Inference Algorithms for the Multiplicative Mixture Mallows Model

Ranjitha Prasad and Vincent Y. F. Tan, Senior Member, IEEE

Abstract—A popular approach to obtain a consensus ranking from ranking data is based on the probabilistic, distance-based Mallows model comprising of a modal permutation and dispersion parameters. Often, the population consists of several subpopulations. As a result, finite mixture models are used to distinguish latent sub-groups of individuals in a heterogeneous population. Given a finite number of subpopulations each based on the Mallows model, a popular inference approach is the computationally intensive expectation maximization algorithm for additive models. We address the drawbacks of this model using a novel multiplicative mixture Mallows model (M4). Given complete and partial ranking observations from a heterogeneous population, we derive inference algorithms for the joint estimation of per-sample weights, parameters, and the ground truth permutations of the component distributions. For the inference of parameters, we adopt the majorization-minimization framework which is reminiscent of stochastic gradient descent. We derive Cramér-Rao lower bounds as a fundamental limit on the mean square error (MSE) parameter estimation performance of the inference algorithms. We numerically validate the MSE and permutation estimation performance of the proposed algorithms on synthetic datasets. We also demonstrate the goodness-of-fit using the Bayesian information criterion and the integrated complete likelihood on the real-world APA and Sushi datasets.

Index Terms—Ranking, Mallows model, Mixture models, Majorization-minimization, Stochastic gradient descent, Goodness-of-fit.

1 INTRODUCTION

Ranking is an essential ingredient of a gamut of applications such as electoral preference learning, personalized advertisement targeting, recommender systems, etc. Typically ranking data is obtained from surveys and market studies where the participants provide complete or partial list of items in the order of preference, which we refer to as a ranking order. The design of a ranking system involves choosing an appropriate ranking model based on the nature of the ranking orders, subsequently learning the model parameters and obtaining the ground truth permutation (the consensus ranking) that best agrees with the given sample permutations. The task of ranking is related to several problems in computer science such as query processing [1], [2], noisy sorting [3] and multi-label classification [4].

Typically, given $n$ items, the survey participants independently generate ranking orders of length $t \leq n$. There is substantial amount of work on models, algorithms and guarantees that consider observations in the form of pairwise preferences ($t = 2$) [5], [6], [7]. Several ranking models such as the Plackett-Luce [8], Mallows [9], non-parametric kernel based models [10], etc consider ranking orders of length $2 \leq t \leq n$. In this work, we are interested in the Mallows model which is a probabilistic distance-based modeling approach to analyze rankings [9], [11]. Mallows models are employed when observations are permutations ($t = n$) and when the observations consist of the $t$ most highly rated items ($t < n$). Here, each observation is regarded as a noisy version of the ground truth permutation whose probability of occurrence is inversely related to the distance between itself and the ground truth permutation [9]. Algorithms to infer the model parameters and the underlying consensus ranking using the Kendall-Tau distance metric [9], [12] and its variants [13] have been proposed [9], [12], [13].

Real world survey data often consists of samples from heterogeneous subpopulations. It is common to assume that the global population consists of a number of subpopulations of rankers who share a common preference behaviour and thus, each cluster is characterized by a unique ground truth permutation and dispersion parameters. Recently, theoretical guarantees have been provided for learning additive mixtures of ranking models [14], [6], [7]. In the context of the permutation-based Mallows model, additive mixture models are commonly employed, where the expectation maximization (EM) algorithm is used for inference [15], [16]. A general EM-based inference framework for the Mallows model with $t \leq n$ was proposed in [17]. The critical drawback of these inference techniques is that the expectation requires weighted summations over all $n!$ possible permutations. For large $n$, the complexity of EM algorithm is prohibitive and the convergence is slow [16]. Furthermore, in [18], the authors note that in high-dimensions, the posterior distribution behaves similarly to the individual components of the additive mixture density, leading to a vacuous modelling of heterogeneous population.

In order to address the computational shortcomings of additive mixture models, we propose a novel, low-complexity model known as the multiplicative mixture Mallows model (M4), where the overall mixture distribution is an exponentially weighted product of the component distributions. Such models are similar to the product of experts model [18]. Moreover, in the theory of hypothesis testing, such a product mixture distribution is a tilting from one component distribution to another [19, Chapter 11]. Based on the concept of tilting, we propose a sample-wise cluster assignment where every sample is a weighted product of...
two out of $M$ component distributions, resulting in a 2-sparse per-sample weights.

**Main Contributions**

We propose the $M^4$ for ranking orders obtained from a heterogeneous population. We consider two scenarios, one in which the observed rankings are permutations of $n$ items, and the other, where the observed rankings consist of the top-$t$ items. For both scenarios, our contributions are as follows:

- We derive an iterative majorization-minimization (M-M) approach to learn the per-sample weights, and the dispersion parameters of the model. Specifically:
  - We employ the concave convex procedure (CCCP) by deriving various convex auxiliary functions for iterative optimization.
  - We employ the low-complexity SGD algorithm in each step of the iterative algorithm, rendering the overall algorithm to be implementation friendly.
- We derive Cramér-Rao lower bounds (CRLBs) as the fundamental lower bounds on the mean square error (MSE) performance of unbiased estimators for the proposed $M^4$.
- We propose a greedy approach for joint estimation of the dispersion parameters, per-sample weights, and ground truth permutation estimation algorithm.
- We validate the mean square error (MSE) performance of the proposed algorithms on synthetic datasets and compare the performance with the corresponding CRLB.
- We use the Bayesian information criterion (BIC) and integrated complete likelihood (ICL) [20] to demonstrate the goodness-of-fit of $M^4$ vis-à-vis additive mixture models and single component distributions. We validate our methods on real-world datasets such as the APA and the Sushi datasets. We also compare and contrast the computational ease of $M^4$ vis-à-vis additive models.

The rest of this paper is organized as follows. In Sec. 2, we describe the Mallows model and the corresponding distance metric for full and partial ranking observations. In Sec. 3 we propose $M^4$ in the context of permutation-based rankings and obtain the inference algorithms in 4. We propose $M^4$ in the context of partial ranking observations and derive corresponding inference algorithms in Sec. 5. The efficacy of $M^4$ is demonstrated on real and synthetic datasets in Sec. 6. We provide concluding remarks and directions for future research in Sec. 7. Technical proofs deferred to the appendices.

