Stochastically Transitive Models for Pairwise Comparisons: Statistical and Computational Issues

Nihar B. Shah[†] NIHAR@EECS.BERKELEY.EDU
Sivaraman Balakrishnan[♯] SIVA@STAT.CMU.EDU
Adityanand Guntuboyina[*] ADITYA@STAT.BERKELEY.EDU
Martin J. Wainwright[†*] WAINWRIG@BERKELEY.EDU

[†]Dept. of EECS, [*]Dept. of Statistics, University of California, Berkeley
[♯] Dept. of Statistics, Carnegie Mellon University

Abstract

There are various parametric models for analyzing pairwise comparison data, including the Bradley-Terry-Luce (BTL) and Thurstone models, but their reliance on strong parametric assumptions is limiting. In this work, we study a flexible model for pairwise comparisons, under which the probabilities of outcomes are required only to satisfy a natural form of stochastic transitivity. This class includes parametric models including the BTL and Thurstone models as special cases, but is considerably more general. We provide various examples of models in this broader stochastically transitive class for which classical parametric models provide poor fits. Despite this greater flexibility, we show that the matrix of probabilities can be estimated at the same rate as in standard parametric models. On the other hand, unlike in the BTL and Thurstone models, computing the minimax-optimal estimator in the stochastically transitive model is non-trivial, and we explore various computationally tractable alternatives. We show that a simple singular value thresholding algorithm is statistically consistent but does not achieve the minimax rate. We then propose and study algorithms that achieve the minimax rate over interesting sub-classes of the full stochastically transitive class. We complement our theoretical results with thorough numerical simulations.

1. Introduction

Pairwise comparison data is ubiquitous and arises naturally in a variety of applications, including tournament play, voting, online search rankings, and ad placement problems. In rough terms, given a set of n objects, and a collection of possibly inconsistent comparisons between pairs of these objects, the goal is to aggregate these comparisons in order to perform effective statistical inference on various underlying properties of the population. A particular property of interest is the underlying pairwise comparison probabilities—that is, the probability that object i is preferred to object j in a pairwise comparison. The Bradley-Terry-Luce (Bradley & Terry, 1952; Luce, 1959) and Thurstone (Thurstone, 1927) models are mainstays in analyzing this type of pairwise comparison data. These models are parametric in nature: more specifically, they assume the existence of an n-dimensional weight vector that measures the quality or strength of each item. The pairwise comparison probabilities are then determined via some fixed (parametric) function of the qualities of the pair of objects. Estimation in these models reduces to estimating the underlying weight vector, and a large body of prior work has focused on these models (see, e.g., Negahban et al. 2012; Hajek et al. 2014; Shah et al. 2016a). These models however, enforce strong relationships on the pairwise comparison probabilities that often fail to hold in real applications. Various papers (Davidson & Marschak, 1959; McLaughlin & Luce, 1965; Tversky, 1972; Ballinger & Wilcox, 1997) have provided experimental results in which these parametric modeling assumptions fail to hold.

Our focus in this paper is on models with roots in social science and psychology (e.g., see Fishburn 1973 for an overview), where the only coherence assumption made on the pairwise comparison probabilities is that of *strong stochastic transitivity*, or SST for short. These models include parametric models as special cases but are considerably more general. The SST model is validated by several empirical analyses in a long line of work (Davidson & Marschak, 1959; McLaughlin & Luce, 1965; Tversky, 1972; Ballinger & Wilcox, 1997). The conclusion of Ballinger & Wilcox (1997) is especially strongly worded:

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

All of these parametric c.d.f.s are soundly rejected by our data. However, SST usually survives scrutiny.

We are thus provided with strong empirical motivation for studying the fundamental properties of pairwise comparison probabilities satisfying SST.

In this paper, we focus on the problem of estimating the matrix of pairwise comparison probabilities—that is, the probability that an item i will beat a second item j in any given comparison. Estimates of these comparison probabilities are useful in various applications. For instance, when the items correspond to players or teams in a sport, the predicted odds of one team beating the other are central to betting and bookmaking operations. In a supermarket or an ad display, an accurate estimate of the probability of a customer preferring one item over another, along with the respective profits for each item, can effectively guide the choice of which product to display. Accurate estimates of the pairwise comparison probabilities can also be used to infer partial or full rankings of the underlying items.

Our contributions: We begin by studying the performance of optimal methods for estimating matrices in the SST class: our first main result (Theorem 1) characterizes the minimax rate in squared Frobenius norm up to logarithmic factors. This result reveals that even though the SST class of matrices is considerably larger than the classical parametric class, surprisingly, it is possible to estimate any SST matrix at the same rate as the classical parametric family. On the other hand, computing this optimal estimator over the SST class is non-trivial, as a brute force approach entails an exhaustive search over permutations. Accordingly, we turn to studying computationally tractable estimation procedures. Our second main result (Theorem 2) applies to a polynomial-time estimator based on thresholding the singular values of the data matrix. We sharpen and generalize a previous analysis of Chatterjee (2014), and give a tight characterization of the rate achieved by both hard and soft thresholding estimators. Our third contribution, formalized in Theorems 3 and 4, is to show that for certain interesting subsets of the full SST class, a combination of parametric maximum likelihood (Shah et al., 2016a) and noisy sorting algorithms (Braverman & Mossel, 2008) leads to a tractable two-stage method that achieves the minimax rate. Our fourth contribution is to supplement our minimax lower bound with lower bounds for various known estimators. These lower bounds show that none of these tractable estimators achieve the minimax rate uniformly over the entire class. The lower bounds also show that the minimax rates for any of these subclasses is no better than the full SST class.

Related work: The literature on ranking and estimation from pairwise comparisons is vast and we refer the reader to various surveys (Fligner & Verducci, 1993; Marden, 1996; Cattelan, 2012) for a more detailed overview. We focus our literature review on papers that are closely related to the contributions of our work. Some recent work (Negahban et al., 2012; Hajek et al., 2014; Shah et al., 2016a) studies procedures and minimax rates for estimating the latent quality vector that underlie such parametric models. Theorem 4 in the present paper provides an extension of these results, in particular by showing that an optimal estimate of the latent quality vector can be used to construct an optimal estimate of the pairwise comparison probabilities. Chatterjee (2014) formally introduced the estimation problem considered in this paper, and analyzed an estimator based on singular value thresholding. We provide a sharper analysis of this estimator, and show that our upper bound is—in fact—unimprovable.

