Learning Mixtures of Random Utility Models with Features from Incomplete Preferences

Zhibing Zhao[1*], **Ao Liu**[2], **Lirong Xia**[2]
[1]Microsoft, 555 110TH Ave NE, Bellevue, WA, 98004
[2]Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY, 12180
zhaozb08@gmail.com, liua6@rpi.edu, xial@cs.rpi.edu

Abstract

Random Utility Models (RUMs), which subsume Plackett-Luce model (PL) as a special case, are among the most popular models for preference learning. In this paper, we consider RUMs with features and their mixtures, where each alternative has a vector of features, possibly different across agents. Such models significantly generalize the standard PL and RUMs, but are not as well investigated in the literature. We extend mixtures of RUMs with features to models that generate incomplete preferences and characterize their identifiability. For PL, we prove that when PL with features is identifiable, its MLE is consistent with a strictly concave objective function under mild assumptions, by characterizing a bound on root-mean-square-error (RMSE), which naturally leads to a sample complexity bound. We also characterize identifiability of more general RUMs with features and propose a generalized RBCML to learn them. Our experiments on synthetic data demonstrate the effectiveness of MLE on PL with features with tradeoffs between statistical efficiency and computational efficiency. Our experiments on real-world data show the prediction power of PL with features and its mixtures.

1 Introduction

Preference learning is a fundamental machine learning problem in a wide range of applications such as discrete choice analysis [McFadden, 1973], marketing [Berry *et al.*, 1995], meta-search engines [Dwork *et al.*, 2001], information retrieval [Liu, 2009], recommender systems [Baltrunas *et al.*, 2010], crowdsourcing [Chen *et al.*, 2013; Mao *et al.*, 2013], social choice [Azari Soufiani *et al.*, 2012], among many others. Plackett-Luce model (PL) [Plackett, 1975; Luce, 1959] is one of the most popular statistical models for preference learning due to its interpretability and computational tractability. In the standard PL, each alternative is parameterized by a real number θ such that e^θ represents the alternative's "quality". The higher the quality is, the more likely the alternative is ranked higher by an agent.

Extensions of the standard PL has been proposed and studied mostly in three dimensions. The first dimension is *PL with features*, where features of the agents and/or the alternatives are given, and the model is parameterized by the relationship (often linear, see Definition 2) between the features and the quality of the alternatives. Examples include the conditional logit model [McFadden, 1973], the BLP model [Berry *et al.*, 1995], and bilinear models [Azari Soufiani *et al.*, 2013; Schäfer and Hüllermeier, 2018; Zhao *et al.*, 2018a].

The second dimension is *PL for partial preferences*, where the data consist of partial preferences, often represented by partial orders over the alternatives. Due to the hardness of tackling general partial orders [Liu *et al.*, 2019], most previous work focused on natural sub-cases, including choice data [Train, 2009], top-ranked orders (top-l) [Mollica and Tardella, 2017], and pairwise preferences [Hüllermeier *et al.*, 2008]. In particular, in a top-l order, the agent reports a linear order over her most-preferred l alternatives. Top-l orders generalize standard PL ($l = m - 1$, where m is the number of alternatives) and choice data ($l = 1$).

The third dimension is *PL mixture models*, where $k \geq 1$ PL models are combined via a *mixing coefficients* $\vec{\alpha} = (\alpha_1, \ldots, \alpha_k)$ and each α_i represents the probability that the data is generated from the i-th PL. Mixtures of PLs provide better fitness to data [Tkachenko and Lauw, 2016] and are a popular method for clustering [Gormley and Murphy, 2008], but they are generally hard to compute and are prone to criticisms on interpretability and trustworthiness due to their (non-)identifiability [Zhao *et al.*, 2016].

While there is a large literature on standard PL and its extensions in each of the three dimensions, little is known about the generalization of PL in all three dimensions simultaneously, i.e. mixtures of PL models with features for partial preferences. The literature on general RUMs and their extensions is far more limited. The problem is already highly challenging for top-l orders—to the best of our knowledge, only one previous work studied mixtures of PL models with features for top-l orders [Tkachenko and Lauw, 2016], where an EM algorithm was proposed yet no theoretical guarantees on the model or the algorithm were given.

Motivated by the lack of theoretical understandings of the general PL extensions and learning algorithms, we ask the fol-

*Work done while at Rensselaer Polytechnic Institute.

	features	top-l	mixtures	identifiability	consistency
This work	✓	✓	✓	✓(Thm. 1, 2, Coro. 1, 2)	✓(Thm. 4)
[Tkachenko and Lauw, 2016]	✓	✓	✓		
[Yıldız *et al.*, 2020]	✓	✓			
[McFadden, 1973], [Berry *et al.*, 1995]	✓	top-1		✓	✓
[Grün and Leisch, 2008]	✓	top-1		✓	
[Schäfer and Hüllermeier, 2018]	✓	linear orders		✓	
[Zhao *et al.*, 2016], [Zhao *et al.*, 2018b], [Zhao and Xia, 2019]		linear orders	✓	✓	✓
[Chierichetti *et al.*, 2018]		✓	✓	✓	

Table 1: The model, identifiability, and consistency of this work compared with previous work.

lowing question. **When and how can preference learning be done for (mixtures of) RUMs with features for incomplete preferences?**

The question is highly challenging and the answer is not fully known even for many well-studied sub-cases, such as (non-mixture) PL with features for top-l orders and mixtures of standard PL (without features) for linear orders. In this paper, we provide the first answers to the question for PL with features for top-l orders by characterizing their *identifiability*, *consistency*, and *sample complexity* of their MLEs. We also provide the first generic identifiability result for its mixture models as well as more general RUMs with features.

