlled under a broad class of analytic functions. The first term is controlled by increasing the precision of $\hat{\boldsymbol{P}}$ and the second term is controlled by choosing larger values for D.

In the following Theorem we will assume no particular structure on $\boldsymbol{P}$, i.e. the probability matrix $\boldsymbol{P}$ need not be consistent with any stochastic transitivity model.

Theorem 7 *Let $\boldsymbol{P^*} = F_{\hat{\boldsymbol{\beta}}}(\Delta\hat{\boldsymbol{\mu}})$ where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ are estimated with* PolyRank. *Then for the estimated probability matrix we have:*

$$||\boldsymbol{P}^* - \boldsymbol{P}|| \leq ||\hat{\boldsymbol{P}} - \boldsymbol{P}|| + L||F^{-1}_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{P}}) - \Delta\hat{\boldsymbol{\mu}}||, \tag{21}$$

If additionally it is assumed that $\hat{\boldsymbol{P}}$ obeys strong stochastic transitivity, then the estimators are order preserving.

An immediate consequence of order preservation is that if $\boldsymbol{P}$ is in the interior of the strong stochastic transitivity set then PolyRank is order-consistent for any consistent estimator $\hat{\boldsymbol{P}}$, i.e., when $\hat{\boldsymbol{P}} \to \boldsymbol{P}$ then the vector $\hat{\boldsymbol{\mu}}$ will correctly recover the underlying order among the items. The error bound in equation (21), though, cannot be controlled as in equation (20), this is so because the second term can be as big as κI^2 for some positive κ even when $\hat{\boldsymbol{P}}$ satisfies strong stochastic transitivity (Shah et al., 2015b).

4. Convergence and Concentration

In this subsection we assume that the model is correctly specified and investigate some properties of the estimators obtained by PolyRank. We start with the case where the comparisons graph is fixed and the number of comparisons per pair, i.e., the m_{ij}'s is allowed to increase. Similar conditions have been considered in literature (Rajkumar and Agarwal, 2014; Shah et al., 2015a).

It is well known that the topology of the comparison graph plays an important role in the quality of the estimators (Shah et al., 2015a; Massey, 2017; Colley, 2002). In particular Shah et al. (2015a) show that, the mean squared errors of the estimated merits from a standard Bradley-Terry-Luce model are proportional to the second eigenvalue of the graph Laplacian. This eigenvalue, referred to as the algebraic connectivity of the graph, measures how "well" the graph is connected (Chung, 1994). In our concentration bounds the number of edges in the comparison graph and the condition number associated with (10) will play a similar role. Let $n = \sum_{i,j} m_{ij}$ be the number of paired comparisons.

Theorem 8 *Let $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\mu}}_n$ be estimated using* PolyRank *with $m_{ij} \equiv w_{ij}n$. Let $\hat{p}_{ij}$ be the usual MLEs. Then for large enough n there are constants K_1 and K_2 such that,*

$$\mathbb{P}\left(\left|\left| \begin{pmatrix} \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \\ \hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} \end{pmatrix} \right|\right| \geq \epsilon \right) \leq K_1 \exp\left(-nK_2\epsilon^2\right), \tag{22}$$

where K_1 and K_2 are discussed bellow.

Theorem 8 shows that the estimators $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\mu}}_n$ converge at an exponential rate and are therefore strongly consistent. Theorem 8 also implies an exponential convergence of $\boldsymbol{P}^*$ to $\boldsymbol{P}$ as well as of $F_{\hat{\boldsymbol{\beta}}}(x)$ to $F_{\boldsymbol{\beta}}(x)$ in the Chebyschev distance. Theorem 8 is proved by first establishing sensitivity bounds for the weighted norm. These are analogues of Theorem 4 and are of the form

$$\left|\left| \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \end{pmatrix} \right|\right| \leq K||\hat{\boldsymbol{P}} - \boldsymbol{P}||_{\boldsymbol{W}}. \tag{23}$$

The constants in (22) are $K_1 = 2|E|$ where $|E|$ is the number of edges in the comparison graph, and $K_2 = 2/(|E|(1+4LU)^2U^2D(I+D)||\boldsymbol{A}_{\boldsymbol{W}}^{-1}||^2)$, where $\boldsymbol{A}_{\boldsymbol{W}}^{-1}$ is defined as in equation (10) with the appropriate modifications for weights. Clearly, $I + D - 1 \leq |E| \leq (I^2 - I)/2$. Of course, smaller values of $|E|$ will provide tighter bounds in equation (22). The value of $||\boldsymbol{A}_{\boldsymbol{W}}^{-1}||$ is a function of, among other things, the topology of the comparison graph. Unfortunately, the condition number of $A_{\boldsymbol{W}}$ is difficult to analyze as it contains a $(I+D-1)\times D$ Vandermonde submatrix which can range from 1 (the best possible condition number) to exponential on the dimensions of the matrix (Pan, 2015). As a rule of thumb Vandermonde matrices are well conditioned when the points $p_{(ij)_1}, ..., p_{(ij)_{D+I-1}}$ are (approximately) spaced over Chebyshev points. Matrix $\boldsymbol{A}_{\boldsymbol{W}}$ also contains a standard $(I + D - 1) \times I$ submatrix of pairings and thus we conjecture that smaller values of the second eigenvalue of the graph Laplacian matrix should also provide tighter estimation bounds.

4.1. Round robin

We now turn our attention to round robin tournaments (Chatterjee and Mukherjee, 2016; Shah et al., 2015b; Simons and Yao, 1999), in which each pair of items is compared m times. If the number of items I is fixed and if $m \to \infty$ then we can use the results described above. A more interesting situations arises when $m = 1$ but the number of items $I \to \infty$. As pointed out earlier, in this setting if $\hat{p}_{ij} \propto Y_{ij}$ then the LS estimator $\hat{\mu}_i$ will be proportional to its Copeland Score (the number of times an item was preferred). Recent papers addressing this setting are by Chatterjee and Mukherjee (2016) and Shah et al. (2015b). In particular they assume strong stochastic transitivity and construct an estimator $\hat{\boldsymbol{P}}_{\text{ISO}}$ which is shown to satisfy:

$$\sup \frac{1}{I^2}\mathbb{E}||\hat{\boldsymbol{P}}_{\text{ISO}} - \boldsymbol{P}||_2^2 \leq C\sqrt{\frac{\log I}{I}}, \tag{24}$$

where the supremum is taken over all matrices that satisfy strong stochastic transitivity. They also show that if the true model was LST then, under some regularity conditions, the upper bound in (24) can be tightened to $O(1/I)$ up to log factors (Shah et al., 2015b). Their estimator is calculated in two steps: (i) first, they sort the items according to their Copeland Score; (ii) then, they perform a two dimensional isotonic regression on the matrix $\boldsymbol{Y}$ assuming the order obtained in (i).

The resulting estimator has two drawbacks when the true model is LST. First, the estimator may be infeasible, i.e., $\hat{\boldsymbol{P}}_{\text{ISO}}$ may not be LST. Our experience indicates that this is frequently the case. In addition the resulting estimator does not fully exploit the benefits of an LST model since the estimated probability matrix is not a Functional of a merit vector and the comparison function. These deficiencies, however, can be addressed by applying PolyRank to their estimator. A trivial consequence of equation (16) is that the refined estimator $\boldsymbol{P}^*$ retains the optimal risk bounds of $\hat{\boldsymbol{P}}_{\text{ISO}}$ and by construction is feasible. We state the full result for completeness:

Theorem 9 *Let $\boldsymbol{P}^*$ be the refinement of $\hat{\boldsymbol{P}}_{ISO}$ using* PolyRank, *where $\hat{\boldsymbol{P}}_{ISO}$ is the estimator of Chatterjee and Mukherjee (2016), then:*

$$\sup \frac{1}{I^2}\mathbb{E}||\boldsymbol{P}^* - \boldsymbol{P}||_2^2 \leq K\frac{\log^2 I}{I}, \tag{25}$$

for some constant K that does not depend on neither I nor D (nor the condition number of A) and the supremum is taken over the set of probability matrices consistent with functions $F \in \mathcal{F}$. This is optimal up to log factors.

