Learning Mixtures of Plackett-Luce Models

Zhibing Zhao ZHAOZ6@RPI.EDU
Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180 USA

Peter Piech PIECHP@RPI.EDU
Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180 USA

Lirong Xia XIAL@CS.RPI.EDU
Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180 USA

Abstract

In this paper we address the identifiability and efficient learning problems of finite mixtures of Plackett-Luce models for rank data. We prove that for any $k \geq 2$, the mixture of k Plackett-Luce models for no more than $2k-1$ alternatives is non-identifiable and this bound is tight for $k = 2$. For generic identifiability, we prove that the mixture of k Plackett-Luce models over m alternatives is *generically identifiable* if $k \leq \lfloor \frac{m-2}{2} \rfloor !$. We also propose an efficient generalized method of moments (GMM) algorithm to learn the mixture of two Plackett-Luce models and show that the algorithm is consistent. Our experiments show that our GMM algorithm is significantly faster than the EMM algorithm by Gormley & Murphy (2008), while achieving competitive statistical efficiency.

1. Introduction

In many machine learning problems the data are composed of rankings over a finite number of *alternatives* (Marden, 1995). For example, meta-search engines aggregate rankings over webpages from individual search engines (Dwork et al., 2001); rankings over documents are combined to find the most relevant document in information retrieval (Liu, 2011); noisy answers from online workers are aggregated to produce a more accurate answer in crowdsourcing (Mao et al., 2013). Rank data are also very common in economics and political science. For example, consumers often give discrete choices data (McFadden, 1974) and voters often give rankings over presidential candidates (Gormley & Murphy, 2008).

Perhaps the most commonly-used statistical model for rank data is the *Plackett-Luce* model (Plackett, 1975; Luce, 1959). The Plackett-Luce model is a natural generalization of multinomial logistic regression. In a Plackett-Luce model, each alternative is parameterized by a positive number that represents the "quality" of the alternative. The greater the quality, the the chance the alternative will be ranked at a higher position.

In practice, *mixtures* of Plackett-Luce models can provide better fitness than a single Plackett-Luce model. An additional benefit is that the learned parameter of a mixture model can naturally be used for clustering (McLachlan & Basford, 1988). The k-mixture of Plackett-Luce combines k individual Plackett-Luce models via a linear vector of *mixing coefficients*. For example, Gormley & Murphy (2008) propose an *Expectation Minorization Maximization (EMM)* algorithm to compute the MLE of Plackett-Luce mixture models. The EMM was applied to an Irish election dataset with 5 alternatives and the four components in the mixture model are interpreted as *voting blocs*.

Surprisingly, the *identifiability* of Plackett-Luce mixture models is still unknown. Identifiability is an important property for statistical models, which requires that different parameters of the model have different distributions over samples. Identifiability is crucial because if the model is not identifiable, then there are cases where it is impossible to estimate the parameter from the data, and in such cases conclusions drawn from the learned parameter can be wrong. In particular, if Plackett-Luce mixture models are not identifiable, then the voting bloc produced by the EMM algorithm of Gormley & Murphy (2008) can be dramatically different from the ground truth.

In this paper, we address the following two important questions about the theory and practice of Plackett-Luce mixture models for rank data.

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

Q1. Are Plackett-Luce mixture models identifiable?

Q2. How can we efficiently learn Plackett-Luce mixture models?

Q1 can be more complicated than one may think because the non-identifiability of a mixture model usually comes from two sources. The first is *label switching*, which means that if we label the components of a mixture model differently, the distribution over samples does not change (Stephens, 2000). This can be avoided by ordering the components and merging the same components in the mixture model. The second is more fundamental, which states that the mixture model is non-identifiable even after ordering and merging duplicate components. Q1 is about the second type of non-identifiability.

The EMM algorithm by Gormley & Murphy (2008) converges to the MLE, but as we will see in the experiments, it can be very slow when the data size is not too small. Therefore, to answer Q2, we want to design learning algorithms that are much faster than the EMM without sacrificing too much statistical efficiency, especially mean squared error (MSE).

1.1. Our Contributions

We answer Q1 with the following theorems. The answer depends on the number of components k in the mixture model and the number of alternatives m.

Theorem 1 and 2. *For any $m \geq 2$ and any $k \geq \frac{m+1}{2}$, the k-mixture Plackett-Luce model (denoted by k-PL) is* **non-identifiable**. *This lower bound on k as a function of m is tight for $k = 2$ ($m = 4$).*

The second half of the theorem is positive: the mixture of two Plackett-Luce models is identifiable for four or more alternatives. We conjecture that the bound is tight for all $k > 2$.

The k-PL is *generically* identifiable for m alternatives, if the Lebesgue measure of non-identifiable parameters is 0. We prove the following positive results for k-PL.

Theorem 3. *For any $m \geq 6$ and any $k \leq \lfloor \frac{m-2}{2} \rfloor!$, the k-PL is generically identifiable.*

We note that $\lfloor \frac{m-2}{2} \rfloor!$ is exponentially larger than the lower bound $\frac{m+1}{2}$ for (strict) identifiability. One interpretation of the theorem is that, when $\frac{m}{2} + 1 \leq k \leq \lfloor \frac{m-2}{2} \rfloor!$, even though k-PL is not identifiable in the strict sense, one may not need to worry too much in practice due to generic identifiability.

For Q2, we propose a generalized method of moments (GMM)[1] algorithm (Hansen, 1982) to learn the k-PL. We illustrate the algorithm for $k = 2$ and $m \geq 4$, and prove that the algorithm is consistent, which means that when the data are generated from k-PL and the data size n goes to infinity, the algorithm will reveal the ground truth with probability that goes to 1. We then compare our GMM algorithm and the EMM algorithm (Gormley & Murphy, 2008) w.r.t. statistical efficiency (mean squared error) and computational efficiency in synthetic experiments. As we will see, in Section 5, our GMM algorithm is significantly faster than the EMM algorithm while achieving competitive statistical efficiency. Therefore, we believe that our GMM algorithm is a promising candidate for learning Plackett-Luce mixture models from big rank data.

1.2. Related Work and Discussions

Most previous work in mixture models (especially Gaussian mixture models) focuses on cardinal data (Teicher, 1961; 1963; McLachlan & Peel, 2004; Kalai et al., 2012; Dasgupta, 1999). Little is known about the identifiability of mixtures of models for rank data.

For rank data, Iannario (2010) proved the identifiability of the mixture of shifted binomial model and the uniform models. Awasthi et al. (2014) proved the identifiability of mixtures of two Mallows' models. Mallows mixture models were also studied by Lu & Boutilier (2014) and Chierichetti et al. (2015). Our paper, on the other hand, focuses on mixtures of Plackett-Luce models.