Identifiability is a fundamental property of statistical models that is important to explainability and trustworthiness of decisions, which are particularly relevant in preference learning scenarios [Gormley and Murphy, 2008]. Identifiability requires that different parameters of the model lead to different distributions over data. If a model is non-identifiable, then sometimes the explanations of the learned parameters and corresponding decisions can be provably wrong, because there may exist another parameter that fits the data equally well, yet whose explanation and corresponding decisions are completely different. See Example 1 for an illustrative example. Additionally, the identifiability of a model is necessary for any algorithm to be consistent or have finite sample complexity.

Our Contributions. In this paper, we provide the first theoretical characterizations of identifiability for mixtures of $k \geq 1$ PLs with features for top-l orders, denoted by k-PL$_{\mathcal{X}}$-TO, where $\mathcal{X}$ is the feature matrix. We note that in k-PL$_{\mathcal{X}}$-TO, each agent submits a top-l order for possibly different l.

Identifiability. We provide necessary and sufficient conditions for PL$_{\mathcal{X}}$-TO and its special case called the bilinear model, denoted by PL$_{Y,Z}$-TO, to be identifiable in Theorem 1 and Corollary 1, respectively. Even though k-PL$_{\mathcal{X}}$-TO is not identifiable for any $k \geq 2$, we provide a sufficient condition in Theorem 2 for a parameter k-PL$_{\mathcal{X}}$-TO to be identifiable, which leads to the *generically identifiability* of 2-PL$_{\mathcal{X}}$-TO in Corollary 2. It suggests that identifiability may not be a practical concern if the condition is satisfied. We also characterize identifiability of RUMs with features in the appendix.

Strict concavity and sample complexity of MLE. We provide a necessary and sufficient condition for the MLE of PL$_{\mathcal{X}}$-TO to be strictly concave in Theorem 3 and bound on the RMSE of MLE of PL$_{\mathcal{X}}$-TO in Theorem 4, which implies the consistency of MLE and a sample complexity bound.

Our experiments on synthetic data demonstrate the performance of MLE and the tradeoffs between statistical efficiency and computational efficiency when learning from different top-l preferences for PL$_{\mathcal{X}}$-TO. For k-PL$_{\mathcal{X}}$-TO, we propose an EM algorithm and show the prediction power of different configurations of k-PL$_{\mathcal{X}}$-TO.

Related Work and Discussions. Table 1 summarizes related works on PL and its extensions that are close to ours. As discussed above, there is a large literature on PL and its extensions in each of the three dimensions, yet no theoretical result is known even for the special non-mixture case PL$_{\mathcal{X}}$-TO for top-l orders in general. [Tkachenko and Lauw, 2016] is the only previous work we are aware of that tackles k-PL$_{\mathcal{X}}$-TO, which does not provide theoretical guarantees. Even for PL$_{\mathcal{X}}$-TO, we are only aware of another recent paper [Yıldız *et al.*, 2020], which proposed an ADMM-based algorithm for computing the MLE, but it is unclear whether their algorithm converges to the ground truth because the consistency of MLE was unknown, which is a direct corollary of Theorem 4. To the best of our knowledge, our identifiability results (Theorems 1, 2, Corollaries 1, 2) and the RMSE bound (Theorem 4) are the first for (k-)PL$_{\mathcal{X}}$-TO even for linear orders ($l = m - 1$).

Mixtures of PLs with features for choice data (top-1) are well studied and can be dated back to the classical conditional logit model [McFadden, 1973] and the BLP model [Berry *et al.*, 1995]. PL with bilinear features, which is a special case of PL$_{\mathcal{X}}$-TO, has been studied in the literature [Azari Soufiani *et al.*, 2013; Schäfer and Hüllermeier, 2018; Zhao *et al.*, 2018a]. There is a large literature on standard PL (without features) and its mixture models [Hunter, 2004; Soufiani *et al.*, 2013; Negahban *et al.*, 2017; Maystre and Grossglauser, 2015; Khetan and Oh, 2016; Zhao and Xia, 2018; Gormley and Murphy, 2008; Oh and Shah, 2014; Zhao *et al.*, 2016; Chierichetti *et al.*, 2018; Zhao and Xia, 2019; Liu *et al.*, 2019]. Our setting is more general.

Recently, Schäfer and Hüllermeier (2018, Proposition 1) provided a sufficient condition for PL with bilinear features to be identifiable. However, their result is flawed due to the missing conditions on the ranks of feature matrices. See Exam-

ple 3 for more details. Our Corollary 1 provides a necessary and sufficient condition for the identifiability of more general models. Zhao *et al.* (2016) characterized the conditions on k (the number of components in the mixture model) and m (the number of alternatives) for mixtures of standard PLs to be (non-)identifiable. Zhao and Xia (2019) characterized (non-)identifiability of mixtures of PLs (without features) for structured partial orders. Grün and Leisch (2008) characterized conditions for mixtures of multinomial logit models to be identifiable. These results do not subsume our results because the model studied in this paper is more general.

Consistency of MLE was proven for the conditional logit model [McFadden, 1973] assuming that each agent provides multiple choice data. Or equivalently, agents with the same features are repeatedly observed. This assumption may not hold in the big data era, where the feature space can be extremely large and it is unlikely that agents would have identical features. Our proof of the RMSE bound on MLE, which implies consistency, tackles exactly this case and is inspired by the proof of [Khetan and Oh, 2016, Theorem 8] for standard (non-mixture) PL. Unlike the standard PL where the Hessian matrix is negative semidefinite with at least one zero eigenvalue [Khetan and Oh, 2016], the Hessian matrix of the model studied in this paper does not have zero eigenvalues. Due to this difference, the techniques in proving the sample complexity bound in [Khetan and Oh, 2016] cannot be directly extended to our setting.

Due to the space constraint, we focus on Plackett-Luce model in the main paper and defer results on RUMs to the appendix.