## 2 The Mallows Model and Preliminaries

In this section, we describe the ranking data and system, hence laying the foundation for inference in the $M^4$. Let $\pi$ denote the ranking order of a survey participant when queried over $n$ items in the decreasing order of preference, i.e., $\pi$ is a permutation over the set $[n] = \{1, \ldots, n\}$, where $\pi(l)$ denotes the rank of $l$ in $\pi$. Let $\Omega$ be the set of all $n!$ possible rankings and $d(\cdot, \cdot)$ be a distance function on $\Omega \times \Omega$, such that $d(\pi_i, \pi_j) \geq 0$ for every $\pi_i, \pi_j \in \Omega$, and $d(\pi_i, \pi_j) = 0$ if and only if $\pi_i = \pi_j$. Furthermore, $d(\pi_i, \pi_j)$ is right invariant, i.e., $d(\pi_i, \pi_j) = d(\pi_i \pi_k, \pi_j \pi_k)$, where $\pi \pi_k$ represents the composition of two permutations.

### 2.1 Mallows Model

In the vanilla Mallows model [11], the rankings are generated from a probability density function given by

$$p_0(\pi) = \frac{\exp(-\theta d(\pi, \pi_0))}{\psi(\theta)}, \quad \pi \in \Omega, \quad \theta > 0. \quad (1)$$

In the above, $\pi_0$ represents the ground truth permutation and $\theta$ is an inverse scale dispersion parameter such that when $\theta \rightarrow \infty$, $p_0(\pi)$ concentrated at the ground truth permutation $\pi_0$ and when $\theta \rightarrow 0$, $p_0(\pi)$ is the uniform distribution over all permutations [9]. For the Kendall-Tau distance, every permutation $\pi$ is uniquely determined from $n - 1$ integers (sufficient statistics), given by

$$V_j(\pi) = \sum_{i > j} 1 \{ i \prec_j j \} \quad \text{and} \quad d(\pi, \pi_0) = \sum_{j=1}^{n-1} V_j(\pi \pi_0^{-1}), \quad (2)$$

where $i \prec_j j$ means that $i$ is ranked before $j$ in $\pi$, $1\{\cdot\}$ is the indicator function, and $V_j(\pi \pi_0^{-1}) \in \{0, \ldots, n-j\}$. In [9], it has been shown that that the model in (1) factors into a product of independent univariate exponential models, one for each $V_j(\pi)$ where $j = 1, \ldots, n-1$.

A parametrized generalization of the Mallows model is

$$\text{Pr}(V_j(\pi \pi_0^{-1}) = v_j) = \frac{\exp(-\theta_j v_j)}{\psi_j(\theta_j)}, \quad v_j = 0, 1, \ldots, n-j, \quad (3)$$

where $\psi_j(\theta_j) = (1 - \exp(-(n-j+1)\theta_j))^{-1}(1 - \exp(-\theta_j))^{-1}$ is the normalization constant. Note that dispersion parameters $\theta_j$ allow us to emphasize ranks individually, hence modelling scenarios where there is consensus for the highly ranked items, but indifference to the ranking of other items. The joint Mallows distribution [9] is given by

$$\text{Pr}(V_1(\pi \pi_0^{-1}) = v_1, \ldots, V_{n-1}(\pi \pi_0^{-1}) = v_{n-1}) = \prod_{j=1}^{n-1} \frac{\exp(-\theta_j v_j)}{\psi_j(\theta_j)}. \quad (4)$$

This is a product of independent univariate exponential models, one for each $V_j(\pi \pi_0^{-1})$.

### 2.2 IGM for Top-t Ranking Framework

In order to incorporate the top-$t$ ranking framework, an infinite generalized Mallows (IGM) model was proposed [13]. In this model, the top-$t$ ordering $\sigma$ is assumed to be a prefix of an infinite length permutation $\pi$. The infinite permutation can be represented uniquely by the reciprocal Kendall-Tau distance, which can be better understood using the permutation matrix representation of $\pi$. The permutation matrix, $P$ corresponding to $\pi$ has $P(i, j) = 1$ if and only if $\pi(i) = j$, and $P(i, j) = 0$ otherwise. Hence, the permutation matrix of
an infinite permutation is an infinite matrix. The reciprocal Kendall-Tau distance is represented by $S_1(\pi), S_2(\pi), \ldots$, where $S_1(\pi)$ is the number of zeros preceding 1 in the first column. We delete this row of the permutation matrix, and compute $S_2(\pi)$ as the number of zeros preceding 1 in the second column, and so on. The procedure above can be succinctly stated as follows:

$$S_j(\pi) = (\pi^{-1}(j) - 1) - \sum_{j' < j} 1\{\pi^{-1}(j') < \pi^{-1}(j)\}. \quad (5)$$

In the above $\pi^{-1}(j) - 1$ represents the number of zeros that precedes 1 in an column. In addition, $\sum_{j' < j} 1\{\pi^{-1}(j') < \pi^{-1}(j)\}$ accounts for the row-deletion operation as discussed earlier. As an example, consider an infinite permutation $\pi = (3, 4, 2, 6, 1, \ldots)$ and its top-4 counterpart $\sigma = (3, 4, 2, 6)$. Here, $S_1(\pi) = 2, S_2(\pi) = 2, S_3(\pi) = 1$ and so on.

The composition $\pi \pi'$ of two infinite permutations $\pi$ and $\pi'$ is given by the product of the corresponding permutation matrices $\Pi \Pi'$ intuitionally, $S_j(\pi | \pi_0)$ depends on the rank of $\pi^{-1}(j)$ in the ground truth permutation $\pi_0$. Accordingly, the reciprocal distance w.r.t. $\pi_0$ is given by

$$S_j(\pi | \pi_0) = (\pi_0^{-1}(j) - 1) - \sum_{j' < j} 1\{\pi_0^{-1}(j') < \pi_0^{-1}(j)\}. \quad (6)$$

Similar to the computation of $S_j(\pi), S_j(\pi | \pi_0)$ is the number of zeros preceding 1 in the first column of $\Pi_j \Pi$ and so on.

The matrix $\Sigma$ of the top-t ordering $\sigma$ is a truncation of its infinite permutation matrix counterpart $\Pi$, retaining $t$ columns of $\Pi$. Further, the matrix used to obtain the reciprocal distance is $\Pi_j^T \Sigma$. Continuing the example, say $\pi_0 = (2, 1, 5, 4, 3, 6, 7, \ldots)$. Hence, $S_1(\pi | \pi_0) = 4, S_2(\pi | \pi_0) = 3, S_3(\pi | \pi_0) = 0$, etc.

