Mixed Membership Mallows Model

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose Mixed Membership Mallows Models (M4) to model noisy preferences of heterogenous and inconsistent users. In our model each Mallows component accounts for noisy preferences; an inconsistent user is modeled by means of a probabilistic mixture of latent Mallows components that are *shared* among all users; and a user-specific mixture models heterogeneity. We propose methods for estimating Mallows components from pairwise comparison data. The technical novelty of our approach is two-fold. We first establish an information-theoretic connection between M4 and topic models. Second, we prove that for a broad class of probabilistic mixtures and Mallows dispersion parameters M4 is inevitably approximately separable. This characterization leads us to propose algorithms based on convex geometry for estimating Mallows components. We prove asymptotic consistency, polynomial sample and computational complexity bounds for our estimates. As a by-product of our approach we also obtain the *first provably consistent and efficient algorithm* for learning special cases considered before.

1 Introduction

We propose *Mixed Membership Mallows Model* (M4) to capture the preference behavior of a diverse user-population who provide noisy and inconsistent comparisons [1–6] as seen in many applications such as restaurants check-ins, clicks from Yelp, movie purchases, and reviews from Netflix [1, 4, 7, 8] data. In these applications, each user can be influenced by multiple ranking factors to different extents at different times resulting in inconsistent and noisy behavior (see Fig 1). In addition, the number of comparisons available from each user is typically very small.

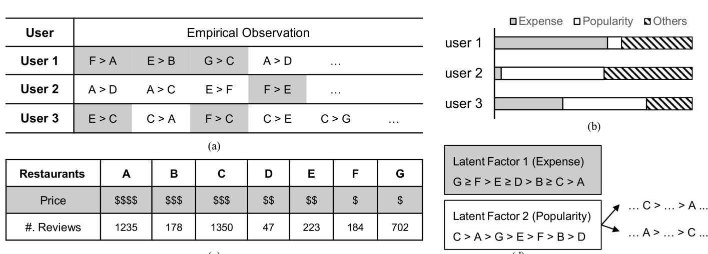


Figure 1: An illustration of how M4 models noisy preferences of heterogenous and inconsistent users. Say a set of ratings from Yelp for restaurants are obtained and anonymized from a local area (subplot (c)). Two example latent factors, "expense" and "popularity" (subplot (d)), influence the three users' behavior (subplot (a)), with different weights (subplot (b)). This models heterogeneity. $A > B$ means A is preferred over B. The shading in subplot (a) indicates the most-likely influencing factor of each observation using the same color coding as in other subplots. This accounts for inconsistency. A and C are very close in "Popularity" and both $C > A$ and $A > C$ are possible when influenced by the same "Popularity" factor which accounts for noise.

A key conceptual contribution of the proposed work is our M4 model. It leverages existing mixed membership ranking models[1, 2] and incorporates the popular Mallows model [5, 9] as shared latent factors. Each user comparison is modeled as a *probabilistic mixture* of a few latent Mallows components that are shared among the population. M4 thus subsumes the popular Mixture of Mallows model [4, 5, 10–12] as well as recent work on mixed membership ranking models [1] as special cases. In doing so it not only accounts for noise in each latent ranking factor, which has been a popular choice in the mixture of ranking models [e.g., 4, 5] but also incorporates user-specific mixed memberships (heterogeneity) which is demonstrably a better fit for real-world comparisons [1, 2].

Our technical contributions involve schemes that estimate Mallows components from noisy pairwise comparison data with provable guarantees on their performance for a broad class of mixture prior distributions. Informally,

Theorem 1. (Informal) *Except for a set with vanishing probability, all the reference rankings in M4 can be estimated (a) consistently as the number of users scales, (b) with polynomial computation complexity and sample complexity bounds.*

As a by-product of our approach we also obtain the *first provably consistent and efficient algorithm* to special cases considered before such as the popular mixture of Mallows models [4, 5, 10–12], whose theoretical guarantees are not yet clear except in some special cases [5].

Our approach relies on establishing an information-theoretic connection between M4 and topic-word matrix [13] by viewing distinct pairs as "words", comparisons of each user as a "document" and each latent Mallows component as a "topic". This connection leads us to a surprising finding, namely, the inevitability of approximate separability of Mallows components. Intuitively, approximate separability requires the existence of ordered pairs that have negligible probability in all-but-one of the Mallows components, i.e., the row entries concentrate predominantly in one column. Formally, we prove that most instances of the ranking matrix in M4, when appropriately sampled, are approximately separable when the number of items is large relative to the number of latent ranking factors.

This result leads us to a geometric approach that is inspired by the recent works in topic modeling with a separable latent structure [e.g., 14–19]. However, we cannot directly apply results from existing work since they rely on *exact* separabilty. In this context we non-trivially generalize the geometry induced by the *exact-separability* [e.g., 14, 16] to handle the *approximate-separability* property which holds for M4 (see Definition 1), and establish provable estimation guarantees (see Theorem 3).

While these technical advances are of independent interest and provide a framework for learning general mixed membership models, for concreteness, we focus on M4 models. Finally, we point out that our results only require the number of users to scale and allow the number of comparisons per user to be a small constant, which is well adapted to sparse real-world data.

Related Work: We describe several research topics that are related to proposed work.