2 Preliminaries

Let $\mathcal{A} = \{a_1, \ldots, a_m\}$ denote the set of m alternatives. Let $\{1, \ldots, n\}$ denote the set of n agents. Given an agent j, each alternative is characterized by a column vector of $d \geq 1$ features $\vec{x}_{ji} \in \mathbb{R}^d$. For any $r = 1, \ldots, d$, let $x_{ji,r}$ denote the rth feature of $\vec{x}_{ji}$. A linear order, which is a transitive, antisymmetric, and total binary relation, is denoted by $R = a_{i_1} \succ \ldots \succ a_{i_m}$, where $a_{i_1} \succ a_{i_2}$ means that the agent prefers a_{i_1} over a_{i_2}. Let $\mathcal{L}(\mathcal{A})$ denote the set of all linear orders. A ranked top-l (top-l for short) order has the form $O = a_{i_1} \succ a_{i_2} \ldots \succ a_{i_l} \succ$ others. It is easy to see that a linear order is a special top-l order with $l = m - 1$. Let $\mathcal{T}(\mathcal{A})$ denote the set of all top-l orders for all $l \in \{1, 2, \ldots, m-1\}$. An l-way order has the form $O = a_{i_1} \succ a_{i_2} \ldots \succ a_{i_l}$. Let $\mathcal{I}(\mathcal{A})$ denote the set of all l-way orders for all $l \in \{2, \ldots, m\}$.

Definition 1 (Plackett-Luce model (PL))**.** *The parameter space is $\Theta = \mathbb{R}^m$. The sample space is $\mathcal{L}(\mathcal{A})^n$. Given a parameter $\vec{\theta} \in \Theta$, the probability of any linear order $R = a_{i_1} \succ a_{i_2} \succ \ldots \succ a_{i_m}$ is $\Pr_{PL}(R|\vec{\theta}) = \prod_{p=1}^{m-1} \frac{\exp(\theta_{i_p})}{\sum_{q=p}^{m} \exp(\theta_{i_q})}$.*

It follows that the marginal probability for any top-l order $R = a_{i_1} \succ a_{i_2} \ldots \succ a_{i_l} \succ$ others is $\Pr_{\text{PL}}(R|\vec{\theta}) = \prod_{p=1}^{l} \frac{\exp(\theta_{i_p})}{\sum_{q=p}^{m} \exp(\theta_{i_q})}$. In the literature, a normalization constraint on $\vec{\theta}$, e.g. $\sum_i \theta_i = 0$, is often required to make the model identifiable. In this paper we do put such a constraint since it is more convenient to extend the current definition to PL with features.

Let $X_j = [\vec{x}_{j1}, \ldots, \vec{x}_{jm}] \in \mathbb{R}^{d \times m}$ denote the feature matrix for agent j, and let $\mathcal{X} = [X_1, \ldots, X_n] \in \mathbb{R}^{d \times mn}$ denote the feature matrix that concatenates the features for all agents.

Definition 2 (Plackett-Luce model with features ($\text{PL}_{\mathcal{X}}$))**.** *Let $\mathcal{X} \in \mathbb{R}^{d \times mn}$ denote a feature matrix. The parameter space is $\Theta = \mathbb{R}^d$. The sample space is $\mathcal{L}(\mathcal{A})^n$. For any parameter $\vec{\beta} \in \Theta$, the probability of any linear order $R_j = a_{i_1} \succ a_{i_2} \succ \ldots \succ a_{i_m}$ given by agent j is $\Pr_{PL_{\mathcal{X}}}(R_j|\vec{\beta}) = \prod_{p=1}^{m-1} \frac{\exp(\vec{\beta} \cdot \vec{x}_{ji_p})}{\sum_{q=p}^{m} \exp(\vec{\beta} \cdot \vec{x}_{ji_q})}$.*

We note that *all feature matrices are assumed given, i.e., not part of the parameter of any model in this paper.* PL with bilinear features [Azari Soufiani *et al.*, 2013] is a special case of $\text{PL}_{\mathcal{X}}$, where each agent $j \in \{1, \ldots, n\}$ is characterized by a column feature vector $\vec{y}_j \in \mathbb{R}^L$ and each alternative $a_i \in \mathcal{A}$ is characterized by a column feature vector $\vec{z}_i \in \mathbb{R}^K$. We note that for any $i \in \{1, \ldots, m\}$, $\vec{z}_i$ is the same across all agents. Let $Y = [\vec{y}_1, \ldots, \vec{y}_n] \in \mathbb{R}^{L \times n}$ denote the agent feature matrix and $Z = [\vec{z}_1, \ldots, \vec{z}_m] \in \mathbb{R}^{K \times m}$ denote the alternative feature matrix.

Definition 3 (Plackett-Luce model with bilinear features $\text{PL}_{Y,Z}$)**.** *Let $Y \in \mathbb{R}^{L \times n}$ denote an agent feature matrix and let $Z \in \mathbb{R}^{K \times m}$ denote an alternative feature matrix. The parameter space consists of matrices $\Theta = \mathbb{R}^{K \times L}$. The sample space is $\mathcal{L}(\mathcal{A})^n$. Given a parameter $B \in \Theta$, the probability of any linear order $R_j = a_{i_1} \succ a_{i_2} \succ \ldots \succ a_{i_m}$ given by agent j is $\Pr_{PL_{Y,Z}}(R_j|B) = \prod_{p=1}^{m-1} \frac{\exp(\vec{z}_{i_p}^{\top} B \vec{y}_j)}{\sum_{q=p}^{m} \exp(\vec{z}_{i_q}^{\top} B \vec{y}_j)}$.*

$\text{PL}_{Y,Z}$ can be viewed as a special case of $\text{PL}_{\mathcal{X}}$ by letting $\mathcal{X} = Y \otimes Z$ and vectorizing B accordingly.