Various papers (Kenyon-Mathieu & Schudy, 2007; Braverman & Mossel, 2008) consider the noisy sorting problem, in which the goal is to infer the underlying order under the assumption that each pairwise comparison has a probability of agreeing with the underlying order that is bounded away from $\frac{1}{2}$ by a fixed constant. These works provide polynomial-time algorithms to infer the true underlying order with a certain accuracy. Part of our analysis leverages an algorithm due to Braverman & Mossel (2008); in particular, we extend their analysis to provide guarantees for estimating pairwise comparison probabilities as opposed to estimating the underlying order.

As will be clarified in the sequel, the assumption of strong stochastic transitivity has close connections to the statistical literature on shape constrained inference (e.g., Silvapulle & Sen 2011), particularly to the problem of bivariate isotonic regression. Some of our analysis leverages metric entropy bounds from past work in this area (e.g., Gao & Wellner 2007; Chatterjee et al. 2015).

While in the present paper we establish guarantees on the recovery of the pairwise-comparison probabilities, in a companion paper (Shah & Wainwright, 2015), we study the problem of identifying the top subset or estimating the total ordering of the items based on noisy pairwise comparisons.

2. Background and problem formulation

Given a collection of $n \geq 2$ items, suppose that we have access to noisy comparisons between any pair $i \neq j$ of distinct items. The full set of all possible pairwise comparisons can be described by a probability matrix $M^* \in [0,1]^{n \times n}$, in which M^*_{ij} is the probability that item i is preferred to item j. The upper and lower halves of the probability matrix M^* are related by the shifted-skew-symmetry condition $M^*_{ji} = 1 - M^*_{ij}$ for all $i, j \in [n]$.

2.1. Estimation of pairwise comparison probabilities

For any matrix $M^* \in [0,1]^{n\times n}$ with $M^*_{ij} = 1 - M^*_{ji}$ for every (i,j), suppose that we observe a random matrix $Y \in \{0,1\}^{n\times n}$ with (upper-triangular) independent Bernoulli entries, in particular, with $\mathbb{P}[Y_{ij} = 1] = M^*_{ij}$ for every $1 \le i \le j \le n$ and $Y_{ji} = 1 - Y_{ij}$. Based on observing Y, our goal in this paper is to recover an accurate estimate, in the squared Frobenius norm, of the matrix M^*.

Our primary focus in this paper will be on the setting where for n items we observe the outcome of a single pairwise comparison for each pair. We will subsequently (in Section 3.5) also address the more general case when we have partial observations, that is, when each pairwise comparison is observed with a fixed probability.

2.2. Strong stochastic transitivity

Beyond the previously mentioned constraints on the matrix M^*—namely that $M^*_{ij} \in [0,1]$ and that $M^*_{ij} = 1 - M^*_{ij}$—more structured and interesting models are obtained by imposing further restrictions on the entries of M^*. We now turn to one such condition, known as *strong stochastic transitivity* (SST), which reflects the natural transitivity of any complete ordering. Formally, suppose that the full collection of items $[n]$ is endowed with a complete ordering π^*. We use the notation $\pi^*(i) < \pi^*(j)$ to convey that item i is preferred to item j in the total ordering π^*. Consider some triple (i,j,k) such that $\pi^*(i) < \pi^*(j)$. A matrix M^* satisfies the SST condition if the inequality $M^*_{ik} \ge M^*_{jk}$ holds for every such triple. The intuition underlying this constraint is as follows: since i dominates j in the true underlying order, when we make noisy comparisons, the probability that i is preferred to k should be at least as large as the probability that j is preferred to k. The SST condition was first described in the psychology literature (e.g., Fishburn 1973; Davidson & Marschak 1959).

The SST condition is characterized by the existence of a permutation such that the permuted matrix has entries that increase across rows and decrease down columns. More precisely, for a given permutation π^*, let us say that a matrix M is π^*-faithful if for every pair (i,j) such that $\pi^*(i) < \pi^*(j)$, we have $M_{ik} \ge M_{jk}$ for all $k \in [n]$. With this notion, the set of SST matrices is given by

$$\mathbb{C}_{\text{SST}} = \big\{M \in [0,1]^{n\times n} \mid M_{ba} = 1 - M_{ab} \;\forall\, (a,b), \text{ and } \exists \text{ perm. } \pi^* \text{ s.t. } M \text{ is } \pi^*\text{-faithful}\big\}. \quad (1)$$

Note that the stated inequalities also guarantee that for any pair with $\pi^*(i) < \pi^*(j)$, we have $M_{ki} \le M_{kj}$ for all k, which corresponds to a form of column ordering. The class $\mathbb{C}_{\text{SST}}$ will be our primary focus in this paper.

2.3. Classical parametric models

Let us now describe a family of classical parametric models, one which includes Bradley-Terry-Luce and Thurstone (Case V) models (Bradley & Terry, 1952; Luce, 1959; Thurstone, 1927). As we will see, this family forms a relatively small subclass of the SST matrices $\mathbb{C}_{\text{SST}}$. In more detail, parametric models are described by a weight vector $w^* \in \mathbb{R}^n$ that corresponds to the notional qualities of the n items. Moreover, consider any non-decreasing function $F : \mathbb{R} \mapsto [0,1]$ such that $F(t) = 1 - F(-t)$ for all $t \in \mathbb{R}$; we refer to any such function F as being valid. Any such pair (F, w^*) induces a particular pairwise comparison model in which

$$M^*_{ij} = F(w^*_i - w^*_j) \qquad \text{for all pairs } (i,j). \quad (2)$$

For each valid choice of F, we define

$$\mathbb{C}_{\text{PAR}}(F) = \{M \in [0,1]^{n\times n} \text{ induced by (2) for some } w^* \in \mathbb{R}^n\}. \quad (3)$$

For any choice of F, it is easy to verify that $\mathbb{C}_{\text{PAR}}(F)$ is a subset of $\mathbb{C}_{\text{SST}}$, meaning that any matrix M induced by the relation (2) satisfies all the constraints defining the set $\mathbb{C}_{\text{SST}}$. As particular important examples, we recover the Thurstone model by setting $F(t) = \Phi(t)$ where Φ is the Gaussian CDF, and the Bradley-Terry-Luce model by setting $F(t) = \frac{e^t}{1+e^t}$, corresponding to the sigmoid function.