Mixture of Rankings Models: The family of mixture of rankings models have demonstrated superior modeling power to capture a heterogeneous and noisy user-population in both full and partial rankings [6, 20]. Here each user is associated with *one* of the multiple latent ranking components and the population can be clustered into heterogeneous preference types. The popular mixture of Mallows models is closely related to our setting where the latent factors are Mallows components [4, 5, 10–12]. Approximation methods such as MCMC have been used for estimation from pairwise comparisons [4], full rankings [11], and in a non-parametric Bayes setting [12]. Only recently, [5] proposed a provable algorithm based on tensor decomposition that can handle a mixture of 2 Mallows components using the top-3 ranked items as the observations which is restrictive and impractical within the context of the target web-scale applications. We note that mixture of Bradley-Terry-Luce (BTL) models [6] and Plackett-Luce (PL) models [21] have also been studied.

Table 1: Comparison of M4 to closely related works.

Method	Observation	Ranking component	Prior	Consistency	Computation
M4	Pairwise	Mallows	General	Provable	Polynomial
[1]	Pairwise	Total Ranking	General	Provable	Polynomial
[2]	Full	Plackett-Luce	Dirichlet	Not Available	Not Available
[3]	Pairwise	Bradely-Terry-Luce	Dirichlet	Not Available	Not Available
[4]	Pairwise	Mallows	Mixture	Not Available	Not Available
[5]	Top-3 rank	Mallows	Mixture	Provable	Polynomial

Mixed Membership Ranking Models: M4 shares the same mixed membership modeling perspective as the recent efforts in [1–3]. [2] and [3] proposed to model the latent ranking factors using score-based PL and BTL models respectively. Approximation based methods are used in estimation with no theoretical guarantees. It is also unclear how the MCMC based approach in [2] can scale to the targeted web-scale applications. [1] proposed to model the latent ranking factors as total rankings and gave an algorithm with provable efficiency guarantees. It is closely related to M4 in the motivating geometry but with fundamental differences. Foremost, [1] is a *special case* of M4 by suppressing the dispersion of Mallows components to zero. Intuitively, while both [1] and M4 can capture the inconsistency as stemming from the influence of multiple latent factors, M4 can further account for the consequence of the randomness in each latent factor. In addition, our new results on approximate separability subsumes the exact separable geometry exploited in[1] as a special instance. Table 1 provides a summarized comparison of M4 with the closely related works.

Separable Topic Discovery: The recent work on consistent and efficient topic discovery with an *exact separability* property [14, 16] forms the starting point of our work. Our result allows the latent factors to be approximately separable, i.e., with a small but finite deviation from exact separability. Recent works in [17, 18] considered similar settings but require much stronger assumptions. [17] requires a significant portion of users to be influenced almost by only one of the latent factors. [18] requires a strict initialization and it is not clear how it can be achieved using only the observations. In contrast, we do not depend on initialization.

Rating-based Approaches: Considerable work in modeling user preferences and choices has focused on numerical and start ratings.[22, 23] The prevalent idea there is also to view the user ratings as being influenced by a small number of latent factors shared by the population [e.g., 8]. This modeling perspective is similar to M4 although it focuses on a different feature space.

2 Mixed Membership Mallows Models

This section introduces the Mixed Membership Mallows Model (M4) and the associated learning problem. It also discusses how M4 is related to other ranking models and an information-equivalent topic model.

Consider a population of M users in which, for simplicity, each user compares N pairs of items. Assume that items are numbered $1, \ldots, Q$. The result of the n-th comparison made by user m is an *ordered pair* of items $w_{m,n} = (i, j)$ if items i, j are compared and user m prefers i over j. Assume that the two items to be compared are sampled according to some distribution μ on all pairs with $\mu_{i,j} = \mu_{j,i} > 0$ being the probability of comparing items i and j. The ordered pairs produced by all M users can be represented using a $W \times M$ matrix $\mathbf{X}$ whose $W = Q(Q-1)$ rows correspond to all the ordered pairs (i, j) and M columns correspond to all the users. The number of times that user m compares and prefers item i over j is then given by $X_{(i,j),m}$.

Mallows Model A Mallows model for rankings is a pmf over permutations (rankings) σ over Q items. A Mallows pmf $p_{\mathrm{M}}(\sigma|\sigma_k, \phi_k)$ is parameterized by a *reference ranking* σ_k and a *dispersion parameter* $\phi_k \in [0, 1)$. Under this pmf, the probability of σ decays exponentially with its Kendall's tau distance[1] $d(\sigma, \sigma_k)$ to σ_k at a rate governed by ϕ_k [9]. Specifically, $p_{\mathrm{M}}(\sigma|\sigma_k, \phi_k) = \phi_k^{d(\sigma,\sigma_k)}/Z_k$ where Z_k is the normalization constant. The closer ϕ_k is to 1, the more spread the Mallows pmf is.[2]

M4 then views the ordered pairs produced by each user as a probabilistic mixture of K latent component Mallows pmfs which capture *heterogeneous* influencing factors. The preferences of M users are characterized by mixing weights $\boldsymbol{\theta}_m$'s. The pmf of $w_{m,n}$ conditioned on $\boldsymbol{\theta}_m$ is given by