Definition 4 (k-PL)**.** *For any $k \geq 1$, the mixture of k Plackett-Luce models is defined as follows. The sample space is $\mathcal{L}(\mathcal{A})^n$. The parameter space has two parts. The first part is the* mixing coefficients *$\vec{\alpha} = (\alpha_1, \cdots, \alpha_k)$ with $\vec{\alpha} \geq 0$ and $\vec{\alpha} \cdot \vec{1} = 1$. The second part is $\left(\vec{\theta}^{(1)}, \cdots, \vec{\theta}^{(k)}\right)$, where $\vec{\theta}^{(r)} \in \Theta$ is the parameter of the r-th PL component. The probability of any linear order R is $\Pr_{k\text{-}PL}(R|\vec{\theta}) = \sum_{r=1}^{k} \alpha_r \Pr_{PL}(R|\vec{\theta}^{(r)})$.*

Definition 5 (Mixtures of k Plackett-Luce models with features (k-$\text{PL}_{\mathcal{X}}$))**.** *Let $\mathcal{X} \in \mathbb{R}^{d \times mn}$ denote the feature matrix. The parameter space Θ has two parts. The first part is the vector of mixing coefficients $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ and the second part is $(\vec{\beta}^{(1)}, \vec{\beta}^{(2)}, \ldots, \vec{\beta}^{(k)})$, where for $r = 1, \ldots, k$, $\vec{\beta}^{(r)} = \{\beta_i^{(r)} | 1 \leq i \leq d\}$. The sample space is $\mathcal{L}(\mathcal{A})^n$. Given a parameter $\vec{\theta} \in \Theta$, the probability of any linear order $R_j = [a_{i_1} \succ a_{i_2} \succ \ldots \succ a_{i_m}]$ given by agent j is $\Pr_{k\text{-}PL(\mathcal{X})}(R_j|\vec{\theta}) = \sum_{r=1}^{k} \alpha_r \Pr_{PL_{\mathcal{X}}}(R_j|\vec{\beta}^{(r)})$.*

$\text{PL}_{\mathcal{X}}$ can be viewed as a special case of k-$\text{PL}_{\mathcal{X}}$ where $k = 1$.

Definition 6 (Identifiability)**.** *Let $\mathcal{M} = \{\Pr(\cdot|\vec{\theta}) : \vec{\theta} \in \Theta\}$ be a statistical model, where Θ is the parameter space and $\Pr(\cdot|\vec{\theta})$ is the distribution over the sample space associated with $\vec{\theta} \in \Theta$. We say that a parameter $\vec{\theta} \in \Theta$ is* identifiable *in*

$\mathcal{M}$, if for any $\vec{\gamma} \in \Theta$ with $\vec{\gamma} \neq \vec{\theta}$, we have $\Pr(\cdot|\vec{\theta}) \neq \Pr(\cdot|\vec{\gamma})$. *$\mathcal{M}$ is* identifiable *if all its parameters are identifiable.*

The following example shows that non-identifiability of a model can lead to unavoidable untrustworthy interpretations and decisions.

Example 1. Suppose an automobile manufacturer is using $\text{PL}_{Y,Z}$ to learn consumers' preferences over car models. For simplicity suppose there are two agents $\{1, 2\}$ and two alternatives (car models) $\{a_1, a_2\}$. $Y = [y_1, y_2] = [0.5, 1]$, where each agent is represented by her normalized income (0.5 for the first agent and 1 for the second agent). $Z = [\vec{z}_1, \vec{z}_2] = \begin{bmatrix} 0.6 & 1 \\ 0.2 & 0.5 \end{bmatrix}$, where each car is represented by its normalized price (0.6 for the first car and 1 for the second car) and its normalized miles per gallon (0.2 for the first car and 0.5 for the second car).

Let $B = [-1, 8/3]^\top$ and $B' = [1, 0]^\top$. We show that B and B' correspond to exactly the same distribution over the two agents' preferences. In fact, it is not hard to verify that $\vec{z}_1^\top B y_1 - \vec{z}_2^\top B y_1 = z_1^\top B' y_1 - \vec{z}_2^\top B' y_1 = -0.2$ and $\vec{z}_1^\top B y_2 - \vec{z}_2^\top B y_2 = \vec{z}_1^\top B' y_2 - \vec{z}_2^\top B' y_2 = -0.4$. Therefore, $\Pr_{\text{PL}_{Y,Z}}(R_1 = a_1 \succ a_2|B) = \frac{1}{1+\exp(\vec{z}_2^\top B y_1 - \vec{z}_1^\top B y_1)} = \frac{1}{1+\exp(\vec{z}_2^\top B' y_1 - \vec{z}_1^\top B' y_1)} = \Pr_{\text{PL}_{Y,Z}}(R_1 = a_1 \succ a_2|B')$. Other probabilities can be calculated similarly.

Therefore, it is impossible for any statistical method to distinguish B from B'. This may not be a big concern if the company uses the learned model to predict the preferences of new customers, as both B and B' would give the same prediction. However, the first components of B and B' have opposite interpretations. The first component of B being positive is often interpreted as the existence of a negative correlation between an agent's income and the car's price, i.e. richer people prefer cheaper cars. The interpretation of the first component of B' is on the opposite. This makes the interpretation and any decision based on it untrustworthy. □

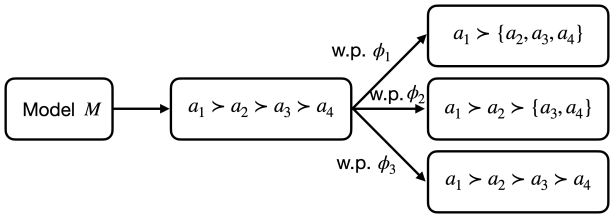

Figure 1: Illustration of model M-TO with the parameter $\vec{\phi}$, where M is a model that generates linear orders. M-TO allows users to report top-l orders of different l's.