The generalization of the top-t Mallows model is given by

$$Pr(S_j(\sigma | \sigma_0) = s_j) = \frac{\exp\{-\gamma_j s_j\}}{\Lambda_j(\gamma_j)}, \quad s_j = 0, 1, 2 \ldots \quad (7)$$

where $\Lambda_j(\gamma_j) = (1 - \exp(-\gamma_j))^{-1}$ is the normalization constant. Similar to the joint distribution of the permutation-based model in (4), the joint distribution of the top-$t$ Mallows model factors into a product of independent univariate exponential models, one for each $S_j(\sigma | \sigma_0)$. That is, if $K \subset \mathbb{N}$ is a finite set, say $K = \{k_1, k_2, \ldots, k_{t-1}\}$, the joint distribution is

$$Pr(S_{k_1}(\sigma | \sigma_0) = s_{k_1}, \ldots, S_{k_{t-1}}(\sigma | \sigma_0) = s_{k_{t-1}}) = \prod_{i=1}^{t-1} \frac{\exp\{-\gamma_{k_i} s_{k_i}\}}{\Lambda_{k_i}(\gamma_{k_i})}. \quad (8)$$

In the sequel, we propose the $M4$ model and derive the inference framework for the permutation and top-$t$ $M4$ models.

### 3 Permutation-based Multiplicative Mixture Mallows Model (M4)

We present the $M4$ model for learning the ground-truth permutations and the associated dispersion parameters of the Mallows model. Consider a heterogeneous population consisting of $M$ clusters, where the $m$-th cluster is characterized by a ground-truth permutation $\pi_m$ and parameters $\theta_m \in \mathbb{R}^{n-1}_{+}$. The Kendall-Tau distance $[12]$ between a given sample $\pi_k$ and the $m$-th ground truth permutation $\pi_m$ is given by

$$V_{mj}(\pi_k, \pi_m) = \sum_{i > j} 1\{i \prec \pi_k, j \prec \pi_m\}, \quad \text{where} \quad \pi_k = \pi_k \pi_m^{-1}. \quad (9)$$

Given $K$ independent and identically distributed (i.i.d.) sample permutations $\pi = [\pi_1, \ldots, \pi_K]$, $M$ component distributions and $v_m(k) = \{v_m(k)\}_{n-1}^{-1}$, a weighted additive mixture distribution is given by [21]

$$p(v_m(k) | \theta) = \sum_{m=1}^{M} w_m p_m(v_m(k) | \theta_m), \quad (10)$$

where $w_m$ represents the mixing weight of the $m$-th component distribution given by $p_m(v_m(k) | \theta_m)$. The Kendall-Tau distance for the $k$-th sample, $v_m(k)$ is an $n-1$ length vector whose $j$-th entry is given by $v_m(\pi_0(k) = V_{mj}(\pi_k \pi_0)^{-1}$ (cf. (9)). Rewriting (10) in terms of the per-sample latent boolean weight vector $z_k = (z_k(1), \ldots, z_k(M))'$ such that $\sum_{m=1}^{M} z_k(m) = 1$, and $p(z_k(1) = 0, \ldots, z_k(m) = 1, \ldots, z_k(M) = 0) = w_m$, we obtain [22]

$$p(v_m(k) | \theta) = \sum_{m=1}^{M} w_m \prod_{m=1}^{M} p_m(v_m(k) | \theta_m)^{z_k(m)}, \quad (11)$$

where $\theta \in \mathbb{R}^{M \times (n-1)}$ consists of $\theta_1, \ldots, \theta_M$ as its columns. Setting $z_k(m) = 1$ implies that the $k$-th sample is a member of the $m$-th subpopulation [22], [23]. Hence the distribution of the $k$-th sample $v_m(k)$ conditioned on $\theta$ and $z_k$ is

$$p(v_m(k) | \theta, z_k) = \frac{1}{c(z_k, \theta)} \prod_{m=1}^{M} p_m(v_m(k) | \theta_m)^{z_k(m)}, \quad (12)$$

where $c(z_k, \theta)$ is the partition function. The model in (12) resembles the product of experts model proposed in [18]. In fact, the ranking model for the homogeneous population in [12] is a special case of the $M4$ model when we set $M = 1$ and $z_k(1) = 1$ for all $k$.

We relax the constraint on the boolean nature of vector $z_k$, and assume that its $m$-th entry $z_k(m) \in [0, 1]$. This model is partly inspired by the tilting of probability distributions, which is an ubiquitous concept in hypothesis testing [19, Ch. 11]. For example, in a binary hypothesis test with i.i.d. observations, the minimum type-I error given an exponential constraint on the type-I error also decays exponentially with the exponent being the relative entropy between a tilting of the two distributions. In Bayesian multiple hypothesis testing, this exponent is the minimum Chernoff distance among all pairs of probability distributions corresponding to the multiple hypotheses [24]. Hence, instead of assigning a given sample to one of the $M$ subpopulations, we assign it to one of the $\binom{M}{2}$ size-2 subpopulations, where each of the $M$ hypotheses are characterized by distinct Mallows model. Accordingly, the weight vector $z_k$ is 2-sparse (i.e., $\|z_k\|_{0} = 2$).

We assume that the distribution of the permutations in each subpopulation is a Mallows distribution defined by

$$p_m(v_m(k) | \theta_m) = \prod_{j=1}^{n-1} \frac{\exp\{-\theta_m v_m(k)\}}{\psi_j(\theta_j)}. \quad (13)$$
Hence, the \( \mathbb{M} \) can be written as

\[
p(v(k)|\theta, z_k) = \frac{\exp \left\{ -\sum_{m=1}^{M} z_k(m) \sum_{j=1}^{n-1} \theta_{mj} v_{mj}(k) \right\}}{c(z_k, \theta)},
\]

where \( v(k) = [v_1(k), \ldots, v_M(k)] \) and \( z_k = [z_k(m), \ldots, z_k(M)] \) and \( z = [z_1, \ldots, z_K] \). To obtain the partition function \( c(z_k, \theta) \), we use the fact that the Mallows model separates into \( n - 1 \) univariate distributions.