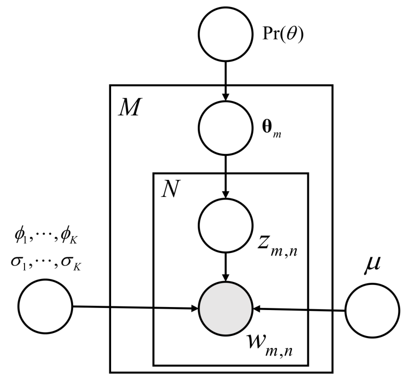
$$p(w_{m,n} = (i, j)|\boldsymbol{\theta}_m, \mu) = \mu_{i,j} \sum_{k=1}^{K} \sum_{\sigma:\ \sigma(i)<\sigma(j)} p_{\mathrm{M}}(\sigma|\sigma_k, \phi_k)\theta_{k,m} \tag{1}$$

In M4, since each user can be influenced by multiple latent factors to different extents, the comparisons produced by each users can potentially be *inconsistent*. In addition, the use of a Mallows model for each latent factor allows one to capture potential *randomness* in the outcomes of item comparisons for items that are very similar. This is because in a realization σ of each Mallows component, item pairs that are close in the reference ranking σ_k are more likely to be reversed in the realization due to nonzero dispersion ϕ_k. Figure 2 summarizes the generative process and its graphical representation.

Learning problem Given $\mathbf{X}$ and K, our primary objective is to develop an algorithmq that can *learn* the parameters of the shared latent Mallows components, i.e., the reference rankings σ_k's and the

[1] The total number of ordered pairs on which two rankings differ.

[2] A Mallows pmf with $\phi_k = 1$ is the uniform distribution over all permutations and is unidentifiable.



For each user $m = 1, \dots, M$,

1) Sample a ranking weight vector $\boldsymbol{\theta}_m \in \triangle^K$ from some prior distribution $\Pr(\boldsymbol{\theta})$

2) For each comparison $n = 1, \dots, N$,

   a. Sample unordered item pair $\{i, j\} \sim \mu$

   b. Sample ranking token $z \in \{1, \dots, K\}$ $\sim$ Multinomial$(\boldsymbol{\theta}_m)$

   (c) Sample ranking $\sigma_{m,n} \sim z$-th Mallows component with parameters (σ_z, ϕ_z)

   d. Output ordered pair $w_{m,n} = (i, j)$ if $\sigma_{m,n}(i) < \sigma_{m,n}(j)$; Otherwise output $w_{m,n} = (j, i)$

Figure 2: Generative process of the Mixed Membership Mallows Model and its Graphical representation. The boxes represent replicates (the outer as users, and the inner as comparisons). $\sigma(i)$ is the position of item i in a ranking σ. We adopt the convention that item i is preferred over j if $\sigma(i) < \sigma(j)$.

dispersion parameters ϕ_k's, with polynomial sample and computational complexity guarantees. For the problem of *inferring* $\boldsymbol{\theta}_m$ [24] and *predicting* preferences for new observations, we use standard tools [25]. Establishing guarantees for the inference and prediction problems is not addressed in this work and remains an open question.

Connection to other ranking models M4 subsumes, as special cases, two important ranking models that have been studied in the literature:

Proposition 1. *a) If $\phi_k \to 0$ for all $k = 1, \dots, K$, then each Mallows component reduces to a pmf with all its mass concentrated on the (single) reference permutation σ_k, i.e., then M4 reduces to the model in [1]. b) If the topic prior $\Pr(\theta)$ has non-zero probability only on the vertices of the probability simplex in K dimensions, then M4 reduces to the mixture of Mallows model in [4, 5]*

Therefore, all learning guarantees for M4, to be discussed, also hold for the mixture of Mallows model in [4, 5] and the mixed membership ranking model in [1]. This leads to the *first asymptotic consistency and polynomial sample and computational complexity learning guarantees* for the mixture of Mallows model from pairwise comparisons in a general setting. Also, unlike related work in Table 1 which are all tied to *one* specific prior on the ranking weights $\boldsymbol{\theta}_m$'s, our approach can be applied to all priors on $\boldsymbol{\theta}_m$'s that have a full-rank correlation matrix.

Approach – Reduction to Topic Model via Ranking Matrix Our solution strategy is to formally associate M4 with a topic model whose topic matrix provides an information-equivalent representation of the parameters of M4. To do this, it is convenient to define a $W \times K$ **ranking matrix** $\boldsymbol{\beta}$ with

$$\beta_{(i,j),k} := \sum_{\sigma:\ \sigma(i)<\sigma(j)} p_{\mathrm{M}}(\sigma|\sigma_k, \phi_k) \tag{2}$$

We note that $\boldsymbol{\beta}$ is completely determined by the σ_k's and the ϕ_k's. Statistically, $\beta_{(i,j),k}$ is the probability that a user prefers item i over j in a ranking sampled from the k-th Mallows component. The ranking matrix $\boldsymbol{\beta}$ is analogous to the topic matrix in a topic modeling problem [13]. The set of all possible pairwise comparisons form the "vocabulary", each user's comparisons form a "document", and the shared Mallows components the "topics". The following proposition shows that the underlying σ_k's and ϕ_k's can be recovered from $\boldsymbol{\beta}$.