3 Models and Their Identifiability for Top-l Orders

Modeling and learning from different partial orders is desirable in the scenarios where users provide different structures of partial orders [Zhao and Xia, 2019]. Following Zhao and Xia (2019), we extend $\text{PL}_{\mathcal{X}}$ to a model that generates top-l partial orders by introducing a new parameter $\vec{\phi} = (\phi_1, \ldots, \phi_{m-1})$,

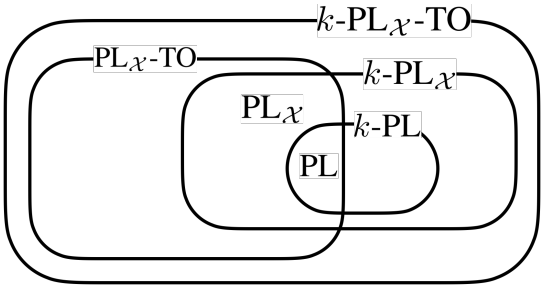

Figure 2: Relations between models in this paper in Venn diagram. k-PL$_{\mathcal{X}}$-TO is the most general model, subsuming all the other models. The two largest submodels of k-PL$_{\mathcal{X}}$-TO are PL$_{\mathcal{X}}$-TO and k-PL$_{\mathcal{X}}$, whose intersection is PL$_{\mathcal{X}}$. PL is the smallest model in this diagram, lying at the intersection of PL$_{\mathcal{X}}$-TO and k-PL.

where $\sum_{i=1}^{m-1} \phi_i = 1$. $\vec{\phi}$ can be viewed as a distribution over $\{1, 2, \ldots, m-1\}$ that represents the probability for the agent to report a top-l order, where $1 \leq l \leq m-1$. Any model defined in the previous section can be extended to a model that generates a top-l order R with probability $\Pr_{\mathcal{M}\text{-TO}}(R|\vec{\theta}, \vec{\phi}) = \phi_l \Pr_{\mathcal{M}}(R|\vec{\theta})$, as is illustrated in Figure 1. For example, $\text{PL}_{\mathcal{X}}$ can be extended to PL$_{\mathcal{X}}$-TO, formally defined as follows.

Definition 7 (PL$_{\mathcal{X}}$-TO)**.** *Let $\mathcal{X} \in \mathbb{R}^{d \times mn}$ denote the feature matrix. The parameter space is $\Theta = \mathbb{R}^d \times \{\vec{\phi} \in \mathbb{R}_{\geq 0}^{m-1} : \vec{\phi} \cdot \vec{1} = 1\}$. The sample space is $\mathcal{T}(\mathcal{A})^n$. Given a parameter $(\vec{\beta}, \vec{\phi})$, the probability of any top-l order $O_j = a_{i_1} \succ a_{i_2} \succ \ldots \succ a_{i_l} \succ$ others given by agent j is $\Pr_{PL_{\mathcal{X}}\text{-}TO}(O_j|\vec{\beta}, \vec{\phi}) = \phi_l \Pr_{PL_{\mathcal{X}}}(O_j|\vec{\beta})$, where $\Pr_{PL_{\mathcal{X}}}(O_j|\vec{\beta})$ is the marginal probability of O_j under PL$_{\mathcal{X}}$ given $\vec{\beta}$.*

Again, $\mathcal{X}$ is assumed given and not part of the parameter of PL$_{\mathcal{X}}$-TO. PL$_{\mathcal{X}}$ is a submodel of PL$_{\mathcal{X}}$-TO (where $\phi_{m-1} = 1$). $\text{PL}_{Y,Z}$-TO and k-PL$_{\mathcal{X}}$-TO can be defined similarly, see Appendix A for their formal definitions. The relations between different models mentioned in this paper are shown in the Venn diagram in Figure 2 ($\text{PL}_{Y,Z}$ and $\text{PL}_{Y,Z}$-TO are omitted for simplicity).

As was illustrated in Example 1, identifiability is important if one wants to interpret the learned parameter. For the rest of this section, we focus on identifiability of PL$_{\mathcal{X}}$-TO and k-PL$_{\mathcal{X}}$-TO.

To characterize the identifiability of PL extensions with features, for each $j \leq n$, we first define agent j's *normalized* feature matrix, denoted by $\text{norm}(X_j)$.

$$\text{norm}(X_j) = [\vec{x}_{j2} - \vec{x}_{j1}, \vec{x}_{j3} - \vec{x}_{j1}, \ldots, \vec{x}_{jm} - \vec{x}_{j1}]. \quad (1)$$

We then define $\text{Norm}(\mathcal{X})$.

$$\text{Norm}(\mathcal{X}) = [\text{norm}(X_1), \text{norm}(X_2), \ldots, \text{norm}(X_n)] \quad (2)$$

In words, $\mathcal{X}$ is normalized by using the feature vector of a_1 as the baseline. Our results still hold if another alternative is used as the baseline. We now present our first identifiability theorem.

Theorem 1. *For any $\mathcal{X}$, PL$_{\mathcal{X}}$-TO is identifiable if and only if Norm($\mathcal{X}$) has full row rank.*

The full proof can be found in Appendix G.2.

Example 2. *Consider a PL$_\mathcal{X}$ (a special case of PL$_\mathcal{X}$-TO), whose feature matrix $\mathcal{X}$ has three rows $\vec{r}_1$, $\vec{r}_2$, and $\vec{r}_3$, where $\vec{r}_1 + \vec{r}_2 = \vec{r}_3$. Therefore, Norm$(X)$ does not have full row rank. Let $\vec{\beta} = [\beta_1, \beta_2, \beta_3]^\top$ be the ground truth parameter. We construct $\vec{\beta}' = [\beta_1 + \beta_3, \beta_2 + \beta_3, 0]^\top$. Then it is easy to see $\vec{\beta}^\top \cdot \mathcal{X} = \vec{\beta}'^\top \cdot \mathcal{X}$, which further means for any order R, we have $\Pr_{PL_\mathcal{X}}(R|\vec{\beta}) = \Pr_{PL_\mathcal{X}}(R|\vec{\beta}')$ by Definition 2. This means this PL$_\mathcal{X}$ is not identifiable.*

Theorem 1 can be applied to characterize the identifiability for PL with bilinear features as in the following corollary.