Technically, part of our (non-)identifiability proofs is motivated by the work of Teicher (1963), who obtained sufficient conditions for the identifiability of finite mixture models. However, technically these conditions cannot be directly applied to k-PL because they work either for finite families (Theorem 1 in (Teicher, 1963)) or for cardinal data (Theorem 2 in (Teicher, 1963)). Neither is the case for mixtures of Plackett-Luce models. To prove our (non-)identifiability theorems, we develop novel applications of the Fundamental Theorem of Algebra to analyze the rank of a matrix $\mathbf{F}_m^k$ that represents k-PL (see Preliminaries for more details). Our proof for generic identifiability is based on a novel application of the tensor-decomposition approach that analyzes the generic *Kruskal's rank* of matrices advocated by Allman et al. (2009).

In addition to being important in their own right, our (non)-identifiability theorems also carry a clear message that has been overlooked in the literature: when using Plackett-Luce mixture models to fit rank data, one must be very careful about the interpretation of the learned parameter. Specifically, when $m \leq 2k - 1$, it is necessary to double-check whether the learned parameter is identifiable (Theo-

[1] This should not be confused with *Gaussian mixture models*.

rem 1), which can be computationally hard. On the positive side, identifiability may not be a big concern in practice under a much milder condition ($k \le \lfloor \frac{m-2}{2} \rfloor !$, Theorem 3).

Gormley & Murphy (2008) used 4-PL to fit an Irish election dataset with 5 alternatives. According to our Theorem 1, 4-PL for 5 alternatives is non-identifiable. Moreover, our generic identifiability theorem (Theorem 3) does not apply because $m = 5 < 6$. Therefore, it is possible that there exists another set of voting blocs and mixing coefficients with the same likelihood as the output of the EMM algorithm. Whether it is true or not, we believe that it is important to add discussions and justifications of the uniqueness of the voting blocs obtained by Gormley & Murphy (2008).

Parameter inference for single Plackett-Luce models is studied in (Cheng et al., 2010) and (Azari Soufiani et al., 2013). Azari Soufiani et al. (2013) proposed a GMM, which is quite different from our method, and cannot be applied to Plackett-Luce mixture models. The MM algorithm by Hunter (2004), which is compared in (Azari Soufiani et al., 2013), is also very different from the EMM that is being compared in this paper.

2. Preliminaries

Let $\mathcal{A} = \{a_i | i = 1, 2, \cdots, m\}$ denote a set of m alternatives. Let $\mathcal{L}(\mathcal{A})$ denote the set of linear orders (rankings), which are transitive, antisymmetric and total binary relations, over $\mathcal{A}$. A ranking is often denoted by $a_{i_1} \succ a_{i_2} \succ \cdots \succ a_{i_m}$, which means that a_{i_1} is the most preferred alternative, a_{i_2} is the second preferred, a_{i_m} is the least preferred, etc. Let $P = (V_1, V_2, \cdots, V_n)$ denote the data (also called a *preference profile*), where for all $j \le n, V_j \in \mathcal{L}(\mathcal{A})$.

Definition 1 *(Plackett-Luce model). The parameter space is* $\Theta = \{\vec{\theta} = \{\theta_i | i = 1, 2, \cdots, m, \theta_i \in [0, 1], \sum_{i=1}^m \theta_i = 1\}\}$. *The sample space is* $\mathcal{S} = \mathcal{L}(\mathcal{A})^n$. *Given a parameter* $\vec{\theta} \in \Theta$, *the probability of any ranking* $V = a_{i_1} \succ a_{i_2} \succ \cdots \succ a_{i_m}$ *is*

$$\Pr_{PL}(V|\vec{\theta}) = \frac{\theta_{i_1}}{1} \times \frac{\theta_{i_2}}{\sum_{p>1} \theta_{i_p}} \times \cdots \times \frac{\theta_{i_{m-1}}}{\theta_{i_{m-1}} + \theta_{i_m}}$$

We assume that data are generated i.i.d. in the Plackett-Luce model. Therefore, given a preference profile P and $\vec{\theta} \in \Theta$, we have $\Pr_{\text{PL}}(P|\vec{\theta}) = \prod_{j=1}^n \Pr_{\text{PL}}(V_j|\vec{\theta})$.

The Plackett-Luce model has the following intuitive explanation. Suppose there are m balls, representing m alternatives in an opaque bag. Each ball a_i is assigned a quality value θ_i. Then, we generate a ranking in m stages. In each stage, we take one ball out of the bag. The probability for each remaining ball being taken out is the value assigned to it over the sum of the values assigned to the remaining balls. The order of drawing is the ranking over the alternatives.

We require $\sum_i \theta_i = 1$ to normalize the parameter so that the Plackett-Luce model is identifiable. It is not hard to verify that for any Plackett-Luce model, the probability for the alternative a_p ($p \le m$) to be ranked at the top of a ranking is θ_p; the probability for a_p to be ranked at the top and a_q ranked at the second position is $\frac{\theta_p \theta_q}{1 - \theta_p}$, etc.

Definition 2 *(k-mixture Plackett-Luce model). Given* $m \ge 2$ *and* $k \ge 2$, *we define the* k-mixture Plackett-Luce model *as follows. The sample space is* $\mathcal{S} = \mathcal{L}(\mathcal{A})^n$. *The parameter space has two parts. The first part is the* mixing coefficients $(\alpha_1, \ldots, \alpha_k)$ *where for all* $r \le k$, $\alpha_r \ge 0$, *and* $\sum_{r=1}^k \alpha_r = 1$. *The second part is* $(\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \ldots, \vec{\theta}^{(k)})$, *where* $\vec{\theta}^{(r)} = [\theta_1^{(r)}, \theta_2^{(r)}, \cdots, \theta_m^{(r)}]^\top$ *is the parameter of the* r*-th Plackett-Luce component. The probability of a ranking* V *is*

$$\Pr_{k\text{-}PL}(V|\vec{\theta}) = \sum_{r=1}^k \alpha_r \Pr_{PL}(V|\vec{\theta}^{(r)}),$$

where $\Pr_{PL}(V|\vec{\theta}^{(r)})$ *is the probability of* V *in the* r*-th Plackett-Luce model given* $\vec{\theta}^{(r)}$.

For simplicity we use k-PL to denote the k-mixture Plackett-Luce model.

Definition 3 *(Identifiability) Let* $\mathcal{M} = \{\Pr(\cdot|\vec{\theta}) : \vec{\theta} \in \Theta\}$ *be a statistical model.* $\mathcal{M}$ *is identifiable if for all* $\vec{\theta}, \vec{\gamma} \in \Theta$, *we have* $\Pr(\cdot|\vec{\theta}) = \Pr(\cdot|\vec{\gamma}) \implies \vec{\theta} = \vec{\gamma}$.