5. Numerical Experiments and An Illustrative Example

In the following we describe four experiments performed to further test and investigate PolyRank. Each simulation is performed 1000 times and we report and discuss the average performance under the specified conditions.

Experiment 1: In this experiment we compare the empirical performance of the estimator of $\boldsymbol{P}$ when using PolyRank with a low degree polynomial with its performance given the correct comparison function. Specifically, this is done by generating $I = 20$ items with merits μ_i sampled uniformly from $[0, 10]$. A total of 50 pairs, selected randomly, were compared assuming a Bradley-Terry-Luce (BTL) model. We refine the estimator $\hat{p}_{ij} = (Y_{ij}+1)/(m_{ij}+2)$ with PolyRank using $D = 5$. We also compute the LS estimated with the known F. Figure 2 shows the average distance $||\boldsymbol{P}^* - \boldsymbol{P}||_2$. As expected, the LS method with the correct comparison function performs best, PolyRank performs almost as well and both substantially outperform the initial estimates. This is consistent with our expectations because the BTL model, despite not belonging to the class of functions $\mathcal{F}$, can be well approximated by this class within the range of choice probabilities generated.

Experiment 2 In this experiment we investigate the empirical performance of PolyRank in the round-robin setting when the number of items is increasing. Specifically, we generate a sequence of round-robin tournaments with an increasing number of items. The data is generated assuming model (4) with $D = 5$. The matrix $\boldsymbol{P}$ is estimated using the isotonic

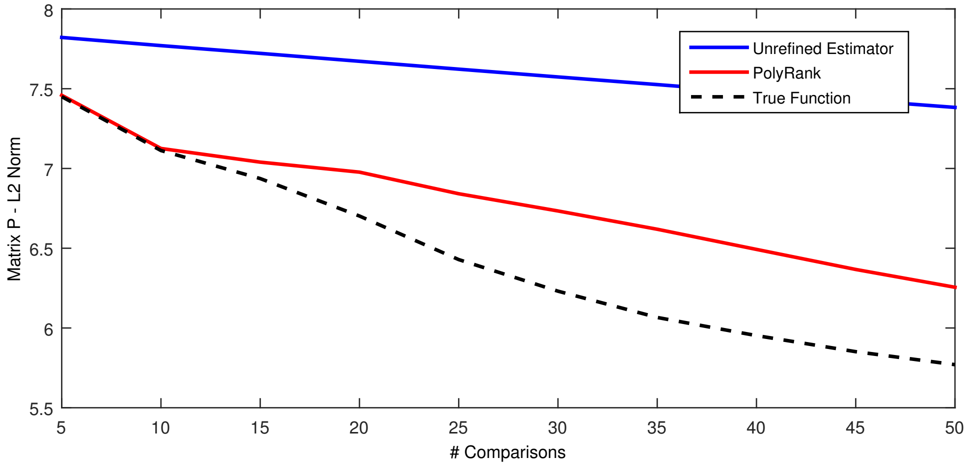

Figure 2: Comparison of refined estimators with low sampling.

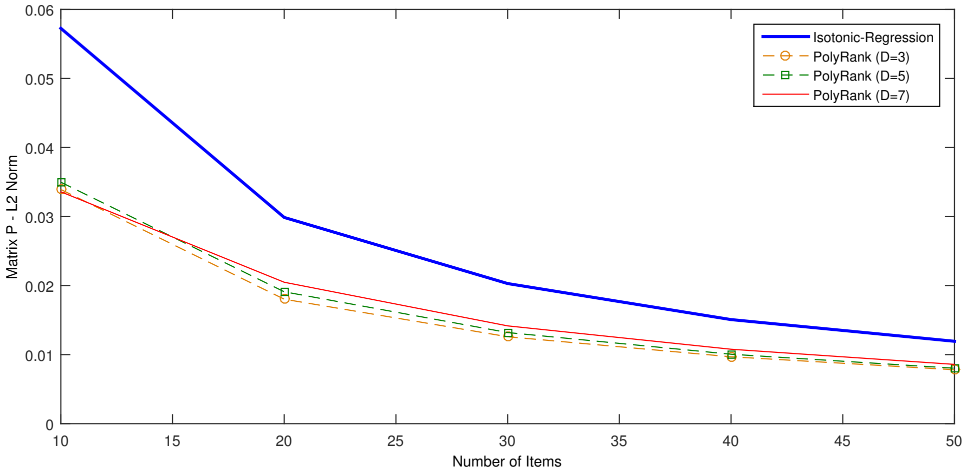

Figure 3: Refined estimators for round-robin tournaments.

regression estimator of Chatterjee and Mukherjee (2016) and refined using PolyRank with $D \in \{3, 5, 7\}$. Figures 3 and 4 display, respectively, the average of $||\boldsymbol{P}^* - \boldsymbol{P}||^2/I^2$ and the average of $||(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \hat{\boldsymbol{\mu}} - \boldsymbol{\mu})||^2/(I + D)$ for $I = 10, 20, 30, 40$ and 50.

Figure 3 shows four curves, all of which decrease with I. The top curve is the risk for the unrefined isotonic-regression based estimator. The estimators refined by PolyRank, which correspond to the lower 4 curves always do better. Notice that over-fitting, i.e., $D = 7$, which corresponds to the second curve from the top, usually results in higher estimation error with no change in the approximation error when compared to $D = 3, 5$. The second curve from the bottom corresponds to the true model. The bottom curve is obtained when $D = 3$, i.e., under under-fitting, and results in the lowest risk. Although this result is somewhat surprising it has been documented also in the context of other models (Claeskens

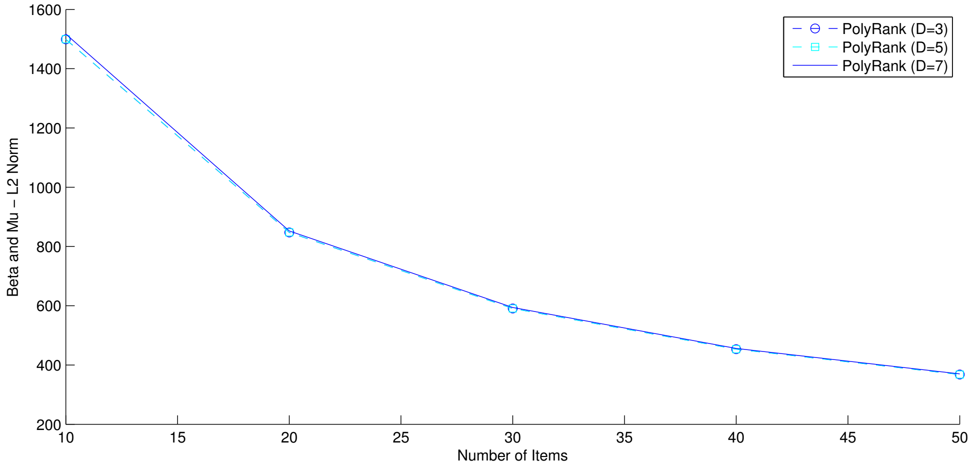

Figure 4: Estimated parameters for round-robin tournaments.

and Hjort, 2006, Chapter 5). This indicates that lower degree polynomial often perform well in practice. In Figure 4 we see that the average error of the estimated parameters decreases as a function of I.

Experiment 3: In this experiment we investigate the performance of PolyRank in the round-robin setting with a fixed number of items and a increasing number of comparisons. we generated a sequence of round-robin tournaments with $I = 10$ and an increasing number of matches. The data is generated assuming model (4) with $D = 3$. The matrix $\boldsymbol{P}$ is estimated using the standard frequency estimator for $\hat{p}_{ij}$ and is refined using PolyRank (with $D = 3$). Figure 5 displays the average of $||\boldsymbol{P}^* - \boldsymbol{P}||^2/I^2$, of $||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2/I$ and of $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2$ for $m_{ij} = 1$ to 5 for all pairs (i, j) and Figure 6 shows the sequence of estimated functions.