Remark: One can impose further constraints on the vector w^* without reducing the size of the class $\{\mathbb{C}_{\text{PAR}}(F), \text{ for some valid } F\}$. In particular, since the pairwise probabilities depend only on the differences $w^*_i - w^*_j$, we can assume without loss of generality that $\langle w^*, 1\rangle = 0$. Moreover, since the choice of F can include rescaling its argument, we can also assume that $\|w^*\|_\infty \le 1$. Accordingly, we assume in our subsequent analysis that w^* belongs to the set $\big\{w \in \mathbb{R}^n \mid \text{such that } \langle w, 1\rangle = 0 \text{ and } \|w\|_\infty \le 1.\big\}$.

2.4. Inadequacies of parametric models

As noted in the introduction, a large body of past work (e.g., Davidson & Marschak 1959; McLaughlin & Luce 1965; Tversky 1972; Ballinger & Wilcox 1997) has shown that parametric models, of the form (3) for some choice of F, often provide poor fits to real-world data. What might be a reason for this phenomenon? Roughly, parametric models impose the very restrictive assumption that the choice of an item depends on the value of a single latent factor (as indexed by w^*)—e.g., that our preference for cars depends only on their fuel economy, or that the probability that one hockey team beats another depends only on the skills of the goalkeepers. This intuition can be formalized

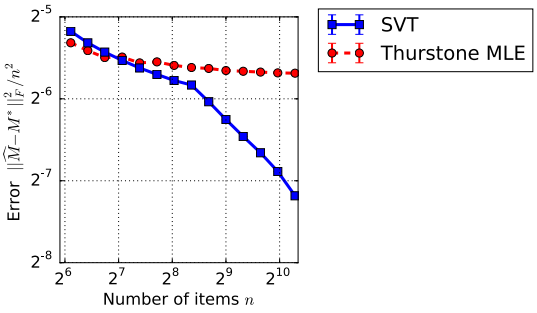


Figure 1: Estimation for a class of SST matrices that are far from parametric models. The parametric model (Thurstone MLE) yields a poor fit, whereas the singular value thresholding (SVT) estimator, which allows for estimates over the full SST class, leads to consistent estimation.

to construct matrices $M^* \in \mathbb{C}_{\text{SST}}$ that are far away from *every valid parametric approximation* as summarized in the following result:

Proposition 1. *For every $n \geq 4$, there exist matrices $M^* \in \mathbb{C}_{SST}$ such that*

$$\inf_{\text{valid } F} \inf_{M \in \mathbb{C}_{PAR}(F)} \frac{1}{n^2} |||M - M^*|||_F^2 \geq c, \tag{4}$$

for some universal constant $c > 0$.

Given that every entry of matrices in $\mathbb{C}_{\text{SST}}$ lies in the interval $[0, 1]$, the Frobenius norm diameter of the class $\mathbb{C}_{\text{SST}}$ is at most n^2, so the scaling of the lower bound (4) is sharp.

What sort of matrices M^* are "bad" in the sense of satisfying a lower bound of the form (4)? To provide some intuition, let us return to the analogy of rating cars. A key property of any parametric model is that if we prefer car 1 to car 2 more than we prefer car 3 to car 4, then we must also prefer car 1 to car 3 more than we prefer car 2 to car 4. (This condition follows from the proof of Proposition 1.) This condition is potentially satisfied if there is a single determining factor across all cars, for instance, the fuel economy. In more generality, in any situation where a single (latent) parameter per item does not adequately explain our preferences, one can expect that the class of parametric models to provide a poor fit to the pairwise preference probabilities.

The lower bound of Proposition 1 means that any parametric estimator of the matrix M^* should perform poorly. This expectation is confirmed by the simulation results in Figure 1. After generating observations from a "bad matrix" over a range of n, we fit the data set using either the maximum likelihood estimate in the Thurstone model, or the singular value thresholding (SVT) estimator to be discussed in Section 3.2. For each estimator $\widehat{M}$, we plot the rescaled Frobenius norm error $\frac{|||\widehat{M}-M^*|||_{\text{F}}^2}{n^2}$ versus the sample size. Consistent with the lower bound (4), the error in the Thurstone-based estimator stays bounded above a universal constant. In contrast, the SVT error goes to zero with n, and as our theory in the sequel shows, the rate at which the error decays is at least as fast as $1/\sqrt{n}$.

3. Main results

Thus far, we have introduced two classes of models for matrices of pairwise comparison probabilities. Our main results provide characterizations of the estimation error associated with different subsets of these classes, using either optimal estimators (that we suspect are not polynomial-time computable), or more computationally efficient estimators that can be computed in polynomial-time. Throughout the section, we let c_u, c_ℓ, c, c_0 denote positive constants whose values do not depend on n. All proofs are available in the appendix.

3.1. Characterization of the minimax risk

We begin by providing a result that characterizes the minimax risk in squared Frobenius norm over the SST class.

Theorem 1. *The minimax risk of estimating $M^* \in \mathbb{C}_{SST}$ is bounded as*

$$\frac{c_\ell}{n} \leq \inf_{\widetilde{M}} \sup_{M^* \in \mathbb{C}_{SST}} \frac{1}{n^2} \mathbb{E}[|||\widetilde{M} - M^*|||_F^2] \leq c_u \frac{\log^2 n}{n},$$

where the infimum ranges over all measurable functions $\widetilde{M}$ of the observed matrix Y.