In this paper, we slightly modify this definition to eliminate the label switching problem. We say that k-PL is identifiable if there do not exist (1) $1 \le k_1, k_2 \le k$, *non-degenerate* $\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \cdots, \vec{\theta}^{(k_1)}, \vec{\gamma}^{(1)}, \vec{\gamma}^{(2)}, \cdots, \vec{\gamma}^{(k_2)}$, which means that these $k_1 + k_2$ vectors are pairwise different; (2) all strictly positive mixing coefficients $(\alpha_1^{(1)}, \ldots, \alpha_{k_1}^{(1)})$ and $(\alpha_1^{(2)}, \ldots, \alpha_{k_2}^{(2)})$, so that for all rankings V we have

$$\sum_{r=1}^{k_1} \alpha_r^{(1)} \Pr_{\text{PL}}(V|\vec{\theta}^{(r)}) = \sum_{r=1}^{k_2} \alpha_r^{(2)} \Pr_{\text{PL}}(V|\vec{\gamma}^{(r)})$$

Throughout the paper, we will represent a distribution over the $m!$ rankings over m alternatives for a Plackett-Luce component with parameter $\vec{\theta}^{(r)}$ as a column vector $\vec{f}_m(\vec{\theta})$ with $m!$ elements, one for each ranking and whose value is the probability of the corresponding ranking. For example, when $m = 3$, we have

$$\vec{f}_3(\vec{\theta}) = \begin{pmatrix} \Pr(a_1 \succ a_2 \succ a_3|\vec{\theta}) \\ \Pr(a_1 \succ a_3 \succ a_2|\vec{\theta}) \\ \Pr(a_2 \succ a_1 \succ a_3|\vec{\theta}) \\ \Pr(a_2 \succ a_3 \succ a_1|\vec{\theta}) \\ \Pr(a_3 \succ a_1 \succ a_2|\vec{\theta}) \\ \Pr(a_3 \succ a_2 \succ a_1|\vec{\theta}) \end{pmatrix} = \begin{pmatrix} \frac{\theta_1\theta_2}{1-\theta_1} \\ \frac{\theta_1\theta_3}{1-\theta_1} \\ \frac{\theta_1\theta_2}{1-\theta_2} \\ \frac{\theta_2\theta_3}{1-\theta_2} \\ \frac{\theta_1\theta_3}{1-\theta_3} \\ \frac{\theta_2\theta_3}{1-\theta_3} \end{pmatrix}$$

Given $\vec{\theta}^{(1)}, \dots, \theta^{(\vec{2k})}$, we define $\mathbf{F}_m^k$ as a $m! \times 2k$ matrix for k-PL with m alternatives

$$\mathbf{F}_m^k = \left[\vec{f_m}(\vec{\theta}^{(1)}) \quad \vec{f_m}(\vec{\theta}^{(2)}) \quad \cdots \quad \vec{f_m}(\vec{\theta}^{(2k)}) \right] \tag{1}$$

We note that $\mathbf{F}_m^k$ is a function of $\vec{\theta}^{(1)}, \dots, \vec{\theta}^{(2k)}$, which are often omitted. We prove the identifiability or non-identifiability of k-PL by analyzing the rank of $\mathbf{F}_m^k$. The reason that we consider $2k$ components is that we want to find (or argue that we cannot find) another k-mixture model that has the same distribution as the original one.

3. Identifiability of Plackett-Luce Mixture Models

We first prove a general lemma to reveal a relationship between the rank of $\mathbf{F}_m^k$ and the identifiability of Plackett-Luce mixture models. We recall that a set of vectors is non-degenerate if its elements are pairwise different.

Lemma 1 *If the rank of $\mathbf{F}_m^k$ is $2k$ for all non-degenerate $\vec{\theta}^{(1)}, \dots, \vec{\theta}^{(2k)}$, then k-PL is identifiable. Otherwise $(2k-1)$-PL is non-identifiable.*

Proof: Suppose for the sake of contradiction the rank of $\mathbf{F}_m^k$ is $2k$ for all non-degenerate $\vec{\theta}^{(1)}, \dots, \vec{\theta}^{(2k)}$ but k-PL is non-identifiable. Then, there exist non-degenerate $\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \cdots, \vec{\theta}^{(k_1)}$, $\vec{\gamma}^{(1)}, \vec{\gamma}^{(2)}, \cdots, \vec{\gamma}^{(k_2)}$ and all strictly positive mixing coefficients $(\alpha_1^{(1)}, \dots, \alpha_{k_1}^{(1)})$ and $(\alpha_1^{(2)}, \dots, \alpha_{k_2}^{(2)})$, such that for all rankings V, we have

$$\textstyle\sum_{r=1}^{k_1} \alpha_r^{(1)} \Pr_{\text{PL}}(V|\vec{\theta}^{(r)}) = \sum_{r=1}^{k_2} \alpha_r^{(2)} \Pr_{\text{PL}}(V|\vec{\gamma}^{(r)})$$

Let $\vec{\delta}^{(1)}, \vec{\delta}^{(2)}, \dots, \vec{\delta}^{(2k-(k_1+k_2))}$ denote any $2k - (k_1 + k_2)$ vectors so that $\{\vec{\theta}^{(1)}, \dots, \vec{\theta}^{(k_1)}, \vec{\gamma}^{(1)}, \dots, \vec{\gamma}^{(k_2)}, \vec{\delta}^{(1)}, \dots, \vec{\delta}^{(2k-(k_1+k_2))}\}$ is non-degenerate. It follows that the rank of the corresponding $\mathbf{F}_m^k$ is strictly smaller than $2k$, because $\sum_{r=1}^{k_1} \alpha_r^{(1)} \Pr_{\text{PL}}(V|\vec{\theta}^{(r)}) - \sum_{r=1}^{k_2} \alpha_r^{(2)} \Pr_{\text{PL}}(V|\vec{\gamma}^{(r)}) + \sum_{r=1}^{(2k-k_1-k_2)} \vec{\delta}^{(r)} \cdot 0 = 0$. This is a contradiction.

On the other hand, if $\text{rank}(\mathbf{F}_m^k) < 2k$ for some non-degenerate $\vec{\theta}$'s, then there exists a nonzero vector $\vec{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_{2k}]^\top$ such that $\mathbf{F}_m^k \cdot \vec{\alpha} = 0$. Suppose in $\vec{\alpha}$ there are k_1 positive elements and k_2 negative elements, then it follows that $\max\{k_1, k_2\}$-mixture model is not identifiable, and $\max\{k_1, k_2\} \leq 2k - 1$. ■

Theorem 1 *For any $m \geq 2$ and any $k \geq \frac{m+1}{2}$, the k-PL is non-identifiable.*

Proof sketch: The proof is constructive and is based on a refinement of the second half of Lemma 1. For any k and $m = 2k - 1$, we will define $\vec{\theta}^{(1)}, \dots, \vec{\theta}^{(2k)}$ and $\vec{\alpha} = [\alpha_1, \dots, \alpha_{2k}]^T$ such that (1) $\mathbf{F}_m^k \cdot \vec{\alpha} = 0$ and (2) $\vec{\alpha}$ has k positive elements and k negative elements. In each $\vec{\theta}^{(r)}$, the value for alternatives $\{a_2, \dots, a_m\}$ are the same. The proof for any $m < 2k - 1$ is similar.