The three decreasing curves of Figure 5 show that the variance of the estimators decreases with the amount of paired comparisons. Figure 6 shows that the estimated comparison function converges to its true value. These results and those of Experiment 2 are consistent with Theorems 8 and 9.

Experiment 4: In practice the degree D of the polynomial (4) may not be known in advance. If we choose D to be too small then we may not fully capture the geometry of F, while if D is too large there is a danger of over-fitting and possible numerical problems. In this experiment we investigate the use of some well known model selection criteria (Claeskens and Hjort, 2006) for choosing D. In particular, we test the empirical performance of the Bayesian Information Criterion (BIC) and two variants of the Akaike Information Criterion (AIC) and contrast these with the performance of (leave-one-out) cross-validation. The classical AIC criteria is

$$\mathcal{AIC}(D) \equiv 2(I + D) + n \log(l(D)) \tag{26}$$

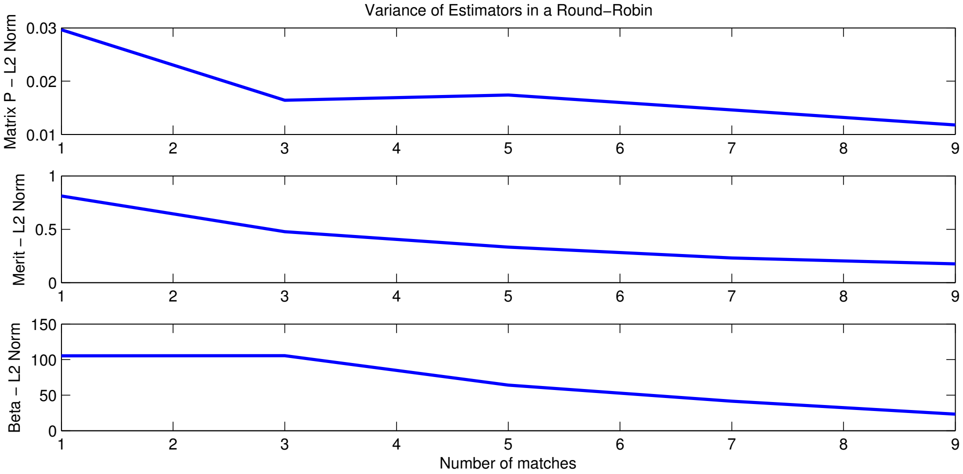

Figure 5: Variance of estimators in a round-robin with increasing number of matches between each pair.

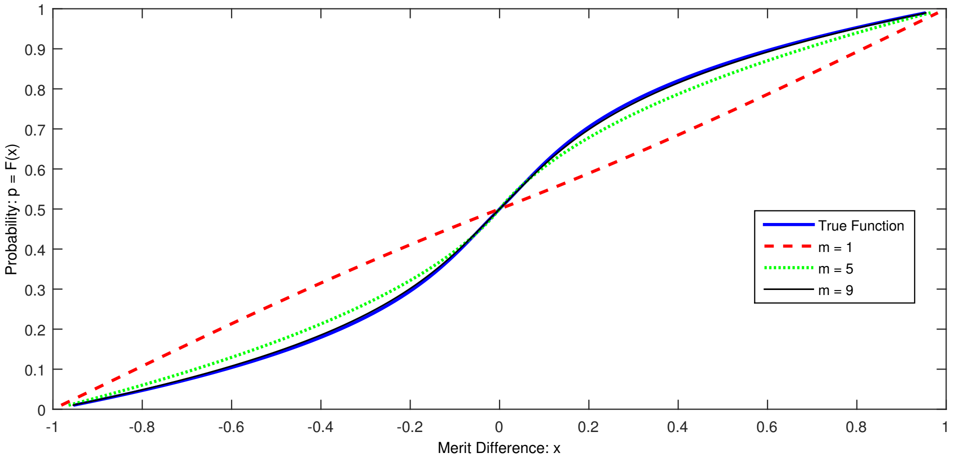

Figure 6: Recovered function graph.

	CV	AICc	AIC	BIC	CV	AICc	AIC	BIC
$D=1$	100%	100%	100%	100%	22%	30.3%	24.4%	37.4%
$D=3$	0%	0%	0%	0%	63.7%	67.3%	69.8%	61.6%
$D=5$	0%	0%	0%	0%	11.8%	1.5%	4.9%	0.1%
$D=7$	0%	0%	0%	0%	2.4%	0.8%	0.8%	0.8%
$D=9$	0%	0%	0%	0%	0.1%	0.1%	0.1%	0.1%
	$m_{ij}=1$				$m_{ij}=3$			

Table 1: Model Selection ($I=6$, $D=5$)

	CV	AICc	AIC	BIC	CV	AICc	AIC	BIC
$D=1$	100%	100%	100%	100%	7.1%	2.8%	2.6%	6.8%
$D=3$	0%	0%	0%	0%	87.4%	89%	85.2%	92.5%
$D=5$	0%	0%	0%	0%	5.3%	8.2%	12.2%	0.7%
$D=7$	0%	0%	0%	0%	.2%	0%	0%	0%
$D=9$	0%	0%	0%	0%	0%	0%	0%	0%
	$m_{ij}=1$				$m_{ij}=3$			

Table 2: Model Selection ($I=12$, $D=5$)

where $l(D)$ is the least-squares loss function, given in (6) and evaluated at the estimated parameters, $I+D$ is the number of parameters in the model and n is the sample-size. The corrected AIC (AICc) is

$$\mathcal{AIC}_c(D) \equiv \mathcal{AIC}(D) + \frac{2(I+D+1)(I+D+2)}{n-I-D-2} \tag{27}$$

and is designed to correct for small sample-sizes. The BIC method penalizes more the number of parameters and is defined by

$$\mathcal{BIC}(D) \equiv (I+D)\log(n) + n\log(l(D)). \tag{28}$$

Tables 1 and 2 compares the AIC, the AICc and the BIC methods in a round-robin data generated with $I=6$ and $I=12$ objects and with $m_{ij}=1$ for all ij, as well as $m_{ij}=3$. The data was generated monotone polynomials of degree 5 randomly selected with uniform coefficients and projected to the monotone cone. We display the frequency in which the methods correctly identify the degree of the polynomial as opposed to overfit/underfit.

Tables 1 and 2 show the empirical performance of the model selection criteria as a function of number of items I and the number of paired comparisons m_{ij}. For low values of m_{ij} all criteria select the lowest degree polynomial, i.e., $D=1$. When the number of comparisons increases the procedures tend to select larger values of D. In general the performance of the procedures are comparable. However, AIC and Cross validation do seem to (slightly) outperform the other methods. Cross validation is significantly more demanding computationally than AIC and thus as a rule of thumb we recommend the use of the AIC method when no further information is available.

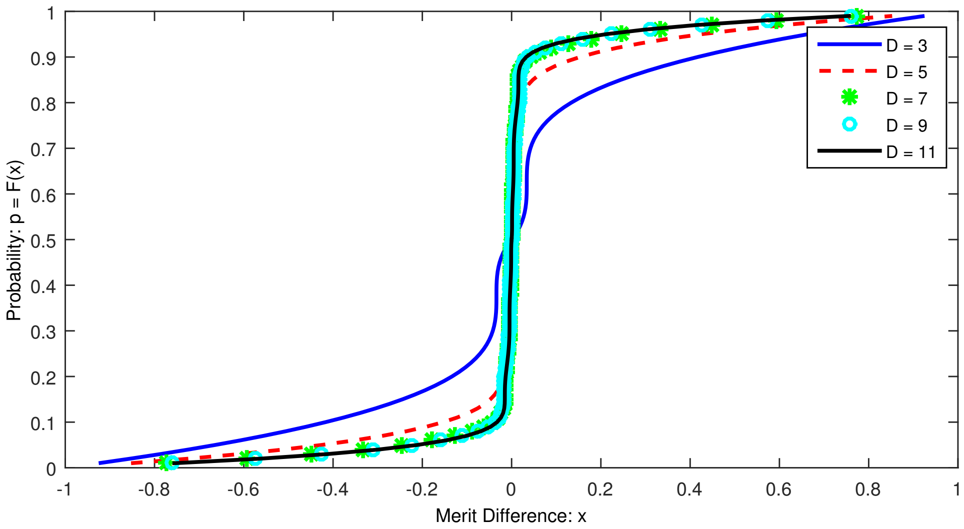

Figure 7: The effect of increasing the dimension D in estimating function F.