The proof of the lower bound is based on extracting a particular subset of $\mathbb{C}_{\text{SST}}$ such that any matrix in this subset has at least n positions that are unconstrained, apart from having to belong to the interval $[\frac{1}{2}, 1]$. We can thus conclude that estimation of the full matrix is at least as hard as estimating n numbers belonging to the interval $[\frac{1}{2}, 1]$ based on a single observation per number, and this leads to an $\Omega(n^{-1})$ lower bound, as stated.

Proving an upper bound requires substantially more effort. In particular, we establish it via careful analysis of the constrained least-squares estimator

$$\widehat{M} \in \arg\min_{M \in \mathbb{C}_{\text{SST}}} |||Y - M|||_{\text{F}}^2. \tag{5a}$$

In particular, we prove that there are universal constants (c_0, c_1, c_2) such that, for any matrix $M^* \in \mathbb{C}_{\text{SST}}$, this estimator satisfies the high probability bound

$$\mathbb{P}\Big[\frac{1}{n^2} |||\widehat{M} - M^*|||_{\text{F}}^2 \geq c_0 \frac{\log^2(n)}{n}\Big] \leq c_1 e^{-c_2 n}. \tag{5b}$$

Since the entries of $\widehat{M}$ and M^* all lie in the interval $[0,1]$, this tail bound implies the upper bound on the expected mean-squared error stated in Theorem 1. Proving the high probability bound (5b) requires sharp control on a quantity known as the localized Gaussian complexity of the class $\mathbb{C}_{\text{SST}}$. We use Dudley's entropy integral in order to derive an upper bound that is sharp up to a logarithmic factor; doing so in turn requires deriving upper bounds on the metric entropy of the class $\mathbb{C}_{\text{SST}}$ for which we leverage the prior work of Gao & Wellner (2007).

We do not know whether the constrained least-squares estimator (5a) is computable in time polynomial in n, but we suspect not. This complexity is a consequence of the fact that the set $\mathbb{C}_{\text{SST}}$ is not convex, but is a union of $n!$ convex sets. Given this issue, it becomes interesting to consider the performance of alternative estimators that can be computed in polynomial-time.

3.2. Sharp analysis of singular value thresholding (SVT) estimator

The first polynomial-time estimator that we consider is a simple estimator based on thresholding singular values of the observed matrix Y, and reconstructing its truncated SVD. For the full class $\mathbb{C}_{\text{SST}}$, Chatterjee (2014) analyzed the performance of such an estimator and proved that the squared Frobenius error decays as $\mathcal{O}(n^{-\frac{1}{4}})$ uniformly over $\mathbb{C}_{\text{SST}}$. We prove that its error decays as $\mathcal{O}(n^{-\frac{1}{2}})$, again uniformly over $\mathbb{C}_{\text{SST}}$, and moreover, that this upper bound is unimprovable.

Let us begin by describing the estimator. Given the observation matrix $Y \in \mathbb{R}^{n\times n}$, we can write its singular value decomposition as $Y = UDV^T$, where the $(n \times n)$ matrix D is diagonal, whereas the $(n \times n)$ matrices U and V are orthonormal. Given a threshold level $\lambda_n > 0$, the soft-thresholded version of a diagonal matrix D is the diagonal matrix $T_{\lambda_n}(D)$ with values

$$[T_{\lambda_n}(D)]_{jj} = \max\{0, D_{jj} - \lambda_n\} \quad \text{for every } j \in [1, n].$$

With this notation, the soft singular-value-thresholded (soft-SVT) version of Y is given by $T_{\lambda_n}(Y) = UT_{\lambda_n}(D)V^T$. The following theorem provides a bound on its squared Frobenius error:

Theorem 2. *For any $M^* \in \mathbb{C}_{\text{SST}}$, the soft-SVT estimator $\widehat{M}_{\lambda_n} = T_{\lambda_n}(Y)$ with $\lambda_n = 2.1\sqrt{n}$ satisfies the bound*

$$\mathbb{P}\Big[\frac{1}{n^2}\|\widehat{M}_{\lambda_n} - M^*\|_F^2 \geq \frac{c_u}{\sqrt{n}}\Big] \leq c_0 e^{-cn}.$$

A few comments on this result are in order. Since the matrices $\widehat{M}_{\lambda_n}$ and M^* have entries in the unit interval $[0,1]$, the normalized squared error $\frac{1}{n^2}\|\widehat{M}_{\lambda_n} - M^*\|_{\text{F}}^2$ is at most 1. Consequently, the tail bound of Theorem 2 guarantees that $\sup_{M^*\in\mathbb{C}_{\text{SST}}} \mathbb{E}[\frac{1}{n^2}\|\widehat{M}_{\lambda_n} - M^*\|_{\text{F}}^2] \leq \frac{c'_u}{\sqrt{n}}$. In an extended version of this paper (Shah et al., 2015), we prove a matching lower bound showing that this upper bound on the error of the SVT estimator is sharp up to constant factors.

To be clear, Chatterjee (2014) analyzed the hard-SVT estimator, which is based on the hard-thresholding operator

$$[H_{\lambda_n}(D)]_{jj} = D_{jj}\, \mathbf{1}\{D_{jj} \geq \lambda_n\}.$$

Here $\mathbf{1}\{\cdot\}$ denotes the 0-1-valued indicator function. In this setting, the hard-SVT estimator is simply, $H_{\lambda_n}(Y) = UH_{\lambda_n}(D)V^T$. With essentially the same choice of λ_n as above, Chatterjee showed that the estimate $H_{\lambda_n}(Y)$ has a mean-squared error of $\mathcal{O}(n^{-1/4})$. One can verify that the proof of Theorem 2 in our paper goes through for the hard-SVT estimator as well. Consequently the performance of the hard-SVT estimator is of the order $\Theta(n^{-1/2})$, and is identical to that of the soft-thresholded version up to universal constants.