Formally, let $m = 2k - 1$. For all $i \geq 2$ and $r \leq 2k$, we let $\theta_i^{(r)} = \frac{1-\theta_1^{(r)}}{2k-2}$, where $\theta_i^{(r)}$ is the parameter corresponding to the ith alternative of the rth model. We use e_r to represent $\theta_1^{(r)}$ and we use b_r to represent $\frac{1-\theta_1^{(r)}}{2k-2}$. It is not hard to check that the probability for a_1 to be ranked at the ith position in the rth Plackett-Luce model is

$$\frac{(2k-2)!}{(2k-1-i)!} \frac{e_r(b_r)^{i-1}}{\prod_{p=0}^{i-1}(1-pb_r)} \tag{2}$$

Then $\mathbf{F}_m^k$ can be reduced to a $(2k-1) \times (2k)$ matrix. Because $\text{rank}(\mathbf{F}_m^k) \leq 2k - 1 < 2k$, Lemma 1 immediately tells us that $(2k-1)$-PL is non-identifiable for $2k-1$ alternatives, but this is much weaker than what we are proving in this theorem. We now define a new $(2k-1) \times (2k)$ matrix $\mathbf{H}^k$ obtained from $\mathbf{F}_m^k$ by performing the following linear operations on row vectors. (i) Make the first row of $\mathbf{H}^k$ to be $\vec{1}$; (ii) for any $2 \leq i \leq 2k - 1$, the ith row of $\mathbf{H}^k$ is the probability for a_1 to be ranked at the $(i-1)$-th position according to (2); (iii) remove all constant factors.

More precisely, for any e_r we define the following function.

$$\vec{f^*}(e_r) = \begin{pmatrix} 1 \\ e_r \\ \frac{e_r(1-e_r)}{e_r+2k-3} \\ \vdots \\ \frac{e_r(1-e_r)^{2k-3}}{(e_r+2k-3)\cdots((2k-3)e_r+1)} \end{pmatrix}$$

Then we define $\mathbf{H}^k = [\vec{f^*}(e_1), \vec{f^*}(e_2), \cdots, \vec{f^*}(e_{2k})]$.

Lemma 2 *If there exist all different $e_1, e_2, \cdots, e_{2k} < 1$ and a non-zero vector $\vec{\beta^*} = [\beta_1^*, \beta_2^*, \cdots, \beta_{2k}^*]^\top$ such that (i) $\mathbf{H}^k \vec{\beta^*} = 0$ and (ii) $\vec{\beta^*}$ has k positive elements and k negative elements, then k-PL for $2k-1$ alternatives is not identifiable.*

Then, the theorem is proved by showing that the following $\vec{\beta^*}$ satisfies the conditions in Lemma 2. For any $r \leq 2k$,

$$\beta_r^* = \frac{\prod_{p=1}^{2k-3}(pe_r + 2k - 2 - p)}{\prod_{q \neq r}(e_r - e_q)} \tag{3}$$

Note that the numerator is always positive. ■

Theorem 2 *For $k = 2$, and any $m \geq 4$, the 2-PL is identifiable.*

Proof sketch: We will apply Lemma 1 to prove the theorem. That is, we will show that for all non-degenerate $\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \vec{\theta}^{(3)}, \vec{\theta}^{(4)}$ such that rank$(\mathbf{F}_4^2) = 4$. We recall that $\mathbf{F}_4^2$ is a 24×4 matrix. Instead of proving rank$(\mathbf{F}_4^2) = 4$ directly, we will first obtain a 4×4 matrix $\mathbf{F}^* = T \times \mathbf{F}_4^2$ by linearly combining some row vectors of $\mathbf{F}_4^2$ via a 4×24 matrix T. Then, we show that rank$(\mathbf{F}^*) = 4$, which implies that rank$(\mathbf{F}_4^2) = 4$.

For simplicity we use $[e_r, b_r, c_r, d_r]^\top$ to denote the parameter of r-th Plackett-Luce component for a_1, a_2, a_3, a_4 respectively. Namely, $\begin{bmatrix} \vec{\theta}^{(1)} & \vec{\theta}^{(2)} & \vec{\theta}^{(3)} & \vec{\theta}^{(4)} \end{bmatrix} = \begin{bmatrix} e_1 & e_2 & e_3 & e_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ d_1 & d_2 & d_3 & d_4 \end{bmatrix} = \begin{bmatrix} \vec{\omega}^{(1)} \\ \vec{\omega}^{(2)} \\ \vec{\omega}^{(3)} \\ \vec{\omega}^{(4)} \end{bmatrix}$, where for each $r \leq 4$, $\vec{\omega}^{(r)}$ is a row vector. We further let $\vec{1} = [1, 1, 1, 1]$.

Clearly we have $\sum_{i=1}^4 \vec{\omega}^{(i)} = \vec{1}$. Therefore, if there exist three $\vec{\omega}$'s, for example $\{\vec{\omega}^{(1)}, \vec{\omega}^{(2)}, \vec{\omega}^{(3)}\}$, such that $\vec{\omega}^{(1)}, \vec{\omega}^{(2)}, \vec{\omega}^{(3)}$ and $\vec{1}$ are linearly independent, then rank$(\mathbf{F}_4^2) = 4$ because each $\vec{\omega}^{(i)}$ corresponds to the probability of a_i being ranked at the top, which means that $\vec{\omega}^{(i)}$ is a linear combination of rows in $\mathbf{F}_4^2$. Because $\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \vec{\theta}^{(3)}, \vec{\theta}^{(4)}$ is non-degenerate, at least one of $\{\vec{\omega}^{(1)}, \vec{\omega}^{(2)}, \vec{\omega}^{(3)}, \vec{\omega}^{(4)}\}$ is linearly independent of $\vec{1}$. W.l.o.g. suppose $\vec{\omega}^{(1)}$ is linearly independent of $\vec{1}$. This means that not all of e_1, e_2, e_3, e_4 are equal. The theorem will be proved in the following two cases.

Case 1. $\vec{\omega}^{(2)}$, $\vec{\omega}^{(3)}$, and $\vec{\omega}^{(4)}$ are all linear combinations of $\vec{1}$ and $\vec{\omega}^{(1)}$.

Case 2. There exists a $\vec{\omega}^{(i)}$ (where $i \in \{2, 3, 4\}$) that is linearly independent of $\vec{1}$ and $\vec{\omega}^{(1)}$.

We will only show the proof for a subcase of Case 1 to illustrate the main idea. The full proof is quite involved and can be found in the full version on arXiv. In Case 1, for all $i = 2, 3, 4$ we can rewrite $\vec{\omega}^{(i)} = p_i \vec{\omega}^{(1)} + q_i$ for some constants p_i, q_i. Because $\vec{\omega}^{(1)} + \vec{\omega}^{(2)} + \vec{\omega}^{(3)} + \vec{\omega}^{(4)} = \vec{1}$, we have $p_2 + p_3 + p_4 = -1$ and $q_2 + q_3 + q_4 = 1$.