Let \( p_j(v_j(k)|\theta_j, z_k) = \prod_{m=1}^{M} \frac{1}{z_k(m)} \exp(-\theta_{mj} v_{mj}(k)) \) be the \( j \)-th univariate distribution, where \( v_j(k) = [v_1(k), \ldots, v_M(k)]^T \). \( \mathbb{M} \) factors in terms of \( p_j(v_j(k)|\theta_j, z_k) \) as

\[
p(v(k)|\theta, z_k) = \frac{1}{c(z_k, \theta)} \prod_{j=1}^{n-1} p_j(v_j(k)|\theta_j, z_k),
\]

where \( \theta_j = [\theta_{j1}, \ldots, \theta_{jM}]^T \) and \( \theta = [\theta_1, \ldots, \theta_{n-1}] \in \mathbb{R}^{M \times (n-1)} \). The partition function can be written as \( c(z_k, \theta) = \prod_{j=1}^{n-1} c_j(z_k, \theta_j) \), where each factor is \( c_j(z_k, \theta_j) = \sum_{v_{mj}(k)} \prod_{m=1}^{M} p_j(v_m(k) | \theta_j) z_k(m) \). This yields \( \mathbb{M} \)

\[
c_j(z_k, \theta_j) = \psi_j \left( \sum_{m=1}^{M} z_k(m) \theta_{mj} \right) \left( \prod_{m=1}^{M} \psi(\theta_{mj}) z_k(m) \right)^{-1}.
\]

Thus, the overall mixture distribution of the \( k \)-th sample is

\[
p(v(k)|\theta, z_k) = \frac{\exp \left\{ -\sum_{m=1}^{M} z_k(m) \sum_{j=1}^{n-1} \theta_{mj} v_{mj}(k) \right\}}{\prod_{j=1}^{n-1} \psi_j \left( \sum_{m=1}^{M} z_k(m) \theta_{mj} \right)}.
\]

### 4 MajORIZATION-MINIMIZATION (M-M) ALGORITHM

The M-M framework [25] solves difficult optimization problems by iteratively minimizing a majorizing function until a local optimum is obtained. We present an M-M algorithm for parameter estimation for the \( \mathbb{M} \) model. Given \( K \) i.i.d. sample permutations, the goal is to estimate \( (\theta, z) \), assuming that the ground-truth permutations \( \pi_{m0} \) are known. We address the estimation of \( \pi_{m0} \) in Sec. 4.4. The log-likelihood is

\[
\ell(\theta, z) = \sum_{k=1}^{K} \log p(v(k)|\theta, z_k)
\]

\[
= -\sum_{k=1}^{K} \sum_{j=1}^{n-1} \sum_{m=1}^{M} z_k(m) \theta_{mj} v_{mj}(k)
\]

\[
-\log \left( 1 - \exp \left[ -n^j + 1 \right] \sum_{m=1}^{M} z_k(m) \theta_{mj} \right)
\]

\[
+ \log \left( 1 - \exp \left[ -\sum_{m=1}^{M} z_k(m) \theta_{mj} \right] \right).
\]

Rewriting (18) in terms of \( \alpha_{kj} \in \mathbb{R}_+^M \) whose entries are given by \( \alpha_{kj}(m) = z_k(m) \theta_{mj} \), the optimization problem given above can be expressed as

\[
\ell(\theta, z) = -\sum_{k=1}^{K} \sum_{j=1}^{n-1} \ell(\alpha_{kj}),
\]

where

\[
\ell(\alpha_{kj}) = \alpha_{kj}^T v_j(k) - \log \left( 1 - \exp \left[ -\sum_{m=1}^{M} \alpha_{kj}(m) \right] \right)
\]

\[
+ \log \left( 1 - \exp \left[ -(n-j+1) \sum_{m=1}^{M} \alpha_{kj}(m) \right] \right).
\]

**Proposition 1.** The function \( \ell(\alpha_{kj}) \) is a difference of convex (d.c.) functions.

**Proof:** See Appendix A.

#### 4.1 CCCP for Parameter Estimation in the \( \mathbb{M} \)

We now employ the CCCP which is a popular technique to obtain solutions to the difference of convex functions given in (19) [26]. CCCP converts such a function to a sequence of convex functions and iteratively solves the original optimization problem by obtaining the optima to the intermediate convex functions. Specifically, in the \( p \)-th iteration, we construct a convex majorizing function \( Q(\alpha_{kj}; \alpha_{kj}^{(p)}) \), such that the \( (p+1) \)-st iterate of \( \alpha_{kj} \) is given by

\[
\alpha_{kj}^{(p+1)} = \arg \max_{\alpha_{kj} \in \mathbb{R}_+^M} Q(\alpha_{kj}; \alpha_{kj}^{(p)}),
\]

and such that \( Q(\alpha_{kj}; \alpha_{kj}^{(p)}) \geq f(\alpha_{kj}) - \alpha_{kj}^{(p)} \nabla_{\alpha_{kj}} g(\alpha_{kj}) \).

\( Q(\alpha_{kj}; \alpha_{kj}^{(p)}) \geq \ell(\alpha_{kj}) \) and \( Q(\alpha_{kj}; \alpha_{kj}^{(p)}) = \ell(\alpha_{kj}) \). We compute the majorizing function \( Q(\alpha_{kj}; \alpha_{kj}^{(p)}) \) as

\[
Q(\alpha_{kj}; \alpha_{kj}^{(p)}) = \alpha_{kj}^T v_m(k) - c_{kj} \sum_{m=1}^{M} \alpha_{kj}(m)
\]

\[
- \log \left( 1 - \exp \left[ -\sum_{m=1}^{M} \alpha_{kj}(m) \right] \right),
\]

where \( c_{kj} \) is given by

\[
c_{kj} = \frac{(n-j+1) \exp \left[ -(n-j+1) \sum_{m=1}^{M} \alpha_{kj}(m) \right]}{1 - \exp \left[ -(n-j+1) \sum_{m=1}^{M} \alpha_{kj}(m) \right]}
\]

Hence, the CCCP procedure leads to an affine approximation \( g(\alpha_{kj}) \) about \( \alpha_{kj}^{(p)} \), and linearly combines it with the convex function \( f(\alpha_{kj}) \) resulting in a convex majorizing function \( \alpha_{kj} \mapsto Q(\alpha_{kj}; \alpha_{kj}^{(p)}) \). To estimate \( z_k \) and \( \theta \), from \( \alpha_{kj} \) for \( k \in [K] \) and \( j \in [n-1] \), we use the biconvexity property [27] of \( Q(\alpha_{kj}; \alpha_{kj}^{(p)}) \).

**Proposition 2.** The function \( (z_k, \theta_j) \mapsto Q(\alpha_{kj}; \alpha_{kj}^{(p)}) \) is biconvex in \( z_k \) and \( \theta_j \). The feasible set \( \Theta \times Z \) is also biconvex.