Corollary 1. *For any model PL$_{Y,Z}$-TO, where $Y \in \mathbb{R}^{L\times n}$ and $Z \in \mathbb{R}^{K\times m}$, PL$_{Y,Z}$-TO is identifiable if and only if both Y and norm(Z) have full row rank.*

The formal proof can be found in Appendix G.3.

The full row rank condition in Theorem 1 and Corollary 1 is mild as long as n is not too small, which is the case in many real-world applications. For example, the full row rank condition holds on the sushi dataset used in our real-world experiment. In Appendix F.4, we show that the probability of violating the full row rank condition becomes very small when sampling $n = 10$ agents from the sushi dataset, and this probability decays exponentially as n increases.

In the next example, we show that the sufficient conditions for PL$_{Y,Z}$-TO to be identifiable by Schäfer and Hüllermeier (2018, Proposition 1) is unfortunately flawed.

Example 3. *Continuing Example 1, it is not hard to verify that no agent feature or alternative feature is a constant, which implies that PL$_{Y,Z}$ is identifiable according to [Schäfer and Hüllermeier, 2018, Proposition 1]. However, as we showed in Example 1, PL$_{Y,Z}$ is not identifiable. It's easy to see that norm(Z) has two rows but only one column, which does not have full row rank.*

Identifiability of mixtures of PLs with features is at least as challenging as the identifiability of mixtures of standard PLs, which is still an open problem for any $k \geq 3$. In the following theorem, we provide a sufficient condition for a parameter in k-PL$_\mathcal{X}$-TO to be identifiable.

Theorem 2. *If k-PL is identifiable, then for any $\mathcal{X}$ such that Norm$(\mathcal{X})$ has full row rank, any parameter $(\vec{\alpha}, \vec{\beta}, \vec{\phi})$ with $\phi_{m-1} > 0$ is identifiable in k-PL$_\mathcal{X}$-TO.*

The proof is done by contradiction. The full proof is provided in Appendix G.4.

Corollary 2. *For any $\mathcal{X}$ such that Norm$(\mathcal{X})$ has full row rank, 2-PL$_\mathcal{X}$-TO over four or more alternatives is identifiable modulo label switching.*

4 MLE of PL$_\mathcal{X}$-TO and Its Consistency

Let $P = (O_1, \ldots, O_n)$ denote the input data, where for each $j \leq n$, O_j is a top-l_j order. Let $\mathcal{X} \in \mathbb{R}^{d\times mn}$ denote the feature matrix. MLE of PL$_\mathcal{X}$-TO computes the parameter that maximize the following log likelihood function:

$$LL(P|\vec{\phi}, \vec{\beta}) = \sum_{j=1}^{n} \ln \Pr_{\text{PL}_\mathcal{X}\text{-TO}}(O_j|\vec{\beta}) = \sum_{j=1}^{n} (\ln \phi_{l_j} + \ln \Pr_{\text{PL}_\mathcal{X}}(O_j|\vec{\beta}))$$

Note that $\vec{\phi}$ and $\vec{\beta}$ parameter are separated in the log likelihood function, we can compute them separately as follows.

$$\vec{\phi}^* = \arg\max_{\vec{\phi}} \sum_{j=1}^{n} \ln \phi, \text{s.t.} \sum_{l=1}^{m-1} \phi_l = 1 \tag{3}$$

$$\vec{\beta}^* = \arg\max_{\vec{\beta}} \sum_{j=1}^{n} \ln \Pr_{\text{PL}_\mathcal{X}}(O_j|\vec{\beta}) \tag{4}$$

$\vec{\phi}$ can be easily computed by counting the frequencies of each top-l order. The main challenge is to accurately estimate the $\vec{\beta}$ part, which is the main focus of the rest of this section.

The following theorem provides a necessary and sufficient condition for the objective function in (4) to be strictly concave, which turns out to be the same condition for the identifiability of PL$_\mathcal{X}$-TO. Strict concavity is desirable because, combined with boundedness, it guarantees the convergence of MLE.

Theorem 3. *For any PL$_\mathcal{X}$-TO and any data $P = (O_1, \ldots, O_n)$, the log likelihood function in* (4) *is strictly concave if and only if Norm$(\mathcal{X})$ has full row rank.*

The full proof can be found in Appendix G.5.

We now introduce an assumption to guarantee the boundedness of MLE for given data P. Boundedness is important for consistency because a strictly concave function may not converge if it is unbounded. For any $j \leq n$, $i_1 \leq m$, and $i_2 \leq m$, we define ξ_{j,i_1i_2} as follows.

$$\xi_{j,i_1i_2} = \begin{cases} 1 & \text{if agent } j \text{ prefers } a_{i_1} \text{ over } a_{i_2} \\ -1 & \text{if agent } j \text{ prefers } a_{i_2} \text{ over } a_{i_1} \\ 0 & \text{if agent } j\text{'s preference between } a_{i_1} \\ & \text{and } a_{i_2} \text{ is not available} \end{cases}$$

Assumption 1. *Let $\mathcal{X} \in \mathbb{R}^{d\times mn}$ denote a feature matrix and let P denote the data. For any $r \leq d$, there exist $j_1, j_2 \in \{1, \ldots, n\}$ with $j_1 \neq j_2$ and $i_1, i_2 \in \{1, \ldots, m\}$ with $i_1 \neq i_2$ such that $\xi_{j_1,i_1i_2}\xi_{j_2,i_1i_2}(x_{j_1i_1,r} - x_{j_1i_2,r})(x_{j_2i_1,r} - x_{j_2i_2,r}) < 0$.*

At a high level, Assumption 1 is a mild condition that requires sufficient diversity in agents' preferences, which mirrors Hunter's assumption for PL [Hunter, 2004, Assumption 1].