5.1. Illustrative Example

In this subsection we illustrate the use of PolyRank on a computer-chess data set[1]. The data set comprises of matches between 186 free single-CPU chess-engines. Each chess-engine played (roughly) 32 matches against 40 opponents. We use PolyRank to estimate model parameters from observed matches that resulted in a victory or defeat; ties are ignored. Figure 7 shows the estimated comparison function for various values of D when for the first 100 chess-engines. As is observed the estimated function $F_{\hat{\boldsymbol{\beta}}}$ seems to stabilize for D greater than or equal to 7. Figure 8 shows the estimated function when the dimension $D = 7$ is fixed and the number of chess-engines I is gradually increased. For I greater than or equal to 60 our estimated function seems to stabilize. Finally, Figure 9 compares the best fit function recovered by PolyRank to the family of BTL models described by $F_{\text{BTL}}(x) = 1/(1 + \exp(-\kappa x))$ for various values of $\kappa > 0$. Somewhat surprisingly, it seems that the family of BTL functions does not provide a good fit.

To illustrate this point consider three players i, j and k such that $p_{ij} = p_{jk} = \alpha > 0.5$ and $p_{ik} = \beta$, then, from (1) we have that $\beta = F(2F^{-1}(\alpha))$. For low values of α (say $\alpha = 0.55$) the BTL model and the polynomila model estimated by PolyRank virtually agree on the value of β (BTL: $\beta \approx 0.6$; Polynomial: $\beta \approx 0.59$); for intermediate values of α (say, $\alpha = 0.7$) the models begin to diverge $\beta \approx 0.84$; Polynomial: $\beta \approx 0.77$) and for large values of α (say $\alpha = 0.9$) this divergence is even more extreme (BTL: $\beta \approx 0.99$; Polynomial: $\beta \approx 0.92$). The estimated polynomial model seems to be more agreeable with the the data at hand since very few chess engines had a (near) perfect win against any opponent. It will be interesting to investigate whether our findings hold for human chess as well.

[1] Publicly available at `http://kirill-kryukov.com/chess/kcec/games.html`.

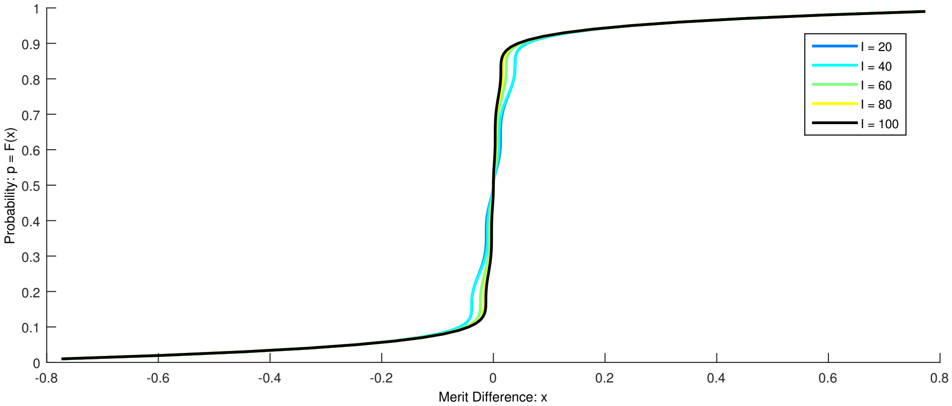

Figure 8: The effect of increasing the amount of data in estimating function F.

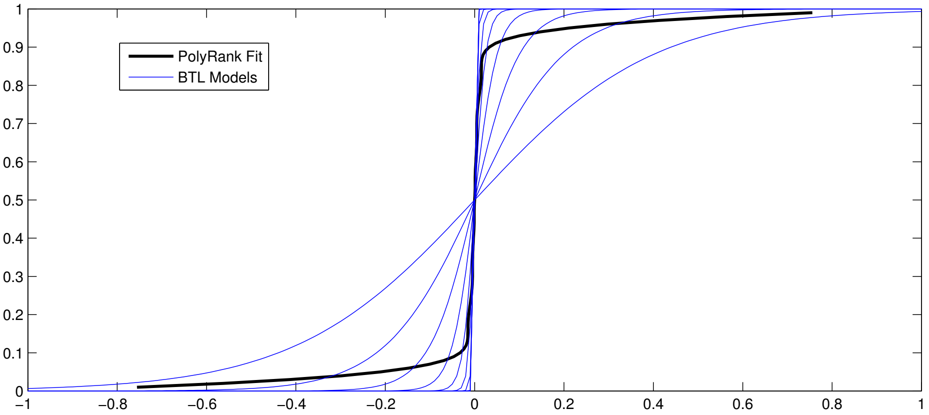

Figure 9: Bradley-Terry-Luce models compared to the best fit comparison function.

6. Summary and Discussion

In this paper we propose a new method for analyzing paired comparison data. Our main contribution is to relax the assumption that the comparison function is known in advance. Instead, we assume that the inverse of the comparison function is a D'th degree polynomial and the comparison function has a bounded support. We show that estimation reduces to a tractable SDP that simultaneously recovers the merit vector and the underlying comparison function F from an initial estimator $\hat{\boldsymbol{P}}$. We refer to this new methodology as PolyRank. We provide non-stochastic as well as stochastic guarantees for our estimators. This includes a thorough sensitivity analysis and additionally convergence and concentration bounds. Our simulation study demonstrates that the method works well in practice. Finally, we investigate a large data set of computer chess matches and provide evidence that the comparison function used for calculating chess ratings for almost nine decades seems to be inadequate, at least for computer chess engines.

Our work shows that PolyRank can be used whenever the existing methods, which assume that the comparison function is known, are used. The only additional requirement is that the comparison graph must have at least $I+D-1$ edges, a condition which is almost always satisfied in practice. Thus, PolyRank provides a flexible and principled alternative to the existing methods for ranking and rating which are based on paired comparisons. Our analysis, however, is just a starting point and many open research problems remain. It is clear that PolyRank can be extended in various directions; these can be grouped into several domains including: (i) modelling issues; (ii) computational/numerical issues; (iii) statistical and inferential problems of varied types.

Modelling. We have assumed that F^{-1} is given by a polynomial. Many other models, in which the polynomial in (4) is replaced by some other set of basis functions, are possible. Monotone splines provide a class of such functions (Ramsay, 1988). One other interesting possibility, with more of a statistical flavor, is to write

$$F^{-1}(p) = \sum_{i=1}^{D} \beta_i K_i(p)$$

where $K_i(p)$ are themselves quantile functions of symmetric random variables. This equation can be viewed as a mixture model on the quantile scale. The family $\{K_i\}$ is then chosen by the investigator; the symmetrized beta family of distributions seems like a suitable family to explore. Another, important issue is the incorporation of covariates, such as time or a "home advantage", as well as many others in the model. This, again, can be done in several ways. The merits can be modeled as regression functions or alternatively one can incorporate the covariates directly into the comparison function. Other issues which deserve attention are the modelling of ties and the comparison of more than two items a time.