Proposition 2. *Let the ranking matrix $\boldsymbol{\beta}$ be defined as in Eq.* (2)*. Then, $\forall (i, j)$ and $\forall k$, we have,*

a) If $\sigma_k(i) < \sigma_k(j)$, then $\beta_{(i,j),k} > 0.5 > \beta_{(j,i),k} \geq 0$ and $\beta_{(i,j),k} + \beta_{(j,i),k} = 1$

b) If $\sigma_k(j) = \sigma_k(i) + 1$, then $1/\beta_{(i,j),k} = 1 + \phi_k$.

Prop. 2 a) shows that σ_k's can be recovered from $\boldsymbol{\beta}$ by rounding its entries to the nearest integer. Prop. 2 b) shows that the dispersion parameter ϕ_k can be recovered from. Thus, $\boldsymbol{\beta}$ does indeed provide an information-equivalent representation of M4.

3 Overview of Algorithm, Key Insights, and Theoretical Results

We have just reduced the learning problem of M4 to the estimation the ranking matrix $\boldsymbol{\beta}$ (Eq. (2)), which plays the same role as the topic matrix in a topic modeling problem. Algorithms for learning the topic matrix with polynomial sample and computational complexity guarantees have been recently developed for topic matrices that are *exactly separable* [14, 16]: in β, if for every column there is at least one row whose occurrence probability is nonzero only in that topic, i.e., it is exactly zero in all other topics. Unfortunately, these results can not be directly applied since the entries of $\boldsymbol{\beta}$ of M4 is strictly positive.

As highlighted in the introduction, two nontrivial technical innovations are needed to overcome this difficulty. First, we consider a general "approximate separability" property (Definition 1) and prove that most instances of the ranking matrix in M4, when appropriately sampled, are approximately separable when $Q \gg K$ Second, we generalize the results based on the solid angle (Eq. (5)) for learning $\boldsymbol{\beta}$ from exact to approximate separability. We introduce these key technical advances in this section and summarize the analysis for our algorithm.

3.1 Approximate Separable Ranking Matrix

Our first conceptual innovation is the following notion of an approximately separable ranking matrix:

Definition 1. **(λ-Approximate Separability)** *A $W \times K$ non-negative matrix $\boldsymbol{\beta}$ is λ-approximately separable for some constant $\lambda \in [0, 1)$, if $\forall$ $k = 1, \dots, K$, there exists at least one row (i.e., ordered pair) (i, j) such that $\beta_{(i,j),k} > 0$ and $\beta_{(i,j),l} \leq \lambda \beta_{(i,j),k}$, $\forall l \neq k$.*

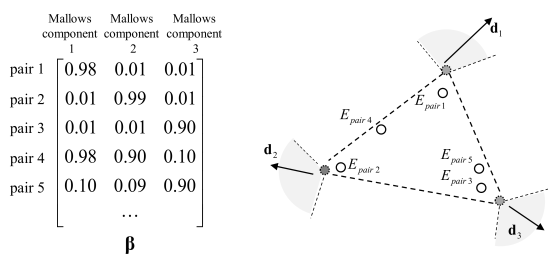


Figure 3: An example of approximate separable $\boldsymbol{\beta}$ with $K = 3$, and the underlying geometry of the row vectors of $\mathbf{E}$. Pair 1, 2, 3 are approximate novel pairs for three Mallows components. The shaded dash circles represent the ideal extreme points with exact separable $\boldsymbol{\beta}$ and the shaded regions depict their solid angles.

Intuitively, λ-approximate separability requires the existence of ordered pairs that have negligible probability in all-but-one of the Mallows components, i.e., the row entries concentrate predominantly in one column. We call such pairs (rows of $\boldsymbol{\beta}$) as λ-approximately novel pairs (rows) for each latent factor. Exact separability studied in [14, 16] corresponds to $\lambda = 0$ and is incompatible with the strict positivity of $\boldsymbol{\beta}$ in M4. Approximate separability is the key.

Figure 3 shows an example where the pairs 1, 2, 3 are, respectively, novel for the first, second, and third Mallows components. Since $\beta_{(i,j),k}$ is a pairwise comparison probability, row (i, j) being approximately novel means that i is preferred over j in only one factor and $i$is mostly likely to be preferred below j in the remaining. To achieve this in the Mallows setting, the position of item i should come before j in one reference ranking (say, $\sigma_1(i) < \sigma_1(j)$) while $\sigma_k(i)$ is after $\sigma_k(j)$ in all the other reference rankings and $L = \sigma_k(i) - \sigma_k(j)$ are large for $k = 2, \dots, K$.

Most ranking matrices of M4 are approximately separable The approximately separability appears to be restrictive. However, it is in fact an inevitable property of M4 when the number of items $Q \gg K$. Concretely, we impose the following prior on the K Mallows components: the K reference rankings σ_k are i.i.d uniformly sampled from the set of all permutations, and the dispersion parameters $\phi_k \leq \phi < 1, \forall k$ are strictly less than 1. We have,

Lemma 1. *Let the reference rankings $\sigma_1, \dots, \sigma_K$ be sampled i.i.d uniformly from the set of all permutations, and the dispersion parameters $\phi_k \leq \phi < 1, k = 1, \dots, K$. Then, the probability that the corresponding ranking matrix $\boldsymbol{\beta}$ being λ-approximately separable for any $\lambda \in (0, 1)$ is at least*

$$1 - K \exp\left(-\frac{Q}{L(\phi, \lambda)^{2K-1}}\right) \tag{3}$$

where $L(\phi, \lambda) = ceil\left(2\frac{\log(\lambda)}{\log(\phi)}\right)$, and $ceil(x)$ is the minimum integer no smaller than x.