**Proof:** See Appendix B.

Several methods have been proposed to solve biconvex optimization problems under different conditions [28]. However, for the case of a non-linear biconvex optimization problem as in (21), the alternate convex search (ACS) algorithm [27] is usually employed.
4.2 Alternate Convex Search

In this subsection, we discuss the ACS algorithm which is an iterative framework to solve a biconvex optimization problem. Here, the set of unknown variables are divided into disjoint sub-blocks. In every iteration of the ACS algorithm, only one sub-block is optimized while the other sub-block is held fixed [27]. Since fixing one of the sub-block results in a convex sub-problem, efficient algorithms can be used to solve biconvex programs.

In the context of $\theta$, the convex sub-problem for optimizing $z_k$ with $\theta_j = \theta_j^{(p)}$ is given by

$$z_k^{(p+1)} = \arg \min_{z_k \in \mathbb{R}^d} Q^\theta(z_k),$$

where $Q^\theta(z_k)$ is obtained by fixing $\theta_j = \theta_j^{(p)}$ in (21), i.e.,

$$Q^\theta(z_k) = \sum_{j=1}^{n-1} z_k^T \text{diag}(\theta_j^{(p)}) v_j(k) - \log(1 - \exp(-z_k^T \theta_j^{(p)})) - c_{kj}^{(p)} z_k^T \theta_j^{(p)},$$

The optimization problem given above is a finite-sum problem, i.e., $Q^\theta(z_k) = \sum_{j=1}^{n-1} Q^\theta_j(z_k)$ where

$$Q^\theta_j(z_k) = z_k^T \text{diag}(\theta_j^{(p)}) v_j(k) - \log(1 - \exp(-z_k^T \theta_j^{(p)})) - c_{kj}^{(p)} z_k^T \theta_j^{(p)}.$$  

Note that the optimization problem in (24) cannot be solved in closed form. However, the finite-sum property allows us to employ the stochastic gradient descent (SGD) technique to obtain $z_k^{(p+1)}$ [29]. Using the SGD approach, the $j$th (where $j \in [n-1]$) update for $z_k$ is

$$z_k^{(p+1)} \triangleq z_k^{(p)} - \eta_{z_k} \nabla_{z_k} Q^\theta_j(z_k^{(p)}),$$

where $\eta_{z_k}$ is a learning rate parameter. Here, $\nabla_{z_k} Q^\theta_j(z_k^{(p)})$ is given by

$$\nabla_{z_k} Q^\theta_j(z_k) = \text{diag}(\theta_j^{(p)}) v_j(k) - \frac{\theta_j^{(p)} \exp(-z_k^T \theta_j^{(p)})}{1 - \exp(-z_k^T \theta_j^{(p)})}.$$  

The convex sub-problem for $\theta_j$ with $z_k = z_k^{(p)}$ is

$$\theta_j^{(p+1)} = \arg \min_{\theta_j \in \mathbb{R}^d} Q^\theta_j(\theta_j),$$

where $Q^\theta_j(\theta_j)$ is obtained by fixing $z_k = z_k^{(p)}$ in (21) as follows:

$$Q^\theta_j(\theta_j) = \sum_{k=1}^{K} \theta_j^T \text{diag}(z_k^{(p)}) v_j(k) - \log(1 - \exp(-\theta_j^T z_k^{(p)})) - c_{kj}^{(p)} \theta_j^T z_k^{(p)}.$$  

Since $Q^\theta_j(\theta_j) = \sum_{k=1}^{K} Q^\theta_k(\theta_j)$, where

$$Q^\theta_k(\theta_j) = \theta_j^T \text{diag}(z_k^{(p)}) v_j(k) - \log(1 - \exp(-\theta_j^T z_k^{(p)})) - c_{kj}^{(p)} \theta_j^T z_k^{(p)},$$

we employ SGD to obtain the $k$-th update for $\theta_j$ given by

$$\theta_j^{(p+1)} \triangleq \theta_j^{(p)} + \frac{\eta_j}{K} \nabla_{\theta_j} Q^\theta_k(\theta_j^{(p)}),$$

where $\eta_j$ is another learning rate and $k \in [K]$. Here

$$\nabla_{\theta_j} Q^\theta_k(\theta_j^{(p)}) \text{ can be computed as}$$

$$\nabla_{\theta_j} Q^\theta_k(\theta_j) = \text{diag}(z_k^{(p)}) v_j(k) - \frac{z_k^{(p)} \exp(-\theta_j^T z_k^{(p)})}{1 - \exp(-\theta_j^T z_k^{(p)})}.$$  

It is known that SGD converges to a local minimizer of the original objective [29], [30]. The proposed ACS algorithm is summarized in Algorithm 1. It is possible to incorporate other convergence criterion. For instance, we can declare that the ACS algorithm has converged if $z$ does not change significantly in successive estimation steps, or when the maximum absolute difference in the estimates of $\theta$ over successive iterations is below a fixed and small threshold.

Algorithm 1 Alternate Convex Search Algorithm

Require: Obtain $z_k$ and $\theta_j$.

Input: Kendall-Tau distance $v$, $M$, $p_{max}$.

Output: $z_k$ and $\theta_j$ after $p$ iterations.

1: Initialization $\theta_j^{(0)}, z_k^{(0)} (m) = 1/M$.
2: while $p \leq p_{max}$ do
3: Update $\{z_k\}_{k=1}^{K}$ using (26) with a fixed $\theta_j^{(p)}$.
4: Update $\{\theta_j\}_{j=1}^{J}$ using (31) with a fixed $z_k^{(p)}$.
5: Update $p = p + 1$.
6: end while

4.3 Cramér-Rao Lower Bounds (CRLBs)

In this section, we derive CRLBs which are the fundamental limits on the mean square error (MSE) performance of unbiased estimators. In particular, we derive the CRLB [31] on the MSE of the full-permutation ranking estimator given by $\hat{\lambda} = [\hat{\theta}_1(v), \ldots, \hat{\theta}_{n-1}(v), \hat{z}_1(v), \ldots, \hat{z}_{K}(v)]$, where $\hat{\theta}_j(v)$ is an estimator of $\theta_j$, $\hat{z}_k(v)$ is an estimator of $z_k$, $v = [v(1), \ldots, v(K)]$. The MSE of such an estimator is defined as

$$\mathbb{E}[\lambda - \hat{\lambda}(v)(\lambda - \hat{\lambda}(v))^T].$$  

where $v$ is distributed as in (17). We compute the Fisher information matrix (FIM) in terms of every univariate component of the Mallows model indexed by $j$. The overall FIM can be stated as

$$\mathbb{I}^\lambda \triangleq -\mathbb{E}_{v} \left[ \nabla_{\lambda} \nabla_{\lambda}^T \log \rho(v|\lambda, z) \right].$$  

In the following theorem, we provide a lower bound on the covariance matrix $\mathbb{E}^\lambda$ using the Fisher Information matrix, $\mathbb{I}^\lambda$.