Computations and Numerics. Compared with traditional methods, where the comparison function is given, PolyRank has higher computational complexity and may suffer from numerical instability. In part, the numerical issues are related to our decision to model the inverse of the comparison function as a polynomial. This in turn entails that the normal equations are associated with Vandermonde matrices. A known way to circumvent this problem is to use a different basis for solving equation (10), this amounts to choosing a different basis for the polynomial regression, such as Chebyshev Polynomials. Another possibility is the use of a different loss function which is less sensitive to numerical issues. Developing a method to uncouple the estimation of F and of $\boldsymbol{\mu}$ as we provided for the unweighted $\mathcal{L}_2$ norm might provide further insight in this direction. One other, future objective, is to extend the practical reach of PolyRank to larger values of D and I while at the same time increasing computational and numerical efficiency. There may be several ways of doing so. One approach is hand crafting a solver for the SDP at hand. Another possibility is developing an online distributed version of PolyRank in which the function and the merits are updated after each pairwise comparison is observed.

Statistics. The current paper leaves many statistical issues unresolved. For example, we did not provide any results on the asymptotic distributions of our estimated parameters. We believe, however, that normal limits are obtained provided $(\boldsymbol{\mu}, \boldsymbol{\beta})$ are in the interior of the parameter space. It is also clear that employing the one step method we can obtain a fully efficient estimator (Fan and Chen, 1999). Other issues of interest are limit theorems for the case where $I \to \infty$ and when paired comparisons are made adaptively. In the adaptive set up one may exploit the fact that function F can be recovered up to arbitrary precision by using a small subset of the items in order to reduce the overall query complexity of the paired comparison experiment.

Acknowledgments

This paper was written when the first author was a graduate student at the Israel Institute of Technology. The research leading to these results has received funding from the European Research Council under European Union's Horizon 2020 Program, ERC Grant agreement no. 682203 "SpeedInfTradeoff" and the research of Ori Davidov was partially supported by the Israeli Science Foundation Grants No. 1256/13 and 457/17.

Appendix A. Proof of Theorems

The following contains the proofs of our main results.

A.1. Proof of Theorem 1

Proof The constraint $F(0) = 1/2$ is equivalent to $F^{-1}(1/2) = \sum_{i=0}^{D} \beta_i \, (1/2)^i = 0$, which is the last equality constraint in (9). Also, F is increasing iff $F^{-1}(p)$ is increasing and for our set of polynomials this is equivalent to $(F^{-1}(p))' = \beta_1 + 2 \cdot \beta_2 p + ... + D \cdot \beta_D p^{D-1} \geq 0$ for every $p \in [0, 1/2]$. In addition, F is L-Lipschitz continuous and so $|F'(x)| \leq L$; which combined with the monotonicity constraint is equivalent to the constraint $(F^{-1}(p))' \geq 1/L$. By Theorem 6 of (Parrilo, 2016) we have that $(F^{-1}(p))' - 1/L = s(x) + x(1/2 - x)t(x)$ where $s(x)$ and $t(x)$ are sum of squares polynomial functions of degree at most $2d$ and $2d-2$ respectively. Now by Lemma 4 of Parrilo (2016) there exists $\boldsymbol{Q^0} \in \mathbb{S}_+^{d+1}$ and $\boldsymbol{Q^1} \in \mathbb{S}_+^{d}$ such that the coefficients of the polynomials $s(x)$ and $t(x)$ are $s_i = \sum_{j+k=i} Q_{jk}^0$ and $t_i = \sum_{j+k=i} Q_{jk}^1$. By combining these conditions on the polynomial $(F^{-1}(p))' - 1/L$ and the constraint $||\boldsymbol{\beta}||_\infty \leq U$ we obtain the desired result. ■

A.2. Proof of Theorem 3

Proof A little algebra show that

$$||\Delta\hat{\boldsymbol{\mu}} - \Delta\boldsymbol{\mu}||_2^2 = 2n||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||_2^2; \tag{29}$$

this is so because

$$||\Delta\hat{\boldsymbol{\mu}} - \Delta\boldsymbol{\mu}||_2^2 = \sum_{ij} (\hat{\mu}_i - \hat{\mu}_j - \mu_i + \mu_j)^2 = 2n||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||_2^2 - 2(\sum_i \hat{\mu}_i - \sum_i \mu_i)^2$$

where the last term is zero by construction. It follows that

$$||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||_2 = \frac{1}{\sqrt{2n}} ||\Delta\hat{\boldsymbol{\mu}} - \Delta\boldsymbol{\mu}||_2 \leq \frac{1}{\sqrt{2n}} ||F^{-1}(\hat{\boldsymbol{P}}) - \Delta\boldsymbol{\mu}||_2$$
$$= \frac{1}{\sqrt{2n}} ||F^{-1}(\hat{\boldsymbol{P}}) - F^{-1}(\boldsymbol{P})||_2 \leq \frac{4U}{\sqrt{2n}} ||\hat{\boldsymbol{P}} - \boldsymbol{P}||_2.$$

The first inequality is a conscequence of the convex projection theorem and the first equality follows from (29). Equation (13) is derived from

$$||\boldsymbol{P}^* - \boldsymbol{P}||_2^2 = ||F(\Delta\hat{\boldsymbol{\mu}}) - F(\Delta\boldsymbol{\mu})||_2^2 \leq L^2 ||\Delta\hat{\boldsymbol{\mu}} - \Delta\boldsymbol{\mu}||_2^2 = L^2 2n ||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||_2^2 \leq L^2 (4U)^2 ||\hat{\boldsymbol{P}} - \boldsymbol{P}||_2^2.$$

Where the last equality is an application of equation (29) and the last inequality an application of (14). If $\hat{\boldsymbol{P}}$ is assumed to obey strong stochastic transitivity, then $\hat{p}_{ij} < 1/2$ implies that $\hat{p}_{ik} \leq \hat{p}_{jk}$ for all k and $\hat{p}_{ik} < \hat{p}_{jk}$ for at least some k. Thus, the identity $\hat{\mu}_i - \hat{\mu}_j = (1/n) \sum_k (F^{-1}(\hat{p}_{ik}) - F^{-1}(\hat{p}_{jk})) < 0$ together with the fact that F^{-1} is strictly monotone assures that the strong stochastic transitivity order is the same as the order of the estimated merits. ■

Remark Theorem 3, applies to other norms with the proper modifications. Assume that the estimator $\boldsymbol{P}^*$ is obtained via (3) under the norm or semi-norm $||\cdot||_\#$. Let $L_\#$ and $4U_\#$ be the Lipschitz constants of F and F^{-1} associated with $||\cdot||_\#$, then

$$||\boldsymbol{P^*}-\boldsymbol{P}||_\# \leq ||\hat{\boldsymbol{P}}-\boldsymbol{P^*}||_\# + ||\hat{\boldsymbol{P}}-\boldsymbol{P}||_\# \leq L_\#||F^{-1}(\hat{\boldsymbol{P}})-\Delta\hat{\boldsymbol{\mu}}||_\# + ||\hat{\boldsymbol{P}}-\boldsymbol{P}||_\# \leq$$

$$L_\#||F^{-1}(\hat{\boldsymbol{P}})-\Delta\boldsymbol{\mu}||_\# + ||\hat{\boldsymbol{P}}-\boldsymbol{P}||_\# \leq 4L_\#U_\#||\hat{\boldsymbol{P}}-\boldsymbol{P}||_\# + ||\hat{\boldsymbol{P}}-\boldsymbol{P}||_\# = (1+4L_\#U_\#)||\hat{\boldsymbol{P}}-\boldsymbol{P}||_\#,$$

as in (13).