In this case for each $r \leq 4$, the r-th column of $\mathbf{F}_4^2$, which is $\vec{f}_4(\vec{\theta}^{(r)})$, is a function of e_r. Because the $\vec{\theta}$'s are non-degenerate, e_1, e_2, e_3, e_4 must be pairwise different. We will show the proof for the following subcase of Case 1.

Case 1.1: $p_2 + q_2 \neq 0$ and $p_2 + q_2 \neq 1$.

For this case we first define a 4×4 matrix $\hat{\mathbf{F}}$ as in Table 1. We use $\vec{1}$ and $\vec{\omega}^{(1)}$ to denote the first two rows of $\hat{F}$. $\vec{\omega}^{(1)}$ corresponds to the probability that a_1 is ranked at the top. We call such a probability a *moment*. Each moment is the sum of probabilities of some rankings. For example, the "$a_1 \succ$ others" moment is the total probability for $\{V \in \mathcal{L}(\mathcal{A}) : a_1 \text{ is ranked at the top of } V\}$. It follows that there exists a 4×24 matrix $\hat{T}$ such that $\hat{\mathbf{F}} = \hat{T} \times \mathbf{F}_4^2$.

$\hat{\mathbf{F}}$	Moments
$\begin{bmatrix} 1 & 1 & 1 & 1 \\ e_1 & e_2 & e_3 & e_4 \\ \frac{e_1 b_1}{1-b_1} & \frac{e_2 b_2}{1-b_2} & \frac{e_3 b_3}{1-b_3} & \frac{e_4 b_4}{1-b_4} \\ \frac{e_1 b_1}{1-e_1} & \frac{e_2 b_2}{1-e_2} & \frac{e_3 b_3}{1-e_3} & \frac{e_4 b_4}{1-e_4} \end{bmatrix}$	$\vec{1}$ $a_1 \succ$ others $a_2 \succ a_1 \succ$ others $a_1 \succ a_2 \succ$ others

Table 1. $\hat{\mathbf{F}}$.

Define $\vec{\theta}^{(b)} = [\frac{1}{1-b_1}, \frac{1}{1-b_2}, \frac{1}{1-b_3}, \frac{1}{1-b_4}]$, where $b_i = p_2 e_i + q_2$. We then define $\vec{\theta}^{(e)} = [\frac{1}{1-e_1}, \frac{1}{1-e_2}, \frac{1}{1-e_3}, \frac{1}{1-e_4}]$, and let $\mathbf{F}^* = \begin{bmatrix} \vec{1} \\ \vec{\omega}^{(1)} \\ \vec{\theta}^{(b)} \\ \vec{\theta}^{(e)} \end{bmatrix}$. It can be verified that $\hat{\mathbf{F}} = T^* \times \mathbf{F}^*$, where

$$T^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{1}{p_2} & -1 & \frac{1-q_2}{p_2} & 0 \\ -(p_2 + q_2) & -p_2 & 0 & p_2 + q_2 \end{bmatrix}$$

Because Case 1.1 assumes that $p_2 + q_2 \neq 0$ and we can select a_2 such that $p_2 \neq 0$, $q_2 \neq 1$, we have that T^* is invertible. Therefore, $\mathbf{F}^* = (T^*)^{-1} \times \hat{\mathbf{F}}$, which means that $\mathbf{F}^* = T \times \mathbf{F}_4^2$ for some 4×24 matrix T.

We now prove that rank$(\mathbf{F}^*) = 4$. For the sake of contradiction, suppose that rank$(\mathbf{F}^*) < 4$. It follows that there exist a nonzero row vector $\vec{t} = [t_1, t_2, t_3, t_4]$, such that $\vec{t} \cdot \mathbf{F}^* = 0$. This means that for all $r \leq 4$,

$$t_1 + t_2 e_r + \frac{t_3}{1 - p_2 e_r - q_2} + \frac{t_4}{1 - e_r} = 0$$

Let $f(x) = t_1 + t_2 x + \frac{t_3}{1 - p_2 x - q_2} + \frac{t_4}{1-x}$. Let $g(x) = (1 - p_2 x - q_2)(1 - x) f(x)$. We recall that e_1, e_2, e_3, e_4 are four roots of $f(x)$, which means that they are also the four roots of $g(x)$. Because in Case 1.1 we assume that $p_2 + q_2 \neq 1$, it can be verified that not all coefficients of $g(x)$ are zero. We note that the degree of $g(x)$ is 3. Therefore, due to the Fundamental Theorem of Algebra, $g(x)$ has at most three different roots. This means that e_1, e_2, e_3, e_4 are not pairwise different, which is a contradiction.

Therefore, rank$(\mathbf{F}^*) = 4$, which means that rank$(\mathbf{F}_4^2) = 4$. This finishes the proof for Case 1.1. All missing proofs can be found in the full version on arXiv. ■

Slightly abusing the notation, we say that a parameter of k-PL is *identifiable*, if there does not exist a different parameter modulo label switching with the same probability distribution over the sample space. The next theorem proves that the Lebesgue measure (in the $km - 1$ dimensional Euclidean space) of non-identifiable parameters of k-PL for

m alternatives is 0 (generic identifiability as is defined in Section 1.1).

Theorem 3 *For $1 \leq k \leq \lfloor \frac{m-2}{2} \rfloor !$, k-PL over $m \geq 6$ alternatives is generically identifiable.*

Proof: The theorem is proved by analyzing the uniqueness of tensor decomposition. We construct a rank-one tensor for each Plackett-Luce component. Then the k-mixture model can be represented by another tensor, which is the weighted sum of k rank-one tensors. If the tensor decomposition is unique, then k-PL is identifiable.

To construct the rank-one tensor $\mathbf{T}_r$ for the r-th Plackett-Luce component, we partition the set of alternatives into three sets. In the rest of the proof we assume that m is even. The theorem can be proved similarly for odd m.

$$\begin{aligned} S_A &= \{a_1, a_2, \cdots, a_{\frac{m-2}{2}}\} \\ S_B &= \{a_{\frac{m}{2}}, a_{\frac{m+2}{2}}, \cdots, a_{m-2}\} \\ S_C &= \{a_{m-1}, a_m\} \end{aligned}$$

There are $n_1 = n_2 = \frac{m-2}{2}!$ rankings over S_A and S_B respectively, and two rankings over S_C ($n_3 = 3$). Let the three coordinates in the tensor $\mathbf{T}_r$ for the r-th Plackett-Luce model (with parameter $\vec{\theta}^{(r)}$ be $\mathbf{p}_A^{(r)}, \mathbf{p}_B^{(r)}, \mathbf{p}_C^{(r)}$ that represent probabilities of all rankings within S_A, S_B, S_C respectively.

Then, for any rankings $V_A \in \mathcal{L}(S_A)$, $V_B \in \mathcal{L}(S_B)$, and $V_C \in \mathcal{L}(S_C)$, we can prove that $\Pr_{\text{PL}}(V_A, V_B, V_C|\vec{\theta}^{(r)}) = \Pr_{\text{PL}}(V_A|\vec{\theta}^{(r)}) \times \Pr_{\text{PL}}(V_B|\vec{\theta}^{(r)}) \times \Pr_{\text{PL}}(V_C|\vec{\theta}^{(r)})$. That is, V_A, V_B and V_C are independent given $\vec{\theta}^{(r)}$. We will prove this result for a more general class of models called random utility models (RUM), of which the Plackett-Luce model is a special case (Thurstone, 1927).