**Theorem 1.** Let $\xi_{kj} = \sum_{m=1}^{M} \theta_{m:j} z_{k}(m)$, $\psi_j(\xi_{kj}) = \frac{\partial \psi_j(\xi_{kj})}{\partial \xi_{kj}}$ and $\psi_j(\xi_{kj}) = \frac{\partial \psi_j(\xi_{kj})}{\partial \xi_{kj}}$. For an unbiased estimator $\hat{\lambda}(v)$, the CRLBs on the MSE matrix $\mathbb{E}^\lambda$ of the unknown deterministic vector $\lambda = [\theta_1, \ldots, \theta_{n-1}, z_1, \ldots, z_K]$ are given by $\mathbb{E}^\lambda \succeq (\mathbb{I}^\lambda)^{-1}$, where

$$\mathbb{I}^\lambda = \begin{bmatrix} \mathbb{I}^\theta & \mathbb{I}^{\theta, z} \\ \mathbb{I}^{z, \theta} & \mathbb{I}^z \end{bmatrix}.$$  

where $\mathbf{I}^0$ is a block diagonal matrix, with the $(p,q)$-th entry of the $j$-th block-diagonal matrix $\mathbf{I}^0_j$, is given by

$$
\mathbf{I}^0_j(p,q) = -\sum_{k=1}^{K} z_k(p) z_k(q) \left( \frac{(\psi_j'(\xi_{kj}))^2}{(\psi_j(\xi_{kj}))^2} - \frac{(\psi_j''(\xi_{kj}))}{\psi_j(\xi_{kj})} \right). 
$$

(36)

The matrix $\mathbf{I}^{0,z}_j$ is block-diagonal, where the $(p,q)$-th entry of the $k$-th diagonal matrix $\mathbf{I}^{0,z}_{j,k}(p,q)$ is given by

$$
\mathbf{I}^{0,z}_j(p,q) = \mu_j \delta(p-q) + \theta_{pj} z_k(q) \left( \frac{(\psi_j'(\xi_{kj}))^2}{(\psi_j(\xi_{kj}))^2} - \frac{(\psi_j''(\xi_{kj}))}{\psi_j(\xi_{kj})} \right) + \frac{(\psi_j'(\xi_{kj}))}{\psi_j(\xi_{kj})} \delta(p-q),
$$

(37)

where $\mu_{pj} = \mathbb{E}[v_{pj}(k)]$ is the expectation of $v_{pj}(k)$ w.r.t. the mixture distribution, $\delta(p-q) = 1$ if $p = q$ and 0 otherwise.

Further, $\mathbf{I}^1$ is a block diagonal matrix, where the $(p,q)$-th entry of the $k$-th diagonal matrix $\mathbf{I}^1_k(p,q)$ is given by

$$
\mathbf{I}^1_k(p,q) = -\sum_{j=1}^{n-1} \theta_{pj} \theta_{qj} \left( \frac{(\psi_j'(\xi_{kj}))^2}{(\psi_j(\xi_{kj}))^2} - \frac{(\psi_j''(\xi_{kj}))}{\psi_j(\xi_{kj})} \right).
$$

(38)

Proof: See Appendix C.

It can be seen from (36), that the lower bound is non-zero only when $z_k(p)$ or $z_k(q)$ is non-zero, i.e., if the sample belongs to the tilted distribution that consists of the $p$-th and the $q$-th component distribution. Note that if $z_k(p)$ takes the value 1 or 0, then the bound (36) is non-zero for the index $p$ for which $z_k(p) = 1$, with $p = q$.

The theorem stated above also allows us to propose novel lower bounds for the homogenous ranking model proposed in [12]. As noted previously, the homogenous ranking model proposed in [12] is as a special case of the $\mathbb{M}$ proposed in this paper, with $M = 1$ and accordingly $z_k(1) = 1$ for all $k$. We state the following corollary which leads to the lower bound on the ML estimator proposed in [12].

**Corollary 1.** For an unbiased estimator $\hat{\theta}_j(v_j)$, the CRBs on the MSE matrix, $\mathbf{E}^0_j$ of the unknown deterministic parameter $\theta_j$ is given by $\mathbf{E}^0_j \succeq (\mathbf{I}^1_j)^{-1}$, where $\mathbf{I}^1_j$ is given by

$$
\mathbf{I}^1_j = -K \left[ \frac{(\psi_j'(\theta_j))^2}{(\psi_j(\theta_j))^2} - \frac{(\psi_j''(\theta_j))}{\psi_j(\theta_j)} \right].
$$

(39)

The proof follows from the proof of Theorem 1, setting $M = 1$ and $z_k(1) = 1$ for all $k$. Large values of the parameters $\theta_j$ result in a Mallows model with smaller variance, leading to more accurate parameter estimation. In (39), we see that the CRLB is determined by the factor $\frac{(\psi_j'(\theta_j))^2}{(\psi_j(\theta_j))^2} - \frac{(\psi_j''(\theta_j))}{\psi_j(\theta_j)}$, which increases as $\theta_j$ increases. Hence, the lower bound given in (39) predicts lower MSE for components with smaller variance.

### 4.4 Estimation of Ground Truth Permutations

In this section, we consider the problem of estimating the ground truth permutations $\pi_{ml}$ for $m \in [M]$. This is a well-known combinatorial optimization problem and several heuristics have been proposed to solve it approximately [9].