Note that the hard and soft-SVT estimators return matrices that may not lie in the SST class $\mathbb{C}_{\text{SST}}$. In a companion paper (Shah et al., 2016b), we provide an alternative computationally-efficient estimator with similar statistical guarantees that is guaranteed to return a matrix in the SST class.

The result of Theorem 2 provides a sharp characterization of the behavior of the soft/hard SVT estimators. On the positive side, these are easily implementable estimators that achieve a mean-squared error bounded by $\mathcal{O}(1/\sqrt{n})$ uniformly over the entire class $\mathbb{C}_{\text{SST}}$. On the negative side, this rate is slower than the $\mathcal{O}(\log^2 n/n)$ rate achieved by the least-squares estimator, as in Theorem 1.

In conjunction, Theorems 1 and 2 raise a natural open question: is there a polynomial-time estimator that achieves the minimax rate uniformly over the family $\mathbb{C}_{\text{SST}}$? We do not know the answer to this question, but the following subsections provide some partial answers by analyzing some polynomial-time estimators that (up to logarithmic factors) achieve the optimal $\tilde{\mathcal{O}}(1/n)$-rate over interesting sub-classes of $\mathbb{C}_{\text{SST}}$. In the next two sections, we turn to results of this type.

3.3. Optimal rates for high SNR subclass

In this section, we describe a multi-step polynomial-time estimator that (up to logarithmic factors) can achieve the optimal $\tilde{\mathcal{O}}(1/n)$ rate over an interesting subclass of $\mathbb{C}_{\text{SST}}$. This subset corresponds to matrices M that have a relatively high signal-to-noise ratio (SNR), meaning that no entries of M fall within a certain window of $1/2$. More formally, for some $\gamma \in (0, \frac{1}{2})$, we define the class

$$\mathbb{C}_{\text{HIGH}}(\gamma) = \{M \in \mathbb{C}_{\text{SST}} \,|\, \max(M_{ij}, M_{ji}) \geq 1/2 + \gamma \;\forall\, i \neq j\}.$$

In terms of estimation difficulty, this SNR restriction does not make the problem substantially easier: as the following theorem shows, the minimax mean-squared error remains lower bounded by a constant multiple of $1/n$. Moreover, we demonstrate a polynomial-time algorithm that achieves this optimal mean squared error up to logarithmic factors.

The following theorem applies to any fixed $\gamma \in (0, \frac{1}{2}]$ independent of n, and involves constants (c_ℓ, c_u, c) that may depend on γ but are independent of n.

Theorem 3. *Any estimator $\widetilde{M}$ incurs an error lower bounded as*

$$\sup_{M^* \in \mathbb{C}_{HIGH}(\gamma)} \frac{1}{n^2} \mathbb{E}\big[|||\widetilde{M} - M^*|||_F^2\big] \geq \frac{c_\ell}{n}. \tag{6a}$$

Moreover, there is an estimator $\widehat{M}$, computable in polynomial-time, such that for any $M^* \in \mathbb{C}_{HIGH}(\gamma)$

$$\mathbb{P}\Big[\frac{1}{n^2}|||\widehat{M} - M^*|||_F^2 \geq \frac{c_u \log^2 n}{n}\Big] \leq \frac{c}{n^2}. \tag{6b}$$

As with our proof of the lower bound in Theorem 1, we prove the lower bound by considering the sub-class of matrices that are free only on the two diagonals just above and below the main diagonal. We now provide a brief sketch for the proof of the upper bound (6b) which is based on analyzing the following two-step estimator:

Step 1: We first find a permutation $\widehat{\pi}_{\text{FAS}}$ of the n items that minimizes the total number of disagreements with the observations. (For a given ordering, we say that any pair of items (i, j) are in disagreement with the observation if either i is rated higher than j in the ordering and $Y_{ij} = 0$, or if i is rated lower than j in the ordering and $Y_{ij} = 1$.) The problem of finding such a disagreement-minimizing permutation $\widehat{\pi}_{\text{FAS}}$ is commonly known as the minimum feedback arc set (FAS) problem. It is known to be NP-hard in the worst-case (Ailon et al., 2008; Alon, 2006), but our set-up has additional probabilistic structure that allows for polynomial-time solutions with high probability. In particular, we call upon a polynomial-time algorithm due to Braverman & Mossel (2008) that, under the model $\mathbb{C}_{\text{SST}}(\gamma)$, is guaranteed to find the exact solution to the FAS problem with high probability. Viewing the FAS permutation $\widehat{\pi}_{\text{FAS}}$ as an approximation to the true permutation π^*, the novel technical work in this first step is show that $\widehat{\pi}_{\text{FAS}}$ is "good enough" for Frobenius norm estimation, in the sense that for any $M^* \in \mathbb{C}_{\text{HIGH}}(\gamma)$, it satisfies the bound

$$\frac{1}{n^2}|||\pi^*(M^*) - \widehat{\pi}_{\text{FAS}}(M^*)|||_{\text{F}}^2 \leq \frac{c \log n}{n} \tag{7a}$$

with high probability. In this statement, for any given permutation π, we have used $\pi(M^*)$ to denote the matrix obtained by permuting the rows and columns of M^* by π. The term $\frac{1}{n^2}|||\pi^*(M^*) - \widehat{\pi}_{\text{FAS}}(M^*)|||_{\text{F}}^2$ can be viewed in some sense as the *bias* in estimation incurred from using $\widehat{\pi}_{\text{FAS}}$ in place of π^*.