Lemma 3 *Given a random utility model $\mathcal{M}(\vec{\theta})$ over a set of m alternatives $\mathcal{A}$, let $\mathcal{A}_1, \mathcal{A}_2$ be two non-overlapping subsets of $\mathcal{A}$, namely $\mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{A}$ and $\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$. Let V_1, V_2 be rankings over $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively, then we have* $\Pr(V_1, V_2|\vec{\theta}) = \Pr(V_1|\vec{\theta}) \Pr(V_2|\vec{\theta})$.

Because S_A, S_B, and S_C are non-overlapping, it follows that $\mathbf{T}_r = \mathbf{p}_A^{(r)} \otimes \mathbf{p}_B^{(r)} \otimes \mathbf{p}_C^{(r)}$. Because $k \leq \lfloor \frac{m-2}{2} \rfloor !$, we have $\min\{k, |\mathcal{L}(S_A)|\} + \min\{k, |\mathcal{L}(S_B)|\} + \min\{k, |\mathcal{L}(S_C)|\} = 2k + 2$. By Corollary 3 in (Allman et al., 2009), k-PL is generically identifiable. For completeness we include Corollary 3 here. Let $\mathcal{M}(k; n_1, n_2, n_3)$ be a k-mixture, 3-feature statistical model, where n_1, n_2, n_3 are the cardinalities of the three sets of events we defined.

The parameters of the model $\mathcal{M}(k; n_1, n_2, n_3)$ are generically identifiable, up to label switching, provided $\min(k, n_1) + \min(k, n_2) + \min(k, n_3) \geq 2k + 2$.

Since $n_1 = n_2 = \frac{m-2}{2}!$, $n_3 \geq 2$, this condition holds. ■

4. A Generalized Method of Moments Algorithm for 2-PL

In a *generalized method of moments* (GMM) algorithm, a set of $q \geq 1$ *moment conditions* $g(V, \vec{\theta})$ are specified. Moment conditions are functions of the parameter and the data, whose expectations are zero at the ground truth. $g(V, \vec{\theta}) \in \mathbb{R}^q$ has two inputs: a data point V and a parameter $\vec{\theta}$. For any $\vec{\theta}^*$, the expectation of any moment condition should be zero at $\vec{\theta}^*$, when the data are generated from the model given $\vec{\theta}^*$. Formally $E[g(V, \vec{\theta}^*)] = \vec{0}$. In practice the observed moment values should match the theoretical values from the model. In our algorithm, each moment condition corresponds to an event in the data, e.g. a_1 is ranked at the top. We use *moments* to denote such events. Given any preference profile P, we let $g(P, \vec{\theta}) = \frac{1}{n} \sum_{V \in P} g(V, \vec{\theta})$, which is a function of $\vec{\theta}$. The GMM algorithm we will use then computes the parameter that minimizes the 2-norm of the empirical moment conditions in the following way.

$$\text{GMM}_g(P) = \inf_{\vec{\theta}} ||g(P, \vec{\theta})||_2^2 \qquad (4)$$

In this paper, we will show results for $m = 4$ and $k = 2$. Our GMM works for other combinations of k and m, if the model is identifiable. Otherwise the estimator is not consistent. For $m = 4$ and $k = 2$, the parameter of the 2-PL is $\vec{\theta} = (\alpha, \vec{\theta}^{(1)}, \vec{\theta}^{(2)})$. We will use the following $q = 20$ moments from three categories.

(i) There are four moments, one for each of the four alternatives to be ranked at the top. Let $\{g_i : i \leq 4\}$ denote the four moment conditions. Let $p_i = \alpha\theta_i^{(1)} + (1-\alpha)\theta_i^{(2)}$. For any $V \in \mathcal{L}(\mathcal{A})$, we have $g_i(V, \vec{\theta}) = 1 - p_i$ if and only if a_i is ranked at the top of V; otherwise $g_i(V, \vec{\theta}) = -p_i$.

(ii) There are 12 moments, one for each combination of top-2 alternatives in a ranking. Let $\{g_{i_1 i_2} : i_1 \neq i_2 \leq 4\}$ denote the 12 moment conditions. Let $p_{i_1 i_2} = \alpha \frac{\theta_{i_1}^{(1)} \theta_{i_2}^{(1)}}{1 - \theta_{i_1}^{(1)}} + (1-\alpha) \frac{\theta_{i_1}^{(2)} \theta_{i_2}^{(2)}}{1 - \theta_{i_1}^{(2)}}$. For any $V \in \mathcal{L}(\mathcal{A})$, we have $g_{i_1 i_2}(V, \vec{\theta}) = 1 - p_{i_1 i_2}$ if and only if a_{i_1} is ranked at the top and a_{i_2} is ranked at the second in V; otherwise $g_{i_1 i_2}(V, \vec{\theta}) = -p_{i_1 i_2}$.

(iii) There are four moments that correspond to the following four rankings $a_1 \succ a_2 \succ a_3 \succ a_4$, $a_2 \succ a_3 \succ a_4 \succ a_1$, $a_3 \succ a_4 \succ a_1 \succ a_2$, $a_4 \succ a_1 \succ a_2 \succ a_3$. The corresponding $g_{i_1 i_2 i_3 i_4}$'s are defined similarly.

To illustrate how GMM works, we give the following example. Suppose the data P contain the following 20 rankings.

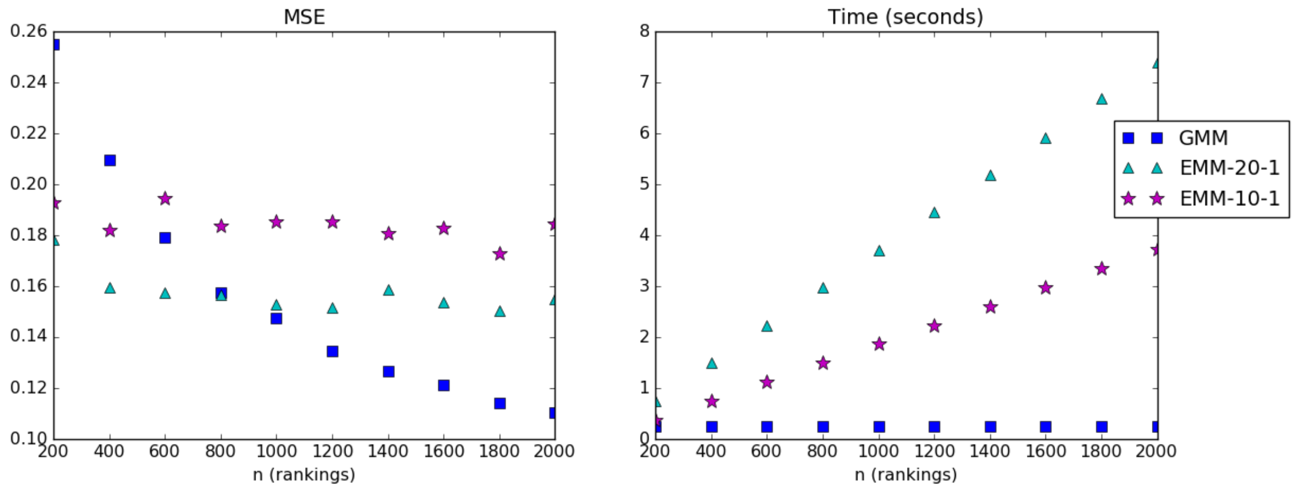

Figure 1. The MSE and running time of GMM and EMM. EMM-20-1 is 20 iterations of EMM overall with 1 iteration of MM for each M step. Likewise, EMM-10-1 is 10 iterations of EMM with 1 iteration of MM each time. Values are calculated over 2000 datasets.