By Eq. 3, the probability of $\boldsymbol{\beta}$ being approximately separable converges to 1 exponentially in Q. Noting the logarithm dependency of L on λ, for very small λ, a reasonably small L would satisfy the convergence in Eq. 3. We further note that the result in Eq. (3) is a loose lower bound on the probability of being separable as evidenced by Table 1 in the appendix.

3.2 Robust Novel Pair Detection with Approximate Separable Ranking Matrix

Our second innovation is to generalize the geometric results [16] from exact to approximate separability. We discuss the representation space, the geometric insights, and our algorithm and results in this section.

Comparison Co-occurrence Matrix We construct a $W \times W$ comparison co-occurrence matrix $\mathbf{E}$ as the representation space in which we develop the geometric intuitions. This statistic can be estimated consistently as $M \to \infty$. Specifically, we split each user's comparisons into two independent halves and denote them as the empirical observation matrices $\mathbf{X}$ and $\mathbf{X}'$. For simplicity, we define a $W \times K$ matrix $\mathbf{B}$ as $B_{(i,j),k} = \mu_{i,j}\beta_{(i,j),k}$. We have,

Lemma 2. *Re-scale the rows of $\mathbf{X}$ and $\mathbf{X}'$ to obtain $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{X}}'$ so that they are row-stochastic, then*

$$M\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}^\top \xrightarrow[\textit{almost surely}]{M\to\infty} \bar{\boldsymbol{B}}\bar{\mathbf{R}}\bar{\boldsymbol{B}}^\top =: \mathbf{E}, \tag{4}$$

where $\bar{\boldsymbol{B}} = \operatorname{diag}^{-1}(\boldsymbol{B}\mathbf{a})\boldsymbol{B}\operatorname{diag}(\mathbf{a})$, $\bar{\mathbf{R}} = \operatorname{diag}^{-1}(\mathbf{a})\mathbf{R}\operatorname{diag}^{-1}(\mathbf{a})$. $\mathbf{a} = \mathbb{E}(\boldsymbol{\theta}_m)$ and $\mathbf{R} = \mathbb{E}(\boldsymbol{\theta}_m\boldsymbol{\theta}_m^\top)$ are the expectation and correlation matrix of the topic prior.

We assume that $\mathbf{R}$ has full rank K which is satisfied by many important priors [14]. We note that $\mathbf{E}$ is not explicitly constructed in our projection-based algorithm.

Exact Separability – Novel Pairs are exact extreme points We focus on the rows of $\mathbf{E}$. In the example in Figure 3, if $\boldsymbol{\beta}$ is exactly separable (i.e., the 0.01 terms of first three rows are ideally 0), the corresponding rows of $\mathbf{E}$ are exactly the extreme points of the convex hull formed by all the row of $\mathbf{E}$ (the gray circles in Figure 3). The normalized solid angle proposed in [16] formally captures this property by defining,

$$q_{(i,j)} \triangleq p\{\forall(s,t) : \|\mathbf{E}_{(i,j)} - \mathbf{E}_{(s,t)}\| \geq \zeta, \mathbf{E}_{(i,j)}\mathbf{d} > \mathbf{E}_{(s,t)}\mathbf{d}\}, \mathbf{d} \in \mathbb{R}^W \sim \text{Isotropic} \tag{5}$$

When $\boldsymbol{\beta}$ is exact separable, one can show that $q_{(i,j)}$ is strictly positive iff (i,j) is a novel pair. Statistically, this solid angle is also the probability that a row vector $\mathbf{E}_{(i,j)}$ has the maximum projection value along an isotropically distributed direction $\mathbf{d}$. This definition provides an algorithm to efficiently approximate the solid angles by first projecting rows of $\mathbf{E}$ onto a few i.i.d isotropic $\mathbf{d}$'s and then calculating the frequency of each row being maximum. The novel pairs are therefore the distinct K points with non-zero solid angle [16].

Approximate Separability – Novel Pairs are the most robust extreme points We still focus on the rows of $\mathbf{E}$. Consider now that $\boldsymbol{\beta}$ is λ-approximate separable with small enough $\lambda > 0$. The rows of $\mathbf{E}$ (empty circles in Figure 3) can be viewed as a small perturbation from the ideal case. As a consequence, (a) The rows of approximately novel pairs – $\mathbf{E}_{\text{pair1}}$, $\mathbf{E}_{\text{pair2}}$, and $\mathbf{E}_{\text{pair3}}$ in empty circle – are inside the ideal convex hull and are close to the ideal extreme points. The corresponding solid angles subtended will be close to that of the ideal extreme points which are lower bounded away from 0. (b) The non-novel rows could become extreme points but would be close to the convex hull formed by the approximate novel rows (e.g., $\mathbf{E}_{\text{pair4}}$ in Figure 3). But in this case the associated solid angles will be very close to 0.

To sum up, the solid angle in Eq. (5) can measure the "robustness" of an extreme point. If we sort the non-zero solid angles for all the rows in $\mathbf{E}$, the distinct K rows with largest solid angles must correspond to $c\lambda$-approximate novel pairs for some constant c and a properly defined ζ in Eq. (5).