A.3. Proof of Theorem 4

Proof Equation (16) is a conscequence of

$$||\boldsymbol{P^*}-\boldsymbol{P}|| \leq ||\hat{\boldsymbol{P}}-\boldsymbol{P^*}|| + ||\hat{\boldsymbol{P}}-\boldsymbol{P}|| \leq L||F_{\hat{\boldsymbol{\beta}}}^{-1}(\hat{\boldsymbol{P}})-\Delta\hat{\boldsymbol{\mu}}|| + ||\hat{\boldsymbol{P}}-\boldsymbol{P}||$$

$$\leq L||F_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{P}})-\Delta\boldsymbol{\mu}|| + ||\hat{\boldsymbol{P}}-\boldsymbol{P}|| \leq 4LU||\hat{\boldsymbol{P}}-\boldsymbol{P}|| + ||\hat{\boldsymbol{P}}-\boldsymbol{P}||;$$

where the last inequality stems from the fact that $F_{\boldsymbol{\beta}}^{-1}(p)$ is $4U$-Lipschitz continuous for every $F_{\boldsymbol{\beta}}(p) \in \mathcal{F}$ and the previous inequality is a consequence of the optimality of PolyRank. In order to prove (17), consider a set of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$ and a perturbed version $(\boldsymbol{A}+\Delta\boldsymbol{A})(\boldsymbol{x}+\Delta\boldsymbol{x}) = b$ where both $\boldsymbol{A}$ and $\boldsymbol{A}+\Delta\boldsymbol{A}$ are non-singular square matrices. Under these conditions one can show that $\Delta\boldsymbol{x} = \boldsymbol{A}^{-1}\Delta\boldsymbol{A}(\boldsymbol{x}+\Delta\boldsymbol{x})$; thus:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}-\boldsymbol{\beta} \\ \hat{\boldsymbol{\mu}}-\boldsymbol{\mu} \end{pmatrix} = \boldsymbol{A}(\boldsymbol{P})^{-1}[\boldsymbol{A}(\boldsymbol{P^*})-\boldsymbol{A}(\boldsymbol{P})]\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\mu}} \end{pmatrix}. \tag{30}$$

Notice that $||\hat{\boldsymbol{\beta}}||_\infty \leq U$ and also $||\hat{\boldsymbol{\mu}}||_\infty \leq U$. The second claim is true for if $\hat{\mu}_j \geq U$ for some j then $|\hat{\mu}_i - \hat{\mu}_j| = |F_{\hat{\boldsymbol{\beta}}}^{-1}(\hat{p}^*_{ij})| \leq |F_{\hat{\boldsymbol{\beta}}}^{-1}(0)| = |\beta_0| \leq U$ which then implies that $\hat{\mu}_i \geq 0$ for every i and so $\sum_i \hat{\mu}_i \geq U > 0$ which violates the constraint $\sum_i \hat{\mu}_i = 0$; therefore we must have $\hat{\mu}_j < U$ for every j (the analogous argument is valid for $\hat{\mu}_j > -U$). Using the equivalence between norms find:

$$\left|\left|\begin{pmatrix} \hat{\boldsymbol{\beta}}-\boldsymbol{\beta} \\ \hat{\boldsymbol{\mu}}-\boldsymbol{\mu} \end{pmatrix}\right|\right| \leq U\sqrt{I+D}||\boldsymbol{A}(\boldsymbol{P})^{-1}||||\boldsymbol{A}(\boldsymbol{P^*})-\boldsymbol{A}(\boldsymbol{P})||.$$

Now notice that

$$||\boldsymbol{A}(\boldsymbol{P^*})-\boldsymbol{A}(\boldsymbol{P})||^2 = \sum_{k=1}^{D+I-1}\sum_{n=1}^{D}(p^{*n}_{(ij)_k}-p^{n}_{(ij)_k})^2 \leq \sum_{k=1}^{D+I-1} D(p^*_{(ij)_k}-p_{(ij)_k})^2 \leq D||\boldsymbol{P^*}-\boldsymbol{P}||^2$$

and therefore:

$$\left|\left|\begin{pmatrix} \hat{\boldsymbol{\beta}}-\boldsymbol{\beta} \\ \hat{\boldsymbol{\mu}}-\boldsymbol{\mu} \end{pmatrix}\right|\right| \leq U(4LU+1)\sqrt{D(I+D)}||\boldsymbol{A}(\boldsymbol{P})^{-1}||||\hat{\boldsymbol{P}}-\boldsymbol{P}||,$$

which completes the proof of (17). Now we will prove equation (18). A bit of algebra shows that for every $F_{\boldsymbol{\beta}} \in \mathcal{F}$ we have that $\max_{\alpha\in[0,1]}|F_{\boldsymbol{\beta}}^{-1}(\alpha)-F_{\hat{\boldsymbol{\beta}}}^{-1}(\alpha)| \leq 2||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}||_\infty$; combining this with (30), (17) and the lipschitz continuity of $F_{\boldsymbol{\beta}}$ we find the desired result. Finally, to prove order preservation, recognize that problem (9) can be solved by minimizing in $\boldsymbol{\mu}$ and in $\boldsymbol{\beta}$ separately. By minimizing on $\boldsymbol{\mu}$ we find the same closed form solution as the least squares refinement procedure, namely $\mu_i = (1/I)\sum_j F_{\boldsymbol{\beta}}^{-1}(\hat{p}_{ij})$. The proof follows by the same arguments as in Theorem 3. ■

A.4. Proof of Theorem 5

Proof We will first prove that $\max_{\alpha\in[0,1]} |F_{\hat{\boldsymbol{\beta}}}^{-1}(\alpha) - F_{\boldsymbol{\beta}}^{-1}(\alpha)| \geq 2|\beta_D|/8^D$. In the following proof $\mathcal{P}_{D-1}$ is the set of polynomials of degree less than or equal to $D-1$.

$$\max_{\alpha\in[0,1]} |F_{\hat{\boldsymbol{\beta}}}^{-1}(\alpha) - F_{\boldsymbol{\beta}}^{-1}(\alpha)| = \max_{\alpha\in[0,1/2]} |F_{\hat{\boldsymbol{\beta}}}^{-1}(\alpha) - F_{\boldsymbol{\beta}}^{-1}(\alpha)|$$

$$\geq \min_{G\in\mathcal{F}_{D'}} \max_{\alpha\in[0,1/2]} |G(\alpha) - F_{\boldsymbol{\beta}}^{-1}(\alpha)| \geq \min_{G^{-1}\in\mathcal{P}_{D-1}} \max_{\alpha\in[0,1/2]} |G^{-1}(\alpha) - F_{\boldsymbol{\beta}}^{-1}(\alpha)|$$

$$= \min_{F_{\tilde{\boldsymbol{\beta}}}^{-1} \in \mathcal{P}_{D-1}} \max_{\alpha\in[0,1/2]} |F_{\tilde{\boldsymbol{\beta}}}^{-1}(\alpha) + \beta_D \alpha^D|$$

$$= |\beta_D| \min_{F_{\tilde{\boldsymbol{\beta}}}^{-1} \in \mathcal{P}_{D-1}} \max_{\alpha\in[0,1/2]} |F_{\tilde{\boldsymbol{\beta}}}^{-1}(\alpha) + \alpha^D|$$

$$= |\beta_D| \min_{F_{\tilde{\boldsymbol{\beta}}}^{-1} \in \mathcal{P}_{D-1}} \max_{\alpha\in[-1,1]} |F_{\tilde{\boldsymbol{\beta}}}^{-1}((\alpha+1)/4) + ((\alpha+1)/4)^D|$$

$$\geq \frac{|\beta_D|}{4^D} \min_{F_{\tilde{\boldsymbol{\beta}}}^{-1} \in \mathcal{P}_{D-1}} \max_{\alpha\in[-1,1]} |F_{\tilde{\boldsymbol{\beta}}}^{-1}(\alpha) + \alpha^D|$$

$$= \frac{|\beta_D|}{4^D} \frac{1}{2^{D-1}}$$

The last equality is a defining property of Chebyshev polynomials (Mason and Handscomb, 2002). Now, notice that the $4U$-Lipschitz continuity of $F_{\boldsymbol{\beta}}^{-1}$ is equivalent to $|F_{\boldsymbol{\beta}}(y) - F_{\boldsymbol{\beta}}(x)| \geq (1/4U)|y-x|$; combining this with $\max_{\alpha\in[0,1]} |F_{\hat{\boldsymbol{\beta}}}^{-1}(\alpha) - F_{\boldsymbol{\beta}}^{-1}(\alpha)| \geq 2|\beta_D|/8^D$ we obtain equation (19). ■