Step 2: Define $\mathbb{C}_{\text{BISO}}$ be the class of matrices $M \in [0,1]^{n \times n}$ that satisfy the linear constraints $M_{ij} = 1 - M_{ji}$ for all $(i, j) \in [n]^2$, and $M_{k\ell} \geq M_{ij}$ whenever $k \leq i$ and $\ell \geq j$. This class is the subset of matrices $\mathbb{C}_{\text{SST}}$ that are faithful with respect to the identity permutation. Letting $\widehat{\pi}_{\text{FAS}}(\mathbb{C}_{\text{BISO}}) = \{\widehat{\pi}_{\text{FAS}}(M),\, M \in \mathbb{C}_{\text{BISO}}\}$ denote the image of this set under $\widehat{\pi}_{\text{FAS}}$, the second step involves computing the constrained least-squares estimate

$$\widehat{M} \in \operatorname*{arg\,min}_{M \in \widehat{\pi}_{\text{FAS}}(\mathbb{C}_{\text{BISO}})} |||Y - M|||_{\text{F}}^2. \tag{7b}$$

Since the set $\widehat{\pi}_{\text{FAS}}(\mathbb{C}_{\text{BISO}})$ is a convex polytope with a number of facets that grows polynomially in n, the constrained quadratic program (7b) can be solved in polynomial-time. The final step in the proof of Theorem 3 is to show that the estimator $\widehat{M}$ also has mean-squared error that is upper bounded by a constant multiple of $\frac{\log^2 n}{n}$.

Our analysis shows that for any fixed $\gamma \in (0, \frac{1}{2}]$, the proposed two-step estimator works well for any matrix $M^* \in \mathbb{C}_{\text{HIGH}}(\gamma)$. Since this two-step estimator is based on finding a minimum feedback arc set (FAS) in the first step, it is natural to wonder whether an FAS-based estimator works well over the full class $\mathbb{C}_{\text{SST}}$. Somewhat surprisingly the answer to this question turns out to be negative: we refer the reader to an extended version of this paper (Shah et al., 2015) for more intuition and details on why the minimal FAS estimator does not perform well over the full class.

3.4. Optimal rates for parametric subclasses

Let us now return to the class of parametric models $\mathbb{C}_{\text{PAR}}(F)$ introduced earlier in Section 2.3. As shown previously in Proposition 1, this class is much smaller than the class $\mathbb{C}_{\text{SST}}$, in the sense that there are models in $\mathbb{C}_{\text{SST}}$ that cannot be well-approximated by any parametric model. Nonetheless, in terms of minimax rates of estimation, these classes differ only by logarithmic factors. An advantage of the parametric class is that it is possible to achieve the $1/n$ minimax rate by using a simple, polynomial-time estimator. In particular, for any log concave function F, the maximum likelihood estimate $\widehat{w}_{\text{ML}}$ can be obtained by solving a convex program. This MLE induces a matrix estimate $M(\widehat{w}_{\text{ML}})$ via Equation (2), and the following result shows that this estimator is minimax-optimal up to constant factors.

Theorem 4. *Suppose that F is strictly increasing, strongly log-concave and twice differentiable. Then the minimax risk over $\mathbb{C}_{PAR}(F)$ is lower bounded as*

$$\inf_{\widetilde{M}} \sup_{M^* \in \mathbb{C}_{PAR}(F)} \frac{1}{n^2}\mathbb{E}[|||\widetilde{M} - M^*|||_F^2] \geq \frac{c_\ell}{n}, \tag{8a}$$

Conversely, the matrix estimate $M(\widehat{w}_{ML})$ induced by the

MLE satisfies the bound

$$\sup_{M^* \in \mathbb{C}_{PAR}(F)} \frac{1}{n^2} \mathbb{E}[\|M(\widehat{w}_{ML}) - M^*\|_F^2] \leq \frac{c_u}{n}. \tag{8b}$$

To be clear, the constants (c_ℓ, c_u) in this theorem are independent of n, but they do depend on the specific properties of the given function F. We note that the stated conditions on F are true for many popular parametric models, including (for instance) the Thurstone and BTL models.

The lower bound (8a) is, in fact, stronger than the the lower bound in Theorem 1, since the supremum is taken over a smaller class. The proof of the lower bound in Theorem 1 relies on matrices that cannot be realized by any parametric model, so that we pursue a different route to establish the bound (8a). On the other hand, in order to prove the upper bound (8b), we make use of bounds on the accuracy of the MLE $\widehat{w}_{\text{ML}}$ from our past work (Shah et al., 2016a).

3.5. Extension to partial observations

We now consider the extension of our results to the setting in which not all entries of Y are observed. Suppose instead that every entry of Y is observed independently with probability p_{obs}. In other words, the set of pairs compared is the set of edges of an Erdős-Rényi graph $\mathcal{G}(n, p_{\text{obs}})$ that has the n items as its vertices.

In this setting, we obtain an upper bound on the minimax risk of estimation by first setting $Y_{ij} = \frac{1}{2}$ whenever the pair (i, j) is not compared, then forming a new matrix Y' as

$$Y' := \frac{1}{p_{\text{obs}}} Y - \frac{1 - p_{\text{obs}}}{2 p_{\text{obs}}} 11^T, \tag{9a}$$

and finally computing the least squares solution

$$\widehat{M} \in \arg\min_{M \in \mathbb{C}_{\text{SST}}} \|Y' - M\|_{\text{F}}^2. \tag{9b}$$

Likewise, the computationally-efficient singular value thresholding estimator is also obtained by thresholding the singular values of Y' with a threshold $\lambda_{n,p_{\text{obs}}} = \frac{2.1\sqrt{n}}{p_{\text{obs}}}$.

The parametric estimators continue to operate on the original (partial) observations, first computing a maximum likelihood estimate $\widehat{w}_{\text{ML}}$ of M^* using the observed data, and then computing the associated pairwise-comparison-probability matrix $M(\widehat{w}_{\text{ML}})$ via (2).

Theorem 5. *In the setting where each pair is observed with a probability p_{obs}:*

(a) The minimax risk is sandwiched as

$$\frac{c_\ell}{p_{\text{obs}} n} \leq \inf_{\widetilde{M}} \sup_{M^* \in \mathbb{C}_{SST}} \frac{1}{n^2} \mathbb{E}[\|\widetilde{M} - M^*\|_F^2] \leq \frac{c_u (\log n)^2}{p_{\text{obs}} n},$$

when $p_{\text{obs}} \geq \frac{c_0}{n}$.