Overall Algorithm We first detect approximately novel pairs for K distinct Mallows components by sorting the solid angles of all pairs using a few i.i.d isotropic random projections. Once the approximate novel pairs for K distinct Mallows components are identified, $\mathbf{B}$ hence $\boldsymbol{\beta}$ can be estimated using constrained linear regression [14, 16]. We then post-process $\boldsymbol{\beta}$ to get σ_k, ϕ_k's of the shared Mallows components by Prop. 2. These steps are outlined in Algorithm 1. We expand the random projection steps in Algorithm 2. The linear regression steps uses the same strategy as in [14, 16] and are deferred in appendix. In Algorithm 3: step 1 estimates all the pairwise relations $\sigma_{(i,j),k} = \mathbb{I}(\sigma_k(i) < \sigma_k(j))$ in σ_k. Step 2 aggregates them to the positions of each item in σ_k. Step 3 estimates ϕ_k.

Computation and Sample Complexity Bounds We summarize the guarantees of our approach. Recall that M, N, Q, K is the number of user, comparisons per user, the number of items, and the number of Mallows components respectively. P is the number of random projections. First, the computation complexity is polynomial in all parameters,

Algorithm 1 M4 Estimation (Main Steps)

Input: Pairwise comparisons $\widetilde{\mathbf{X}}, \widetilde{\mathbf{X}}'(W \times M)$ (defined in Lemma 2); Number of latent components K; Number of projections P; Tolerance parameters $\zeta, \epsilon > 0$

Output: Reference ranking $\widehat{\sigma}_k$ and dispersion $\widehat{\phi}_k$, $k = 1, \ldots, K$

Novel Pairs $\mathcal{I} \leftarrow$DetectNovelPair($\widetilde{\mathbf{X}}, \widetilde{\mathbf{X}}', K, P, \zeta$)

$\widehat{\mathbf{B}} \leftarrow$EstimateRankingMatrix($\mathcal{I}, \mathbf{X}, \epsilon$)

$\widehat{\sigma}_1, \ldots, \widehat{\sigma}_K, \widehat{\phi}_1, \ldots, \widehat{\phi}_K \leftarrow$PostProcess($\widehat{\mathbf{B}}$)

Algorithm 2 DetectNovelPair (via Random Projections)

Input: $\widetilde{\mathbf{X}}, \widetilde{\mathbf{X}}'$; number of rankings K; number of projections P; tolerance ζ;

Output: $\mathcal{I}$: The set of all novel pairs of K distinct rankings.

$\widehat{\mathbf{E}} \leftarrow M\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}^\top$

$\forall (i,j),\ \mathcal{J}_{(i,j)} \leftarrow \{(s,t) : \|\widehat{E}_{(i,j)} - 2\widehat{E}_{(s,t)}\| \geq \zeta/2\}$,

for $r = 1, \ldots, P$ **do**

  Sample $\mathbf{d}_r \in \mathbb{R}^W$ from an isotropic prior

  $\hat{q}_{(i,j),r} \leftarrow \mathbb{I}\{\forall (s,t) \in \mathcal{J}_{(i,j)},\ \widehat{\mathbf{E}}_{(s,t)}\mathbf{d}_r \leq \widehat{\mathbf{E}}_{(i,j)}\mathbf{d}_r\}$, $\forall (i,j)$

end for

$\hat{q}_{(i,j)} \leftarrow \frac{1}{P}\sum_{r=1}^{P} \hat{q}_{(i,j),r}, \forall (i,j)$

$k \leftarrow 0, l \leftarrow 1$, and $\mathcal{I} \leftarrow \emptyset$

while $k \leq K$ **do**

  $(s,t) \leftarrow$ index of the l^{th} largest value among $\hat{q}_{(i,j)}$'s

  if $(s,t) \in \bigcap_{(i,j)\in\mathcal{I}} \mathcal{J}_{(i,j)}$ **then**

    $\mathcal{I} \leftarrow \mathcal{I} \cup \{(s,t)\}, \quad k \leftarrow k+1$

  end if

  $l \leftarrow l+1$

end while

Algorithm 3 PostProcess

Input: $\widehat{\mathbf{B}}$ as the estimate of $\mathbf{B}$ **Output:** $\widehat{\sigma}_k, \widehat{\phi}_k, k = 1, \ldots, K$

$\widehat{\sigma}_{(i,j),k} \leftarrow$ Round $\left[\frac{\widehat{B}_{(i,j),k}}{\widehat{B}_{(i,j),k} + \widehat{B}_{(j,i),k}}\right], \forall i,j \in \mathcal{U}, \forall k$

$\widehat{\sigma}_k(i) \leftarrow$ Position of ith item when sorting $\{\sum_{j\neq i} \widehat{\sigma}_{(i,j),k}\}_{i=1,\ldots,Q}$ in descending order, $\forall\ k$

$\widehat{\phi}_k \leftarrow \frac{1}{Q-1}\sum_{i=1}^{Q-1} 1/\widehat{\beta}_{(\sigma_k^{-1}(i),\sigma_k^{-1}(i+1)),k} - 1, \forall k$

Theorem 2. *The running time of Algorithm 1 is* $\mathcal{O}(MNP + Q^2P + Q^2K^3)$.