In the context of $\mathbb{M}$, the optimization problem for estimating the ground truth permutation is

$$
(\hat{\pi}_1, \ldots, \hat{\pi}_{M}) = \arg\min_{[\pi_{m_1}, \ldots, \pi_{m_M}] \in \Omega^M} \sum_{k=1}^{M} z_k(m) \theta^2_{ml} v_m(k),
$$

(40)

where $v_m(k)$ is a function of $\pi_{ml}$ for all $m$, as given in (3). We propose a greedy approach for joint estimation of the ground truth permutation, $z$ and $\theta$, along the lines of the algorithms proposed in [12]. We assume that we have as input $M$ matrices $Q(1), \ldots, Q(M)$, where the $m$-th matrix is obtained from a homogeneous population with ground truth permutation $\pi_{ml}$ as follows:

$$
Q_{jl}^{(m)}(\pi_{km}) = \frac{1}{K_m} \sum_{k=1}^{K_m} 1\{j < \pi_{kl}\}. \tag{41}
$$

That is, $Q_{jl}^{(m)}(\pi_{km})$ is the empirical probability that $j$ precedes $l$ in the sample $\pi_{km}$, where the sample consists of $K_m$ permutations from the $m$-th component distribution. The mean of $v_m$ under the sampling distribution is given by

$$
\bar{v}_m = \sum_{j \neq \ell} Q_{jl}^{(m)}, \text{ whenever } \pi^{-1}_{m0}(1) = r, \tag{42}
$$

and hence, $\pi^{-1}_{m0}(1) = \arg\min_{j} \sum_{j \neq \ell} Q_{jl}^{(m)}$.

A tree-based search algorithm is derived by extending the above idea to all $j$. The $n!$ nodes of the tree represent all possible orderings of $m$ given by $\rho_{mj} = (r_{m1}, \ldots, r_{mj})$, i.e., each node has $n - j$ children. Hence, a particular level $j$ of the search tree corresponds to the $j$-th position in the sample permutation, where $j \in [n-1]$. Further, any path of length $n$ through the tree starting from the root represents a permutation. Given a ground truth permutation $\pi_{ml}$ and associated parameters of the model $\theta_m = \{\theta_{m1}, \ldots, \theta_{m(n-1)}\}$, the cost at node $\rho_{mj}$ is given by

$$
C_m(r_{m1}, \ldots, r_{mj}) = \sum_{l=1}^{j} z(m) \theta_{ml} v_m(r_{m1}, \ldots, r_{ml}), \tag{43}
$$

where $v_m(r_{m1}, \ldots, r_{ml}) = \sum_{l \neq m} Q_{lj}^{(m)}$. The proposed algorithm chooses the permutation that leads to smallest cost at each level $j$.

The challenge in estimating $\theta_j$ in such a scenario is computing the term $\sum_{k=1}^{K_m} \sum_{m=1}^{M} z_k(m) \theta^2_{ml} v_m(k)$ as $v_m(k)$ depends on the unknown ground-truth permutation. Note that the mean of $\sum_{k=1}^{K_m} \sum_{m=1}^{M} z_k(m) \theta^2_{ml} v_m(k)$ under the sampling distribution is given by $\sum_{j \neq \ell} z(mj) Q_{jl}^{(m)}$, where $z(mj)$ and $z(ml)$ are the weights of the samples from the multiplicative mixture distribution consisting of $m_1$ and $m_2$ component distributions. We employ the mean as a proxy for the term $\sum_{k=1}^{K_m} \sum_{m=1}^{M} z_k(m) \theta^2_{ml} v_m(k)$ in (26) and (31). Further, note that $v_{mj}$ in (26) and (31) can be replaced by $Q_{jl}^{(m)}$, where
the index \( r \) is obtained by minimizing the cost at the \( j \)-th level.

The steps of the proposed \( M4 \)-SEARCHPI algorithm are given in Algorithm 2. Note that unlike the ACS algorithm proposed in the previous section, the per-sample weights \( z_k \) are updated for \( j \) SGD iterations at the \( j \)-th level. The computation of \( A \) is described in [12].

**Algorithm 2** \( M4 \)-SEARCHPI

**Require:** Obtain \((Q^{(1)}, \ldots, Q^{(M)})\) from \((\pi_{10}, \ldots, \pi_{M0}, \theta, z)\).

**Input:** \( Q^{(1)}, \ldots, Q^{(M)}, p_{\text{max}} \).

**Output:** \( \pi_{10}, \ldots, \pi_{M0}, \theta \) and \( z \).

1. **Initialization** \( \theta^{(0)}_m(m) \) for all \( m, j \), \( z^{(0)}_k(m) = 1/M \) for all \( m, k \).
2. Compute \( \hat{V} = \sum_{j \neq c} \sum_{m=1}^{M} z(m)Q^{(m)}_{j\ell} \).
3. while \( |\rho_{m,j}| < n \forall m \) do
4. for \( m = 1 : M \) do
5. for \( r_{m(j+1)} \in [n] \setminus \rho_{m,j} \) do
6. Create node \( \rho' = \rho_{m,j}, r_{m(j+1)} \).
7. Compute \( V_{m(j+1)}(\rho') = \sum_{l \in [n] \setminus \rho'} Q^{(m)}_{l\ell j} \).
8. end for
9. end for
10. while \( \rho \leq p_{\text{max}} \) do
11. \( j = |\rho'| \).
12. Solve (26) using \( \hat{V} \), with a fixed \( \theta^{(j)} \).
13. Solve (31) using \( \hat{V} \), with a fixed \( z^{(j)}_k(m) \).
14. Update \( p = p + 1 \).
15. end while
16. for \( m = 1 : M \) do
17. Compute \( C_m(\rho') \) and \( L_m(\rho') = C_m(\rho') + A \).
18. \( \rho_{m(j+1)} = \arg \min_{\rho \in S_m} L_m(\rho) \).
19. end for
20. end while

5 EXTENSION OF \( M4 \) TO \( TOP-t \)

Often, only the top-\( t \) highly rated items of the consensus ranking are relevant, and this allows us to build simpler models since the observations need not be full permutations. Here, we extend \( M4 \) to the top-\( t \) ranking problem where the goal is the recovery of the \( t \) most highly ranked entries, and the associated dispersion parameters. In [21], the authors propose additive mixture models for obtaining the top-\( t \) consensus ranking. However, we focus on the IGM for ranking in the top-\( t \) ranking problem [13], and derive an inference framework for \( M4 \) for IGM. Subsequently, we show that the multiplicative mixture framework provides a better goodness-of-fit as compared to the additive mixture models [21].