(b) The soft-SVT estimator $\widehat{M}_{\lambda_{n,p_{\text{obs}}}}$ with $\lambda_{n,p_{\text{obs}}} = \frac{2.1\sqrt{n}}{p_{\text{obs}}}$ satisfies the bound

$$\sup_{M^* \in \mathbb{C}_{SST}} \frac{1}{n^2} \mathbb{E}[\|\widehat{M}_{\lambda_{n,p_{\text{obs}}}} - M^*\|_F^2] \leq \frac{c_u}{p_{\text{obs}} \sqrt{n}}.$$

(c) For a parametric sub-class based on a strongly log-concave and smooth F, the estimator $M(\widehat{w}_{ML})$ induced by the maximum likelihood estimate $\widehat{w}_{ML}$ of the parameter vector w^* has error upper bounded as

$$\sup_{M^* \in \mathbb{C}_{PAR}(F)} \frac{1}{n^2} \mathbb{E}[\|M(\widehat{w}_{ML}) - M^*\|_F^2] \leq \frac{c_u}{p_{\text{obs}} n},$$

when $p_{\text{obs}} \geq \frac{c_0 (\log n)^2}{n}$.

We note that we do not have an analogue of the high-SNR result in the partial observations case since having partial observations reduces the SNR. In general, we are interested in scalings of p_{obs} which allow $p_{\text{obs}} \to 0$ as $n \to \infty$. The noisy-sorting algorithm of Braverman & Mossel (2008) for the high-SNR case has computational complexity scaling as $e^{\gamma^{-4}}$, and hence is not computable in time polynomial in n when $\gamma < (\log n)^{-\frac{1}{4}}$. This restriction disallows most interesting scalings of p_{obs} with n.

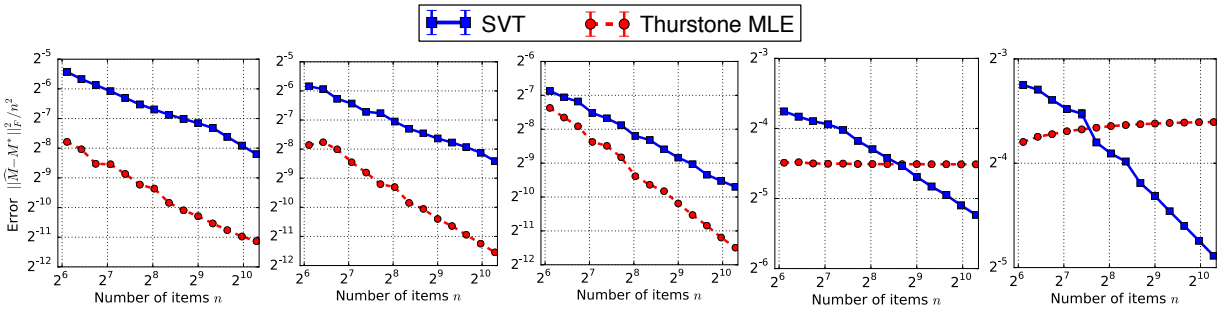
4. Simulations

In this section, we present results from simulations to gain a further understanding of the problem at hand, in particular to understand the rates of estimation under specific generative models. The simulations in this section add to the simulation results of Section 2.4 (Figure 1) demonstrating a large class of matrices in the SST class that cannot be represented by any parametric class. We investigate the performance of the soft-SVT estimator (Section 3.2) and the maximum likelihood estimator under the Thurstone model (Section 2.3).[1] The output of the SVT estimator need not lie in the set $[0, 1]^{n \times n}$ of matrices; in our implementation, we take a projection of the output of the SVT estimator on this set, which gives a constant factor reduction in the error.

In our simulations, we generate the ground truth M^* in the following five ways:

- Uniform: $\binom{n}{2}$ values in $[\frac{1}{2}, 1]$ are chosen independently and uniformly at random, and sorted in descending or-

[1] We could not compare the algorithm that underlies Theorem 3, since it is not easily implementable. In particular, it relies on the algorithm due to Braverman & Mossel (2008) to compute the feedback arc set minimizer. The computational complexity of this algorithm, though polynomial in n, has a large polynomial degree which precludes it from being implemented for matrices of any reasonable size.



(a) Uniform (b) Thurstone (c) BTL (d) High SNR (e) Independent bands

Figure 2: Error incurred by the singular value thresholding (SVT) estimator and the Thurstone MLE under different methods to generate M^*.

der. The values are inserted above the diagonal of an $(n \times n)$ matrix such that the entries decrease down a column or left along a row. We then make the matrix skew-symmetric and permute the rows and columns.

- Thurstone: M^* is generated by first choosing $w^* \in [-1, 1]^n$ uniformly at random from the set satisfying $\langle w^*, 1\rangle = 0$, and then applying Equation (2) with F chosen as the CDF of the standard normal distribution.
- Bradley-Terry-Luce (BTL): Identical to the Thurstone case, except that F is given by the sigmoid function.
- High SNR: A setting studied previously by Braverman & Mossel (2008), in which the noise is independent of the items being compared. Some global order of the n items is fixed, and $M^*_{ij} = 0.9 = 1 - M^*_{ji}$ for every pair (i, j) where i is ranked above j in the underlying ordering. The entries on the diagonal are 0.5.
- Independent bands: The diagonal entries of M^* are first set to $\frac{1}{2}$. Entries on the band immediately above the diagonal itself are chosen independently and uniformly at random from $[\frac{1}{2}, 1]$. The band above is then chosen uniformly at random from the allowable set, and so on. The choice of any entry in this process is only constrained to be upper bounded by 1 and lower bounded by the entries to its left and below. The entries below the diagonal are chosen to make the matrix skew-symmetric.