A.5. Proof of Theorem 6

Proof Let $\hat{l}(D)$ be the empirical loss of a D dimensional fit provided by PolyRank, then:

$$\hat{l}(D) = \min_{\beta\in\mathbb{R}^D, \mu\in\mathbb{R}^I} \sqrt{\sum_{(ij)\in\mathcal{S}} (\beta_0 + ... + \beta_D \hat{p}_{ij}^D - \mu_i + \mu_j)^2}$$

where the minimum is taken over the sets specified by PolyRank for some triplet (D, U, L). It is also true that:

$$\hat{l}(D) = \min_{\beta\in\mathbb{R}^{D+1},\mu\in\mathbb{R}^I} \sqrt{\sum_{(ij)\in\mathcal{S}} (\beta_0 + ... + \beta_D \hat{p}_{ij}^D - \mu_i + \mu_j)^2}$$

for the set specified by the triplet $(D+1, U, L)$, and so

$$\hat{l}(D) \leq \min_{\beta\in\mathbb{R}^{D+1},\mu\in\mathbb{R}^I} \sqrt{\sum_{(ij)\in\mathcal{S}} (\beta_0 + ... + \beta_D \hat{p}_{ij}^D + \beta_{D+1}\hat{p}_{ij}^{D+1} - \mu_i + \mu_j)^2} + \sqrt{\sum_{(ij)\in\mathcal{S}} (\beta_{D+1}\hat{p}_{ij}^{D+1})^2}$$

$$\leq \min_{\beta\in\mathbb{R}^{D+1},\mu\in\mathbb{R}^I} \sqrt{\sum_{(ij)\in\mathcal{S}} (\beta_0 + ... + \beta_D \hat{p}_{ij}^D + \beta_{D+1}\hat{p}_{ij}^{D+1} - \mu_i + \mu_j)^2} + U\sqrt{\sum_{(ij)\in\mathcal{S}} ((1/2)^{D+1})^2}$$

$$\leq \hat{l}(D+1) + U\,(1/2)^{D+1}\sqrt{|\mathcal{S}|}.$$

We have shown that $\hat{l}(D) \leq \hat{l}(D+1) + U\sqrt{|\mathcal{S}|}/2^{D+1}$ which implies that:

$$\hat{l}(D) \leq \hat{l}(D+K) + U\sqrt{|\mathcal{S}|}\sum_{i=D+1}^{D+K}\frac{1}{2^i} = \hat{l}(D+K) + \frac{U\sqrt{|\mathcal{S}|}}{2^D}\sum_{i=1}^{K}\frac{1}{2^i}.$$

Therefore taking the limit of $K\to\infty$ we find that $\hat{l}(D) \leq \hat{l}(\infty) + U\sqrt{|\mathcal{S}|}/2^D$; combining this with the optimality of the estimated parameters we obtain

$$||F_{\hat{\boldsymbol{\beta}}}^{-1}(\hat{\boldsymbol{P}}) - \Delta\hat{\boldsymbol{\mu}}|| \leq ||F^{-1}(\hat{\boldsymbol{P}}) - \Delta\boldsymbol{\mu}|| + \frac{U\sqrt{|\mathcal{S}|}}{2^D}. \tag{31}$$

To complete the proof of (20) notice that

$$||\boldsymbol{P}^* - \boldsymbol{P}|| \leq ||\hat{\boldsymbol{P}} - \boldsymbol{P}|| + ||\hat{\boldsymbol{P}} - \boldsymbol{P}^*|| \leq ||\hat{\boldsymbol{P}} - \boldsymbol{P}|| + L||F_{\hat{\boldsymbol{\beta}}}^{-1}(\hat{\boldsymbol{P}}) - \Delta\hat{\boldsymbol{\mu}}||$$

$$\leq ||\hat{\boldsymbol{P}} - \boldsymbol{P}|| + L||F^{-1}(\hat{\boldsymbol{P}}) - \Delta\boldsymbol{\mu}|| + \frac{LU\sqrt{|\mathcal{S}|}}{2^D} \leq (1+4LU)||\hat{\boldsymbol{P}} - \boldsymbol{P}|| + \frac{1}{2^D}LUI.$$

■

Remark One could equivalently prove that PolyRank defined with the $\mathcal{L}_1$ norm satisfies $||\boldsymbol{P^*} - \boldsymbol{P}||_1 \leq (1+4LU)||\hat{\boldsymbol{P}} - \boldsymbol{P}||_1 + (1/2^D)LUI^2$ for analytic functions with bounded coefficients and with the $\mathcal{L}_\infty$ norm one finds that $||\boldsymbol{P^*} - \boldsymbol{P}||_\infty \leq (1+4LU)||\hat{\boldsymbol{P}} - \boldsymbol{P}||_\infty + (1/2^D)LU$. We provide a sketch of the proof for a generic norm $||\cdot||_\#$. Again, we take $\mathcal{L}_\#$ to be the Lipschitz constant of F associated to $||\cdot||_\#$ and $4U_\#$ the Lipschitz constant of F^{-1} associated to $||\cdot||_\#$, then, we find that

$$||\boldsymbol{P}^* - \boldsymbol{P}||_\# \leq ||\hat{\boldsymbol{P}} - \boldsymbol{P}||_\# + ||\hat{\boldsymbol{P}} - \boldsymbol{P}^*||_\# \leq ||\hat{\boldsymbol{P}} - \boldsymbol{P}||_\# + L_\#||F_{\hat{\boldsymbol{\beta}}}^{-1}(\hat{\boldsymbol{P}}) - \Delta\hat{\boldsymbol{\mu}}||_\#;$$

then an inequality similar to (31) is obtained for the norm $||\cdot||_{\#}$. Norm equivalence guarantees that this can be done up to constant factors. Then, combining the two inequalities one finds that for some $k_1 \geq 1$ and some $k_2 \geq 0$ the following inequality holds:

$$||\boldsymbol{P}^* - \boldsymbol{P}||_{\#} \leq k_1 ||\hat{\boldsymbol{P}} - \boldsymbol{P}||_{\#} + \frac{k_2}{s^D}.$$

This completes the proof.

A.6. Proof of Theorem 7

Proof Equation (21) is a conscequence of the L-Lipschitz continuity of functions in $\mathcal{F}$:

$$||\boldsymbol{P}^* - \boldsymbol{P}|| \leq ||\hat{\boldsymbol{P}} - \boldsymbol{P}|| + ||\hat{\boldsymbol{P}} - \boldsymbol{P}^*|| \leq ||\hat{\boldsymbol{P}} - \boldsymbol{P}|| + L||F^{-1}_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{P}}) - \Delta\hat{\boldsymbol{\mu}}||.$$

■

A.7. Proof of Theorem 8

We will use of the following lemma whose proof is virtually identical to that of Theorem 4:

Lemma 10 *Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ be estimated using* PolyRank. *Then,*

$$\left|\left|\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \end{pmatrix}\right|\right| \leq K||\hat{\boldsymbol{P}} - \boldsymbol{P}||_{\boldsymbol{W}}, \tag{32}$$

where $K \leq U(1 + 4LU)\sqrt{D(I+D)}||\boldsymbol{A}_{\boldsymbol{W}}^{-1}||$ and $\boldsymbol{A}_{\boldsymbol{W}}$ is defined as in 10 with each line multiplied by its respective weight.