$$\begin{aligned}
8 \ @ \ & a_1 \succ a_2 \succ a_3 \succ a_4 \\
4 \ @ \ & a_2 \succ a_3 \succ a_4 \succ a_1 \\
3 \ @ \ & a_3 \succ a_4 \succ a_1 \succ a_2 \\
3 \ @ \ & a_4 \succ a_1 \succ a_2 \succ a_3 \\
2 \ @ \ & a_3 \succ a_1 \succ a_4 \succ a_2
\end{aligned}$$

Then, for example, $g_1(P, \vec{\theta}) = \frac{8}{20} - (\alpha\theta_1^{(1)} + (1-\alpha)\theta_1^{(2)})$, corresponding to the moment that a_1 is ranked at top (category i). $g_{34}(P, \vec{\theta}) = \frac{3}{20} - (\frac{\alpha\theta_3^{(1)}\theta_4^{(1)}}{\theta_1^{(1)}+\theta_2^{(1)}+\theta_4^{(1)}} + \frac{(1-\alpha)\theta_3^{(2)}\theta_4^{(2)}}{\theta_1^{(2)}+\theta_2^{(2)}+\theta_4^{(2)}})$, corresponding to the moment of $a_3 \succ a_4 \succ$ others (category ii). $g_{2341}(P, \vec{\theta}) = \frac{4}{20} - (\frac{\alpha\theta_2^{(1)}\theta_3^{(1)}\theta_4^{(1)}}{(\theta_3^{(1)}+\theta_4^{(1)}+\theta_1^{(1)})(\theta_4^{(1)}+\theta_1^{(1)})} + \frac{(1-\alpha)\theta_2^{(2)}\theta_3^{(2)}\theta_4^{(2)}}{(\theta_3^{(2)}+\theta_4^{(2)}+\theta_1^{(2)})(\theta_4^{(2)}+\theta_1^{(2)})})$, corresponding to the moment of $a_2 \succ a_3 \succ a_4 \succ a_1$ (category iii). Remember that $\sum_{i=1}^{4}\theta_i^{(1)} = \sum_{i=1}^{4}\theta_i^{(2)} = 1$.

The choices of these moment conditions are based on the proof of Theorem 2, so that the 2-PL is strictly identifiable w.r.t. these moment conditions. Therefore, our simple GMM algorithm is the following.

```
Algorithm 1 GMM for 2-PL
Input: Preference profile $P$ with $n$ full rankings.
Compute the frequency of each of the 20 moments
Compute the output according to (4)
```

The theoretical guarantee of our GMM is its consistency, as we defined in Section 1.1.

Theorem 4 *Algorithm 1 is consistent w.r.t.* 2*-PL, where there exists* $\epsilon > 0$ *such that each parameter is in* $[\epsilon, 1]$.

Originally all parameters lie in open intervals (0, 1]. The ϵ requirement in the theorem is introduced to make the parameter space compact, i.e. all parameters are chosen from closed intervals. The proof is done by applying Theorem 3.1 in (Hall, 2005). The main hardness is the identifiability of 2-PL w.r.t. the moment conditions used in our GMM. Our proof of the identifiability of 2-PL (Theorem 2) only uses the 20 moment conditions described above.[2]

Complexity of GMM. For learning k-PL with m alternatives and n rankings with EMM, each E-step performs $O(nk^2)$ operations and each iteration of the MM algorithm for the M-step performs $O(m^2nk)$ operations. Our GMM for $k = 2$ and $m = 4$ has overall complexity $O(n)$. The complexity of calculating moments is $O(n)$ and the complexity of optimization depends only on m and k.

5. Experiments

The performance of our GMM algorithm (Algorithm 1) is compared to the EMM algorithm (Gormley & Murphy, 2008) for 2-PL with respect to running time and statistical efficiency for synthetic data. The synthetic datasets are generated as follows.

- Generating the ground truth: for $k = 2$ mixtures and $m = 4$ alternatives, the mixing coefficient α^* is generated uniformly at random and the Plackett-Luce components $\vec{\theta}^{(1)}$ and $\vec{\theta}^{(2)}$ are each generated from the Dirichlet distribution Dir$(\vec{1})$.

- Generating data: given a ground truth $\vec{\theta}^*$, we generate

[2] In fact our proof only uses 16 of them (4 out of the 12 moment conditions in category (ii) are redundant). However, our synthetic experiments show that using 20 moments improves statistical efficiency without sacrificing too much computational efficiency.

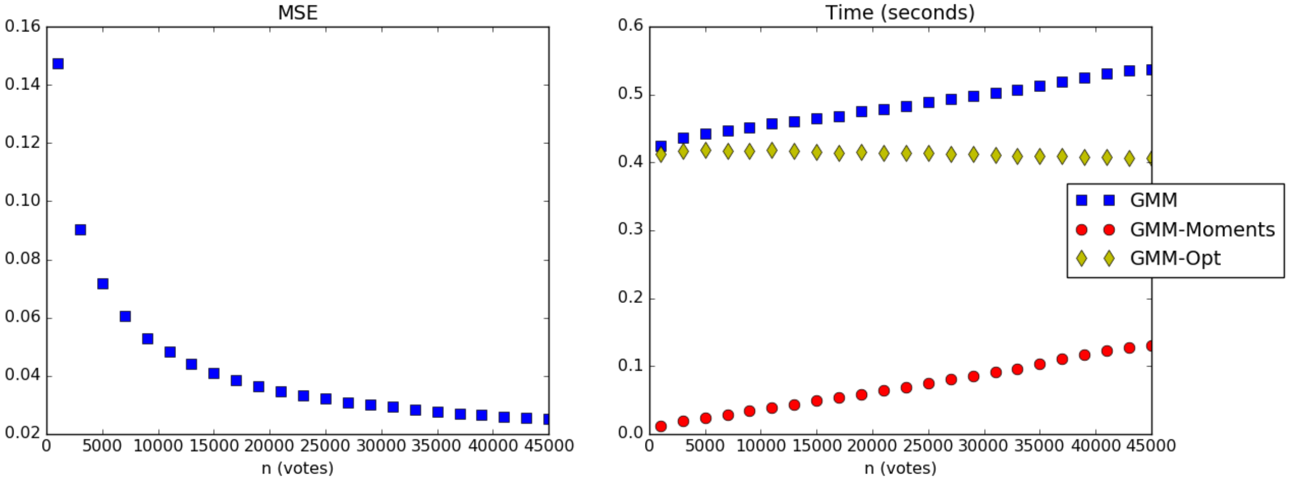

Figure 2. The MSE and running time of GMM. GMM-Moments is the time to calculate the moment condition values observed in the data and GMM-Opt is the time to perform the optimization. Values are calculated over 50000 trials.

each ranking with probability α^* from the PL model parameterized by $\vec{\theta}^{(1)}$ and with probability $1 - \alpha^*$ from the PL model parameterized by $\vec{\theta}^{(2)}$ up to 45000 full rankings.