The following lemma shows that Assumption 1 is sufficient for MLE to be bounded.

Lemma 1. *For any PL$_\mathcal{X}$-TO and data P, if Assumption 1 holds then the MLE in* (4) *is bounded.*

The proof is provided in Appendix G.6. Finally, the following theorem provides a bound on the RMSE of MLE for PL$_\mathcal{X}$-TO given that Norm$(\mathcal{X})$ has full row rank.

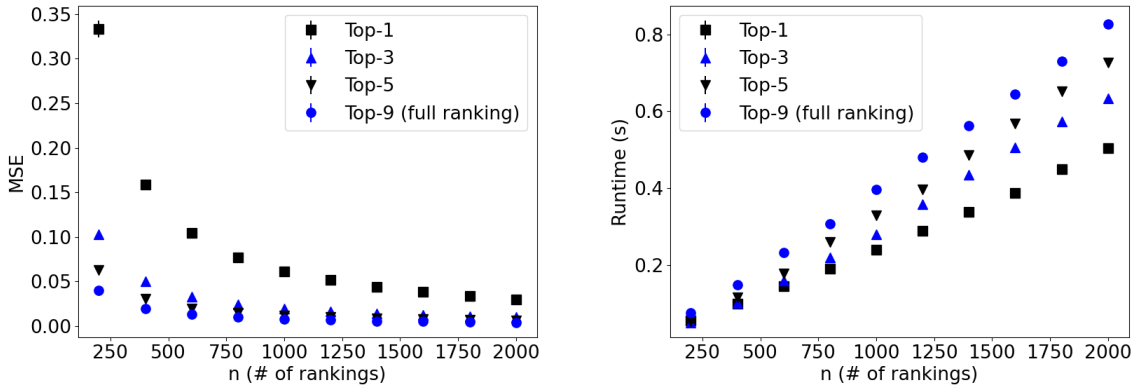

Figure 3: MSE (left) and Runtime (right) with 95% confidence intervals for MLE on PL$_\mathcal{X}$-TO given top-1 only, top-3 only, top-5 only, and top-9 (full rankings) over 2000 trials. Results for Top-7 are very close to those for top-9, and therefore omitted.

Theorem 4. *Given any PL$_\mathcal{X}$-TO over m alternatives and n agents with the feature matrix $\mathcal{X} \in \mathbb{R}^{d\times mn}$. Define $L(\vec{\beta}) = \frac{1}{n}\sum_{j=1}^{n} \ln \Pr_{PL_\mathcal{X}}(O_j|\vec{\beta})$, which is $\frac{1}{n}$ of the objective function in* (4)*. Let $H(\vec{\beta})$ denote the Hessian matrix of $L(\vec{\beta})$ and $\lambda_1(\vec{\beta})$ be the smallest eigenvalue of $-H(\vec{\beta})$. Let $\vec{\beta}_0$ denote the ground truth parameter and $\vec{\beta}^*$ denote the estimated parameter that is computed using* (4)*. Define $\lambda_{\min} = \min_{0\le\sigma\le1} \lambda_1(\sigma\vec{\beta}^* + (1-\sigma)\vec{\beta}^0)$.*

If Norm($\mathcal{X}$) has full row rank and Assumption 1 holds, then for any $0 < \delta < 1$, with probability $1-\delta$,

$$||\vec{\beta}^* - \vec{\beta}_0||_2 \le \frac{\sqrt{8(m-1)^2c^2d\ln(\frac{2d}{\delta})}}{\lambda_{\min}\sqrt{n}}, \qquad (5)$$

where c is the difference between the largest and the smallest entries of $\mathcal{X}$.

The full proof is in Appendix G.7.

The RMSE (root-mean-square-error) bound given by Theorem 4 is not tight since we make no assumptions on the distribution of features. While Theorem 4 does provide insights on convergence of MLE:

1. *Consistency*: as n increases, RMSE $||\vec{\beta}_0 - \vec{\beta}^*||_2$ decreases at the rate of $\frac{1}{\sqrt{n}}$. When n approaches infinity, RMSE approaches 0.
2. *Sample complexity*: for any $\epsilon > 0$, $0 < \delta < 1$, $\Pr(||\vec{\beta}^* - \vec{\beta}_0|| \le \epsilon) \ge 1-\delta$ when $n \ge \frac{8(m-1)^2c^2d\ln\frac{2d}{\delta}}{\lambda_{\min}^2\epsilon^2}$. This is obtained by letting $\epsilon \ge \frac{\sqrt{8(m-1)^2c^2d\ln(\frac{2d}{\delta})}}{\lambda_{\min}\sqrt{n}}$.
3. *Approximation of $\lambda_{\min}$*: in practice, when the size of data is not too small, $\lambda_{\min}$ can be approximated by $\lambda_1(\vec{\beta}^*)$ because $\vec{\beta}^*$ approaches $\vec{\beta}^0$ as n increases. This gives a practical way of computing the RMSE bound when the ground truth is unknown.

5 Experiments

We show experiments on synthetic data in this section and provide additional experiments on mixture models and on real-world data in the appendix.

Setup. Fix $m = 10$ and $d = 10$. For each agent and each alternative, the feature vector is generated in $[-1, 1]$ uniformly at random. Each component in $\vec{\beta}$ is generated uniformly at random in $[-2, 2]$. MLE for PL$_\mathcal{X}$-TO was implemented in MATLAB with the built-in `fminunc` function and tested on a Ubuntu Linux server with Intel Xeon E5 v3 CPUs each clocked at 3.50 GHz. We use mean squared error (MSE) and runtime to measure the statistical efficiency and computational efficiency of algorithms, respectively. Results are shown in Figure 3. All values are averaged over 2000 trials.

Observations. Figure 3 shows the performance of MLE for PL$_\mathcal{X}$-TO. We observe that MSE decreases as the number of agents increases, which demonstrates consistency of MLE for PL$_\mathcal{X}$-TO. Moreover, learning from top-l preferences with different l values provides tradeoffs between statistical efficiency and computational efficiency.