Figure 2 depicts the results of the simulations based on observations of the entire matrix Y. Each point is an average across 20 trials. The error bars in most cases are too small and hence not visible. We see that the uniform case (Figure 2a) is favorable for both estimators, with the error scaling as $\mathcal{O}(\frac{1}{\sqrt{n}})$. With data generated from the Thurstone model, both estimators continue to perform well, and the Thurstone MLE yields an error of the order $\frac{1}{n}$ (Figure 2b). Interestingly, the Thurstone model also fits relatively well when data is generated via the BTL model (Figure 2c). This behavior is likely a result of operating in the near-linear regime of the logistic and the Gaussian CDF where the two curves are similar. In these two parametric settings, the SVT estimator has squared error strictly worse than order $\frac{1}{n}$ but better than $\frac{1}{\sqrt{n}}$. The Thurstone model, however, yields a poor fit for the model in the high-SNR (Figure 2d) and the independent bands (Figure 2e) cases, incurring a constant error as compared to an error scaling as $\mathcal{O}(\frac{1}{\sqrt{n}})$ for the SVT estimator. We recall that the poor performance of the Thurstone estimator was also described previously in Proposition 1 and Figure 1.

In summary, we see that while the Thurstone MLE estimator gives minimax optimal rates of estimation when the underlying model is parametric, it can be inconsistent when the parametric assumptions are violated. On the other hand, the SVT estimator is robust to violations of parametric assumptions, and while it does not necessarily give minimax-optimal rates, it remains consistent across the entire SST class. Finally, we remark that our theory predicts that the least squares estimator, if implementable, would outperform both these estimators in terms of statistical error.

5. Discussion

We analyzed a flexible model based on stochastic transitivity for pairwise comparison data that includes various parametric models, including the BTL and Thurstone models, as special cases. We analyzed various estimators for this broader matrix family, ranging from optimal estimators through to various polynomial-time estimators, including forms of singular value thresholding, as well as multi-stage method based on a noisy sorting routine. We show that this SST model permits far more robust estimation as compared to popular parametric models, while surprisingly, incurring little penalty for this significant generality. Our results thus present a strong motivation towards the use of such general stochastic transitivity based models. Establishing the best possible rates for polynomial-time algorithms over the full class $\mathbb{C}_{\text{SST}}$ is a challenging open problem.

Acknowledgments

This work was partially supported by ONR-MURI grant DOD-002888, AFOSR grant FA9550-14-1-0016, NSF grant CIF-31712-23800, and ONR MURI grant N00014-11-1-0688. The work of NBS was also supported in part by a Microsoft Research PhD fellowship.

References

Ailon, N., Charikar, M., and Newman, A. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):23, 2008.

Alon, N. Ranking tournaments. *SIAM Journal on Discrete Mathematics*, 20(1):137–142, 2006.

Ballinger, T. P. and Wilcox, N. T. Decisions, error and heterogeneity. *The Economic Journal*, 107(443):1090–1105, 1997.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, pp. 324–345, 1952.

Braverman, M. and Mossel, E. Noisy sorting without resampling. In *Proc. ACM-SIAM symposium on Discrete algorithms*, pp. 268–276, 2008.

Brouwer, A. E. and Haemers, W. H. *Spectra of graphs*. Springer, 2011.

Cattelan, M. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 27 (3):412–433, 2012.

Chatterjee, S. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2014.

Chatterjee, S., Guntuboyina, A., and Sen, B. On matrix estimation under monotonicity constraints. *arXiv:1506.03430*, 2015.

Chung, F. and Radcliffe, M. On the spectra of general random graphs. *The electronic journal of combinatorics*, 18 (1):P215, 2011.

Davidson, D. and Marschak, J. Experimental tests of a stochastic decision theory. *Measurement: Definitions and theories*, pp. 233–69, 1959.

Fishburn, P. C. Binary choice probabilities: on the varieties of stochastic transitivity. *Journal of Mathematical psychology*, 10(4):327–352, 1973.

Fligner, M. A. and Verducci, J. S. *Probability models and statistical analyses for ranking data*, volume 80. Springer, 1993.

Gao, F. and Wellner, J. A. Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98(9):1751–1764, 2007.

Gilbert, E. N. A comparison of signalling alphabets. *Bell System Technical Journal*, 31(3):504–522, 1952.

Hajek, B., Oh, S., and Xu, J. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems*, pp. 1475–1483, 2014.

Kenyon-Mathieu, C. and Schudy, W. How to rank with few errors. In *Symposium on Theory of computing (STOC)*, pp. 95–103. ACM, 2007.

Klein, T., Rio, E., et al. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

Kolokolnikov, T., Osting, B., and Von Brecht, J. Algebraic connectivity of Erdös-Rényi graphs near the connectivity threshold. Available online http://www.mathstat.dal.ca/ tkolokol/papers/braxton-james.pdf, 2014.

Ledoux, M. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

Luce, R. D. *Individual choice behavior: A theoretical analysis*. New York: Wiley, 1959.

Marden, J. I. *Analyzing and modeling rank data*. CRC Press, 1996.

McLaughlin, D. H. and Luce, R. D. Stochastic transitivity and cancellation of preferences between bitter-sweet solutions. *Psychonomic Science*, 2(1-12):89–90, 1965.

Negahban, S., Oh, S., and Shah, D. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, pp. 2474–2482, 2012.

Oliveira, R. I. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *arXiv preprint:0911.0600*, 2009.

Shah, N. B. and Wainwright, M. J. Simple, robust and optimal ranking from pairwise comparisons. *arXiv preprint arXiv:1512.08949*, 2015.

Shah, N. B., Balakrishnan, S., Guntuboyina, A., and Wainright, M. J. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *arXiv preprint arXiv:1510.05610*, 2015.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., and Wainwright, M. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 2016a.

Shah, N. B., Balakrishnan, S., and Wainwright, M. J. Feeling the Bern: Adaptive estimators for Bernoulli probabilities of pairwise comparisons. *arXiv preprint arXiv:1603.06881*, 2016b.

Silvapulle, M. J. and Sen, P. K. *Constrained statistical inference: Order, inequality, and shape constraints*, volume 912. John Wiley & Sons, 2011.

Thompson, R. The behavior of eigenvalues and singular values under perturbations of restricted rank. *Linear Algebra and its Applications*, 13(1):69–78, 1976.

Thurstone, L. L. A law of comparative judgment. *Psychological Review*, 34(4):273, 1927.

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

Varshamov, R. Estimate of the number of signals in error correcting codes. In *Dokl. Akad. Nauk SSSR*, 1957.