The GMM algorithm is implemented in Python 3.4 and termination criteria for the optimization are convergence of the solution and the objective function values to within 10^{-10} and 10^{-6} respectively. The optimization of (4) uses the fmincon function through the MATLAB Engine for Python.

The EMM algorithm is also implemented in Python 3.4 and the E and M steps are repeated together for a fixed number of iterations without convergence criteria. The running time of EMM is largely determined by the total number of iterations of the MM algorithm and we look to compare the best EMM configuration with comparable running time to GMM. We have tested all configurations of EMM with 10 and 20 overall MM iterations, respectively. We found that the optimal configurations are EMM-10-1 and EMM-20-1 (shown in Figure 1, results for other configurations are omitted), where EMM-20-1 means 20 iterations of E step, each of which uses 1 MM iteration.

We use the Mean Squared Error (MSE) as the measure of statistical efficiency defined as MSE $= E(\| \vec{\theta} - \vec{\theta}^* \|_2^2)$.

All experiments are run on an Ubuntu Linux server with Intel Xeon E5 v3 CPUs each clocked at 3.50 GHz.

Results. The comparison of the performance of the GMM algorithm to the EMM algorithm is presented in Figure 1 for up to $n = 2000$ rankings. Statistics are calculated over 2000 trials (datasets). We observe that:

• GMM is significantly faster than EMM. The running time of GMM grows at a much slower rate in n.

• GMM achieves competitive MSE and the MSE of EMM does not improve over n. In particular, when n is more than 1000, GMM achieves smaller MSE.

The implication is that GMM may be better suited for reasonably large datasets where running time becomes infeasibly large with EMM. Moreover, it is possible that the GMM algorithm can be further improved by using a more accurate optimizer or another set of moment conditions. GMM can also be used to provide a good initial point for other methods such as the EMM.

For larger datasets, the performance of the GMM algorithm is shown in Figure 2 for up to $n = 45000$ rankings calculated over 50000 trials. As the data size increases, GMM converges toward the ground truth, which verifies our consistency theorem (Theorem 4). The overall running time of GMM shown in the figure is comprised of the time to calculate the moments from data (GMM-Moments) and the time to optimize the objective function (GMM-Opt). The time for calculating the moment values increases linearly in n, but it is dominated by the time to perform the optimization.

6. Future Work

There are many directions for future research. An obvious open question is whether k-PL is identifiable for $2k$ alternatives for $k \geq 3$, which we conjecture to be true. It is important to study how to efficiently check whether a learned parameter is identifiable for k-PL when $m < 2k$. Can we further improve the statistical efficiency and computational efficiency for learning k-PL? We also plan to develop efficient implementations of our GMM algorithm and apply it widely to various learning problems with big rank data.

Acknowledgments

This work is supported by the National Science Foundation under grant IIS-1453542 and a Simons-Berkeley research fellowship. We thank all reviewers for helpful comments and suggestions.

References

Allman, Elizabeth S., Matias, Catherine, and Rhodes, John A. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

Awasthi, Pranjal, Blum, Avrim, Sheffet, Or, and Vijayaraghavan, Aravindan. Learning Mixtures of Ranking Models. In *Proceedings of Advances in Neural Information Processing Systems*, pp. 2609–2617, 2014.

Azari Soufiani, Hossein, Chen, William, Parkes, David C., and Xia, Lirong. Generalized method-of-moments for rank aggregation. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, USA, 2013.

Cheng, Weiwei, Dembczynski, Krzysztof J., and Hüllermeier, Eyke. Label ranking methods based on the plackett-luce model. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 215–222, 2010.

Chierichetti, Flavio, Dasgupta, Anirban, Kumar, Ravi, and Lattanzi, Silvio. On learning mixture models for permutations. *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pp. 85–92, 2015.

Dasgupta, Sanjoy. Learning mixtures of gaussians. *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pp. 634–644, 1999.

Dwork, Cynthia, Kumar, Ravi, Naor, Moni, and Sivakumar, D. Rank aggregation methods for the web. In *Proceedings of the 10th World Wide Web Conference*, pp. 613–622, 2001.

Gormley, Isobel Claire and Murphy, Thomas Brendan. Exploring voting blocs within the Irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, 103(483):1014–1027, 2008.

Hall, Alastair R. *Generalized Method of Moments*. Oxford University Press, 2005.

Hansen, Lars Peter. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50 (4):1029–1054, 1982.

Hunter, David R. MM algorithms for generalized Bradley-Terry models. In *The Annals of Statistics*, volume 32, pp. 384–406, 2004.

Iannario, Maria. On the identifiability of a mixture model for ordinal data. *Metron*, 68(1):87–94, 2010.

Kalai, Adam Tauman, Moitra, Ankur, and Valiant, Gregory. Disentangling gaussians. In *Communications of the ACM*, volume 55, pp. 113–120, 2012.

Liu, Tie-Yan. *Learning to Rank for Information Retrieval*. Springer, 2011.

Lu, Tyler and Boutilier, Craig. Effective sampling and learning for mallows models with pairwise-preference data. *The Journal of Machine Learning Research*, 15 (1):3783–3829, 2014.

Luce, Robert Duncan. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

Mao, Andrew, Procaccia, Ariel D., and Chen, Yiling. Better human computation through principled voting. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Bellevue, WA, USA, 2013.

Marden, John I. *Analyzing and Modeling Rank Data*. Chapman & Hall, 1995.

McFadden, Daniel. Conditional logit analysis of qualitative choice behavior. In *Frontiers of Econometrics*, pp. 105–142, New York, NY, 1974. Academic Press.

McLachlan, Geoffrey and Peel, David. *Finite Mixture Models*. John Wiley & Sons, 2004.

McLachlan, Geoffrey J. and Basford, Kaye E. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, 1988.

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

Stephens, Matthew. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

Teicher, Henry. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248, 1961.

Teicher, Henry. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269, 1963.

Thurstone, Louis Leon. A law of comparative judgement. *Psychological Review*, 34(4):273–286, 1927.