Proof Note that:

$$\mathbb{P}\left\{\left|\left|\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \end{pmatrix}\right|\right| \geq \epsilon\right\} = \mathbb{P}\left\{\left|\left|\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \end{pmatrix}\right|\right|^2 \geq \epsilon^2\right\} \leq \mathbb{P}\left\{K^2 \sum w_{ij}|\hat{p}_{ij} - p_{ij}|^2 \geq \epsilon^2\right\}$$

$$\leq \mathbb{P}\left\{|E| \max_{ij}\{w_{ij}|\hat{p}_{ij} - p_{ij}|^2\} \geq \frac{\epsilon^2}{K^2}\right\} \leq \sum_{ij} \mathbb{P}\left\{w_{ij}|\hat{p}_{ij} - p_{ij}|^2 \geq \frac{\epsilon^2}{|E|K^2}\right\}$$

$$= \sum_{ij} \mathbb{P}\left\{|\hat{p}_{ij} - p_{ij}| \geq \frac{\epsilon}{K\sqrt{w_{ij}|E|}}\right\} \leq 2\sum_{ij} \exp\left\{-2m_{ij}\left(\frac{\epsilon}{K\sqrt{w_{ij}|E|}}\right)^2\right\}.$$

Where the last inequality follows from Hoeffding's bound; thus, taking $m_{ij} = nw_{ij}$ we have:

$$= 2\sum_{ij} \exp\left\{-2nw_{ij}\left(\frac{\epsilon}{K\sqrt{w_{ij}|E|}}\right)^2\right\} \leq 2|E| \exp\left\{-2n\frac{\epsilon^2}{K^2|E|}\right\},$$

which completes our proof. ■

References

I. Adler, Y. Cao, R. Karp, E.A. Pekoz, and S.M. Ross. Random knockout tournaments. *Operations Research*, 2017.

N. Ailon. An active learning algorithm for ranking from pairwise preferences with almost optimal query complexity. *Journal of Machine Learning Research*, 2012.

P.D. Allison and N.A. Christakis. Logit models for sets of ranked items. *Sociological Methodology*, 1994.

S.K. Baek, I.G. Yi, H.J. Park, and B.J. Kim. Universal statistics of the knockout tournament. *Nature, Scientific Reports*, 2013.

M. Balinski and R. Laraki. *Majority Judgment, Measuring, Ranking, and Electing.* The MIT Press, 2010.

W.H. Batchelder, N.J. Bershal, and R.S. Simpson. Dynamic paired-comparison scaling. *Journal of Mathematical Psychology*, 1992.

R.A. Bradley and M.E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.

S. Chatterjee and S. Mukherjee. On estimation in tournaments and graphs under monotonicity constraints. *arXiv:1603.04556 [math.ST]*, 2016.

F.R.K. Chung. Spectral graph theory. *AMS and CBMS*, 1994.

F.R.K. Chung and F.K. Hwang. Do stronger players win more knockout tournaments? *Journal of the American Statistical Association*, 1978.

G. Claeskens and N.L. Hjort. Model selection and model averaging. *Cambridge Series in Statistical and Probabilistic Mathematics*, 2006.

W.N. Colley. Colley's bias free college football ranking method: The colley matrix explained, 2002. URL `http://www.colleyrankings.com/matrate.pdf`.

P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. *Proceedings of the fourth ACM conference on recommender systems*, 2010.

H.A. David. *The method of paired comparisons.* Hodder Arnold, 1988.

J.S. deCani. Maximum likelihood paired comparison ranking by linear programming. *Biometrika*, page 537, 1969.

J. Fan and J. Chen. One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society*, 1999.

P. Favardin, D. Lepelley, and J. Serais. Borda rule, copeland method and strategic manipulation. *Rev.Econ.Design*, 2002.

A.Y. Govan. Ranking theory with application to popular sports, 2008.

R. Heckel, N.B. Shah, K. Ramchandran, and M.J. Wainwright. Active ranking from pairwise comparisons and when parametric assumptions don't help. *arxiv:1606.08842v2 [cs.LG]*, 2016.

R. Herbrich, T. Minka, and T. Graepel. Trueskill tm: A bayesian skill rating system. *Advances in Neural Information Processing Systems, MIT Press*, 2007.

P.J. Huber. Pairwise comparison and ranking: optimum properties of the row sum procedure. *Annals of Mathematical Statistics*, 1963.

S.H. Hwang. Contest success functions: Theory and evidence. *Economics Department Working Paper Series. 11.*, 2009. URL `http://scholarworks.umass.edu/econ_workingpaper/11`.

R.B. Israel. Stronger players need not win more knockout tournaments. *Journal of the American Statistical Association*, 1981.

H. Jia, S. Skaperdas, and S. Vaidya. Contest functions: Theoretical foundations and issues in estimation. *International Journal of Industrial Organization*, 2013.

X. Jiang, L.H. Lim, Y. Yao, and Y. Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 2010.

J. Levin and B. Nalebuff. An introduction to vote-counting schemes. *Journal of Economic Perspectives*, 1995.

J.I. Marden. *Analyzing and Modeling Rank Data*. Chapman and Hall/CRC, 1996.

J.C. Mason and D.C. Handscomb. *Chebyshev Polynomials*. CRC Press, 2002.

K. Massey. Massey ratings, 2017. URL `https://www.masseyratings.com/theory/massey97.pdf`.

H.W. Morrison. Testable conditions for triads of paired comparison choices. *Psychometrika*, 1963.

A. Mutapcic and S. Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods & Software*, 2009.

A.S. Nemirovski and M.J. Todd. Interior-point methods for optimization. *Acta Numerica*, 2009.

I.F.D. Oliveira, S. Zehavi, and O. Davidov. Stochastic transitivity: Axioms and models. *Journal of Mathematical Psychology*, 2018.

E. Pacuit. Voting methods. *In The Stanford Encyclopedia of Philosophy*, 2012.

V. Pan. How bad are vandermonde matrices? *SIAM Journal on Matrix Analysis and Applications*, 2015.

P.A. Parrilo. Algebraic techniques and semidefinite optimization, 2016. URL `http://stellar.mit.edu/S/course/6/sp10/6.256/courseMaterial/topics/topic2/lectureNotes/lecture-10/lecture-10.pdf`.

A. Rajkumar and S. Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. *Proceedings of JMLR*, 2014.

J.O. Ramsay. Monotone regression splines in action. *Statistical Science*, 1988.

M. Regenwetter and C.P. Davis-Stober. There are many models of transitive preference: a tutorial review and current perspective. *Decision Modeling and Behavior in Complex and Uncertain Environments*, 2008.

M. Regenwetter, J. Dana, and C.P. Davis-Stober. Transitivity of preferences. *Psychological Review*, 2011.

D.G. Saari. Basic geometry of voting. *Springer-Verlag Berlin Heidelberg*, 1995.

N.B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M.J. Wainwright. Estimation from pairwise comparisons: sharp minimax bounds with topology dependence. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015a.

N.B. Shah, S. Balakrishnan, and A. Guntuboyina. Stochastically transitive models for pairwise comparisons: statistical and computational issues. *arXiv:1510.05610 [stat.ML]*, 2015b.

G. Simons and Y.C. Yao. Asymptotics when the number of parameters tends to infinity in the bradley-terry model for paired comparisons. *The Annals of Statistics*, 1999.

O. Stein. How to solve a semi-infinite optimization problem. *European Journal of Operational Research*, 2012.

S. Su. Fleximble parametric quantile regression model. *Stat Compt*, 2015.

I. Takeuchi, Q.V. Le, T.D. Sears, and A.J. Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 2006.

L.L. Thurstone. A law of comparative judgment. *Psychology Review*, 1927.

K. Tsukida and M.R. Gupta. How to analyze paired comparison data, 2011. URL `http://www.dtic.mil/dtic/tr/fulltext/u2/a543806.pdf`.

L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 1996.

J.I. Yellott. The relationship between thurstone's, luce's and dawkins' models for paired comparisons. *Annual Meeting of Mathematical Psychology*, 1970.

J.I. Yellott. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential. *Journal of Mathematical Psychology*, 1977.

E. Zermelo. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 1928.