6 Summary and Future Work

We provide the first set of theoretical results on the identifiability of mixtures of PL with features for top-l preferences. We also identify conditions for the MLE of PL$_\mathcal{X}$-TO to be consistent, and propose an EM algorithm to handle general k-PL$_\mathcal{X}$-TO. In the full version of this paper [Zhao *et al.*, 2022], we provide a generalized Rank-Breaking then Composite Marginal Likelihood algorithm for learning RUMs beyond PL from incomplete preferences and show its performance on synthetic data. [Zhao *et al.*, 2022] also includes additional experiments on mixture models as well as missing proofs. Generic identifiability and efficient algorithms for k-PL$_\mathcal{X}$-TO are natural questions for future work.

Acknowledgements

We thank anonymous reviewers for helpful comments. Lirong Xia is supported by NSF #1453542 and a gift fund from Google. Ao Liu is supported by RPI-IBM AIHN scholarship.

References

[Azari Soufiani *et al.*, 2012] Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. Random utility theory for social choice.

In *Proceedings of Advances in Neural Information Processing Systems*, pages 126–134, 2012.

[Azari Soufiani *et al.*, 2013] Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. Preference Elicitation For General Random Utility Models. In *Proceedings of Uncertainty in Artificial Intelligence*, 2013.

[Baltrunas *et al.*, 2010] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 119–126, 2010.

[Berry *et al.*, 1995] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, pages 841–890, 1995.

[Chen *et al.*, 2013] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 193–202, 2013.

[Chierichetti *et al.*, 2018] Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. Learning a mixture of two multinomial logits. In *International Conference on Machine Learning*, pages 961–969, 2018.

[Dwork *et al.*, 2001] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

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

[Grün and Leisch, 2008] Bettina Grün and Friedrich Leisch. Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of classification*, 25(2):225–247, 2008.

[Hüllermeier *et al.*, 2008] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16-17):1897–1916, 2008.

[Hunter, 2004] David R Hunter. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406, 2004.

[Khetan and Oh, 2016] Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. *The Journal of Machine Learning Research*, 17(1):6668–6721, 2016.

[Liu *et al.*, 2019] Ao Liu, Zhibing Zhao, Chao Liao, Pinyan Lu, and Lirong Xia. Learning plackett-luce mixtures from partial preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4328–4335, 2019.

[Liu, 2009] Tie-Yan Liu. Learning to rank for information retrieval. *Information Retrieval*, 3(3):225–331, 2009.

[Luce, 1959] Duncan R. Luce. *Individual choice behavior.* John Wiley, 1959.

[Mao *et al.*, 2013] Andrew Mao, Ariel D Procaccia, and Yiling Chen. Better human computation through principled voting. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[Maystre and Grossglauser, 2015] Lucas Maystre and Matthias Grossglauser. Fast and accurate inference of plackett–luce models. In *Advances in neural information processing systems*, pages 172–180, 2015.

[McFadden, 1973] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In *Frontiers of Econometrics*, pages 105–142, New York, NY, 1973. Academic Press.

[Mollica and Tardella, 2017] Cristina Mollica and Luca Tardella. Bayesian plackett–luce mixture models for partially ranked data. *Psychometrika*, 82(2):442–458, 2017.

[Negahban *et al.*, 2017] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017.

[Oh and Shah, 2014] Sewoong Oh and Devavrat Shah. Learning mixed multinomial logit model from ordinal data. In *Advances in Neural Information Processing Systems*, pages 595–603, 2014.

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

[Schäfer and Hüllermeier, 2018] Dirk Schäfer and Eyke Hüllermeier. Dyad ranking using plackett–luce models based on joint feature representations. *Machine Learning*, 107(5):903–941, 2018.

[Soufiani *et al.*, 2013] Hossein Azari Soufiani, William Chen, David C Parkes, and Lirong Xia. Generalized method-of-moments for rank aggregation. In *Advances in Neural Information Processing Systems*, pages 2706–2714, 2013.

[Tkachenko and Lauw, 2016] Maksim Tkachenko and Hady W Lauw. Plackett-luce regression mixture model for heterogeneous rankings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 237–246, 2016.

[Train, 2009] Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition, 2009.

[Yıldız *et al.*, 2020] İlkay Yıldız, Jennifer Dy, Deniz Erdoğmuş, Jayashree Kalpathy-Cramer, Susan Ostmo, J Peter Campbell, Michael F Chiang, and Stratis Ioannidis. Fast and accurate ranking regression. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[Zhao and Xia, 2018] Zhibing Zhao and Lirong Xia. Composite marginal likelihood methods for random utility models. In *International Conference on Machine Learning*, pages 5922–5931, 2018.

[Zhao and Xia, 2019] Zhibing Zhao and Lirong Xia. Learning mixtures of plackett-luce models from structured partial orders. In *Advances in Neural Information Processing Systems*, pages 10143–10153, 2019.

[Zhao *et al.*, 2016] Zhibing Zhao, Peter Piech, and Lirong Xia. Learning mixtures of plackett-luce models. In *International Conference on Machine Learning*, pages 2906–2914, 2016.

[Zhao *et al.*, 2018a] Zhibing Zhao, Haoming Li, Junming Wang, Jeffrey Kephart, Nicholas Mattei, Hui Su, and Lirong Xia. A Cost-Effective Framework for Preference Elicitation and Aggregation. In *Proceedings of Uncertainty in Artificial Intelligence*, 2018.

[Zhao *et al.*, 2018b] Zhibing Zhao, Tristan Villamil, and Lirong Xia. Learning mixtures of random utility models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Zhao *et al.*, 2022] Zhibing Zhao, Ao Liu, and Lirong Xia. Learning mixtures of random utility models with features from incomplete preferences. *arXiv preprint arXiv:2006.03869*, 2022.