We note that this bounds, although being the first polynomial result, can be further improved but it is not the main focus of this paper. Next, we show consistency and sample complexity bounds in estimating all the reference rankings, Formally,

Theorem 3. *Let the ranking matrix $\boldsymbol{\beta}$ be λ-approximate separable and the second order moments $\mathbf{R}$ of ranking prior to be full rank. If*

$$\lambda \leq \frac{a_{\min}\kappa(1-\phi)q_\wedge}{8K^2 a_0\sqrt{\log(W/q_\wedge)}} \tag{6}$$

and $M, P \to \infty$, then, Algorithm 1 can consistently recover all the reference rankings of the latent Mallows distributions. Moreover, $\forall \delta > 0$, if

$$M \geq \max\left\{\frac{640W^2\log(3W/\delta)}{N\eta^4 d^2 q_\wedge^2}, \frac{320W\log(3W/\delta)}{N\eta^4\lambda_{\min}^2 a_{\min}^2(1-\phi)^2}\right\} \qquad P \geq 32\frac{\log(3W/\delta)}{q_\wedge^2} \tag{7}$$

the proposed algorithm fails with probability at most δ. The other model parameters are defined as follows: $\eta = \min_{1\leq w\leq W}[\mathbf{Ba}]_w$; $a_{\max}$, $a_{\min}$ are the max/min of entries of $\mathbf{a}$; $a_0 = \max_{i,j} a_i/a_j$; $\mathbf{Y} = \bar{\mathbf{R}}\bar{\mathbf{B}}$; $\kappa = \lambda_{\min}/\lambda_{\max}$ is the condition number of $\bar{\mathbf{R}}$; $q_\wedge$ be the minimum normalized solid angle formed by row vectors of $\mathbf{Y}$; $d = 6\kappa/K$; $\phi_k \leq \phi < 1$.

All proofs are deferred in supplementary. In Theorem 3, the Eq. (6) provides an explicit sufficient condition on the required λ. In this bound, λ is inverse polynomial in K. Therefore, in the Eq. 3, the margin L required to achieve a high separable probability would scale as $L(\phi, \lambda) = c_1 + c_2\log(K)$ which is small.

4 Experimental Validation

We conduct semi-synthetic simulation to validate our approach when M4 is true. We also conduct real-world experiments to demonstrate that M4 can effectively capture real-world preference. Star rating datasets are used for its large public availability. We used the suggested settings by [16]. Specifically, $P = 150 \times K, \zeta = 0.05$ in Alg. 2. More detailed settings are in supplementary.

Semi-synthetic Simulation We validate the performance of proposed on semi-synthetic dataset. The ground-truth reference rankings are obtained from a real world movie rating dataset, Movielens, [3] using the same approach as in [1] over the $Q = 100$ most rated items and $K = 10$. We set the same dispersion parameters $\phi_k = \phi \in \{0, 0.1, 0.2, 0.5\}$. The ground-truth $\boldsymbol{\beta}$ is $\lambda = 0, 0.01, 0.05, 0.20$ approximate separable. We use a symmetric Dirichlet prior with concentration $\alpha_0 = 0.1$ on $\boldsymbol{\theta}_m$'s. $N = 300$. $\mu_{i,j} = 1/\binom{Q}{2}, \forall i, j$ is uniform. We evaluate the performance using **Kendall's tau distance** between the estimated $\widehat{\sigma}_k$ and the ground-truth after a bipartite matching. The error is normalized by $W = Q(Q-1)$ and averaged across the K rankings.

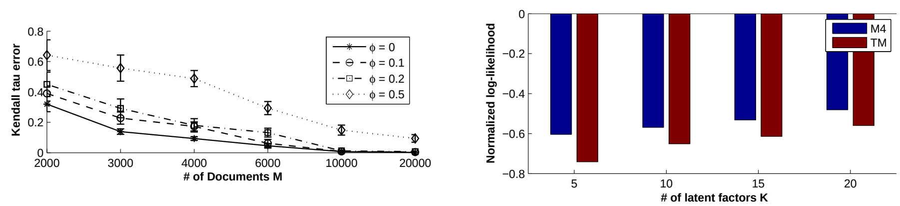


Figure 4: Left: the normalized Kendall's tau distance of the estimated reference rankings, as functions of M, from the semi-synthetic dataset with $Q = 100, N = 300, K = 10$ and different ϕ. Right: the normalized predictive log-likelihood for various K on the truncated Movielens dataset.

Fig. 4 (left) depicts how the estimation error varies with the number of users M for dispersion ϕ. We can see that the reconstruction error converges to zero at different rates in M when λ is small ($\phi = 0, 0.1, 0.2$), and converges to a small but non-zero number when λ is mild ($\phi = 0.5$).

Comparison Prediction on Movielens We predict pairwise comparisons in the real-world Movielens dataset. The star rating based dataset is selected due to public availability and widespread use, but we convert it to pairwise comparisons as suggested in the ranking literature [1, 4, 7]. We focus on the $Q = 200$ most frequently rated movies in the Movielens, split the first $M = 4000$ users for training, and use the remaining users for testing [4]. We convert the training and test ratings into comparisons independently: for all pairs of movies i, j user m rating, $w_{m,n} = (i, j)$ is added if the star ratings for i is higher than j, and all ties are ignored. The prior is set to be Dirichlet. We evaluate the performance by the **held-out log-likelihood**, i.e., $\Pr(\mathbf{w}_{test}|\widehat{\boldsymbol{\beta}})$ using standard tools in [25]. We compared our new model (M4) against the model in [1] (TM) with closest settings to our model. As shown in Figure 4 (right), M4 improves the prediction accuracy of TM for different choice of K.

Table 2: Testing RMSE on the Movielens dataset

K	PMF	BPMF	BPMF-int	TM	M4
10	1.0491	0.8254	0.8723	0.8840	0.8509
15	0.9127	0.8236	0.8734	0.8780	0.8296
20	0.9250	0.8213	0.8678	0.8721	0.8241

Rating prediction via ranking model on Movielens We consider a standard task in recommendation system, star rating prediction, to illustrate our model can better capture real-world behavior [22]. The same training/testing rating split in [8] is used, and we focus on the $Q = 100$ most rated movies following [1]. We train our M4 model on the training comparisons and use that to predict the testing star ratings which induces the most likely testing comparisons. The same K is used for different algorithms since it is the number of latent factors in all the models considered. Detailed settings are in supplementary. We evaluate the performance using the standard root-mean-square-error (**RMSE**) metric [22]. As shown in Table 2, M4 improves upon TM and matches the rating-based benchmarks BPMF [8], PMF[26] although they are coming from a different feature space. We note that the BPMF typically provides robust and benchmark results on real-world problems. This demonstrates that our approach can accommodate noisy real-world user behavior.

[3] `http://grouplens.org/datasets/movielens/`

References

[1] W. Ding, P. Ishwar, and V. Saligrama. A Topic Modeling approach to Ranking. In *Proc. of the 18th International Conference on Artificial Intelligence and Statistics*, San Diego, CA, USA, May. 2015.

[2] I. Gormley and T. Murphy. A grade of membership model for rank data. *Bayesian Analysis*, 4(2):265–295, 2009.

[3] Y. Kim, W. Kim, and K. Shim. Latent ranking analysis using pairwise comparisons. In *2014 IEEE International Conference on Data Mining*.

[4] T. Lu and C. Boutilier. Effective Sampling and Learning for Mallows Models with Pairwise-Preference Data. *Journal of Machine Learning Research*, 15:3783–3829, 2014.

[5] P. Awasthi, A. Blum, O. Sheffet, and A .Vijayaraghavan. Learning mixtures of ranking models. In *Advances in Neural Information Processing Systems*. Montreal, Canada, Dec. 2014.

[6] S. Oh and D. Shah. Learning mixed multinomial logit model from ordinal data. In *Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 2014.

[7] M. Volkovs and R. Zemel. New learning methods for supervised and unsupervised preference aggregation. *Journal of Machine Learning Research*, 15:1135–1176, 2014.

[8] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proc. of the 25th International Conference on Machine Learning*, Helsinki, Finland, Jun. 2008.

[9] C. L Mallows. Non-null ranking models. i. *Biometrika*, pages 114–130, 1957.

[10] G. Lebanon and J. D. Lafferty. Cranking: Combining rankings using conditional probability models on permutations. In *Proc. of the 19th International Conference on Machine Learning*, Sydney, Australia, Jul. 2002.

[11] L. Busse, P. Orbanz, and J. Buhmann. Cluster analysis of heterogeneous rank data. In *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, USA, Jul. .

[12] M. Meila and H. Chen. Dirichlet process mixtures of generalized mallows models. In *Proc. of The Conference on Uncertainty in Artificial Intelligence*, Catalina Island, CA, USA,, Jul. 2010.

[13] D. Blei. Probabilistic topic models. *Commun. of the ACM*, 55(4):77–84, 2012.

[14] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M.l Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proc. of the 30th International Conference on Machine Learning*, Atlanta, GA, USA, Jun. 2013.

[15] W. Ding, M. H. Rohban, P. Ishwar, and V. Saligrama. Topic discovery through data dependent and random projections. In *Proc. of the 30th International Conference on Machine Learning*, Atlanta, GA, USA, Jun. 2013.

[16] W. Ding, M. H. Rohban, P. Ishwar, and V. Saligrama. Efficient Distributed Topic Modeling with Provable Guarantees. In *Proc. ot the 17th International Conference on Artificial Intelligence and Statistics*, Reykjavik, Iceland, Apr. 2014.

[17] T. Bansal, C. Bhattacharyya, and R. Kannan. A provable svd-based algorithm for learning topics in dominant admixture corpus. In *NIPS*, pages 1997–2005, 2014.

[18] P. Awasthi and A. Risteski. On some provably correct cases of variational inference for topic models. *ArXiv e-prints*, 2015.

[19] A. Kumar, V. Sindhwani, and P. Kambadur. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *the 30th Int. Conf. on Machine Learning*, Atlanta, GA, Jun. 2013.

[20] V. Farias, S. Jagabathula, and D. Shah. A data-driven approach to modeling choice. In *Advances in Neural Information Processing Systems*. Vancouver, Canada, Dec. 2009.

[21] H. Azari Soufiani, H. Diao, Z. Lai, and D. C Parkes. Generalized random utility models with multiple types. In *Advances in Neural Information Processing Systems*, pages 73–81. Lake Tahoe, NV, USA, Dec. 2013.

[22] A. Toscher, M. Jahrer, and R. M. Bell. The bigchaos solution to the netflix grand prize, 2009.

[23] F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[24] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, Mar. 2003.

[25] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proc. of the 26th International Conference on Machine Learning*, Montreal, Canada, Jun. 2009.

[26] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2008.