5.1 Mallows Model for Top-\( t \) Ranking Problem

We now derive the \( M4 \) using IGM as the ranking model. The IGM model is given in Sec. 2.2 where recall that \( \sigma \) and \( \sigma_0 \) are length-\( t \) ranking orders. Using the multiplicative mixture framework explored in the context of \( M4 \) for full permutations in Sec. 3, the distribution of the \( k \)-th sample \( s_m(k) \) conditioned on \( \gamma \) and \( z_k \) is given by

\[
p(s_m(k) | \gamma, z_k) = \frac{1}{c^*(z_k, \gamma)} \prod_{m=1}^{M} p_m(s_m(k) | \gamma_m) z_k(m),
\]

where \( c^*(z_k, \gamma) \) is the partition function independent of \( s_m(k) \). Here \( \gamma = [\gamma_0, \ldots, \gamma_M]^T \) and \( \gamma = [\gamma_1, \ldots, \gamma_{M-1}] \in \mathbb{R}^{M \times (t-1)} \). Furthermore, the partition function can be written as \( c^*(z_k, \gamma) = \prod_{j=1}^{M} c^*_j(z_k, \gamma_j) \) where each factor is

\[
c^*_j(z_k, \gamma_j) = \sum_{s(m)(k)} \prod_{m=1}^{M} p_j(s_m(k)) z_k(m).
\]

This results in the following partition function

\[
c^*_j(z_k, \gamma_j) = \Lambda_j \left( \sum_{m=1}^{M} z_k(m) \gamma_m \right) \prod_{m=1}^{M} \Lambda_j(\gamma_m) z_k(m)^{-1}.
\]

Thus, the overall mixture distribution of the \( k \)-th sample is

\[
p(s(k) | \gamma, z_k) = \frac{\exp \left\{ - \sum_{m=1}^{M} z_k(m) \sum_{j=1}^{t-1} \gamma_m s_j(m) \right\}}{\prod_{j=1}^{t} \Lambda_j \left( \sum_{m=1}^{M} z_k(m) \gamma_m \right)},
\]

where \( s(k) = [s_1(k), \ldots, s_M(k)] \) and \( z_k = [z_1, \ldots, z_K] \). The top-\( t \) IGM ranking model was proposed for a homogeneous population [13]. As expected, (46) is a generalization of the IGM in which \( M = 1 \) and \( z_k(1) = 1 \) for all \( k \).

5.2 Inference in the Top-\( t \) Ranking Model

In the context of top-\( t \) ranking problem, we propose novel inference techniques for the top-\( t \) \( M4 \) model proposed in the previous section. The log-likelihood for the top-\( t \) \( M4 \) model is

\[
\ell(z, \gamma) = \sum_{k=1}^{K} \log p(s(k) | \gamma, z_k)
\]

\[
= - \sum_{k=1}^{K} \sum_{m=1}^{M} z_k(m) \gamma_m s_j(m) + \log \left( 1 - \exp \left[ - \sum_{m=1}^{M} z_k(m) \gamma_m \right] \right) .
\]

Using an exponential model with infinite permutations (IGM) simplifies the hitherto intractable partition function, and as a result, the log-likelihood function. Hence, the log-likelihood given above has fewer terms compared to (18).

Rewriting (47) in terms of the \( \beta_k \in \mathbb{R}_+^M \) whose entries are given by \( \beta_k(m) = z_k(m) \gamma_m \), the optimization problem given above can be expressed as \( \ell(z, \gamma) = - \sum_{k=1}^{K} \sum_{j=1}^{t-1} \ell(\beta_k) \), where

\[
\ell(\beta_k) = \beta_k^T s_j(k) - \log \left( 1 - \exp \left[ - \sum_{m=1}^{M} \beta_k(m) \right] \right) .
\]

**Proposition 3.** The function \( \ell(\beta_k) \) is convex in \( \beta_k(m) \).

**Proof:** The proof follows from the proof of Proposition 1 since \( \ell(\beta_k) \) is similar to the convex component of \( \ell(\alpha_k) \) given by \( f(\alpha_k) \).

In addition to the convexity property stated above, we exploit the bi-convexity of \( \ell(\beta_k) \) in order to obtain the estimates of \( z_k(m) \) and \( \gamma_m \), and iteratively solve for \( z_k(m) \) and \( \gamma_m \). The property is stated in the following proposition.
Proposition 4. The function \( (z_k, \gamma_j) \mapsto \ell(\beta_k, \beta_j) \) is biconvex in \( z_k \) and \( \gamma_j \). The feasible set \( \Gamma \times Z \) is also biconvex.

Proof: The proof follows from the proof of Proposition 2.

As stated in the previous section, we solve the biconvex problem using the ACS approach. In the context of the top-\( t \) M4, the convex sub-problem for optimizing \( z_k \) given \( \gamma_j = \gamma_j^{(p)} \) is

\[
z_k^{(p+1)} = \arg\min_{z_k \in \mathbb{R}^{d_t}} \ell^t(z_k), \quad \text{where} \quad \ell^t(z_k) = \sum_{j=1}^{t-1} z_k^T \text{diag}(\gamma_j^{(p)}) s_j(k) - \log(1 - \exp(-z_k^T \gamma_j^{(p)})).
\]

The optimization problem given above is a finite-sum problem, i.e., \( \ell^t(z_k) = \sum_{j=1}^{t-1} \ell_j^t(z_k) \) where

\[
\ell_j^t(z_k) = z_k^T \text{diag}(\gamma_j^{(p)}) s_j(k) - \log(1 - \exp(-z_k^T \gamma_j^{(p)})),
\]

Using the SGD approach, the \( j \)-th (where \( j \in [t-1] \)) update for \( z_k \) is

\[
z_k^{(p+1/2)} = z_k^{(p)} - \eta_z \nabla_{z_k} \ell_j^t(z_k) = \left(z_k^{(p)} - \eta_z \nabla_{z_k} \ell_j^t(z_k)\right),
\]

where \( \eta_z \) is a learning rate parameter. In the above,

\[
\nabla_{z_k} \ell_j^t(z_k) = \text{diag}(\gamma_j^{(p)}) s_j(k) - \frac{\gamma_j^{(p)} \exp(-z_k^T \gamma_j^{(p)})}{1 - \exp(-z_k^T \gamma_j^{(p)})}
\]

In the next step of the iterative algorithm, we solve the convex sub-problem for \( \gamma_j \) with \( z_k = z_k^{(p)} \) which is given by

\[
\gamma_j^{(p+1)} = \arg\min_{\gamma_j \in \mathbb{R}^d} \ell^t(\gamma_j), \quad \text{where} \quad \ell^t(\gamma_j) = \sum_{k=1}^K \ell_k^t(\gamma_j) = \sum_{k=1}^K \gamma_j^T \text{diag}(z_k^{(p)}) s_j(k) - \log(1 - \exp(-z_k^T \gamma_j^{(p)})),
\]

Since the optimization problem given above is a finite-sum problem, i.e., \( \ell^t(\gamma_j) = \sum_{k=1}^K \ell_k^t(\gamma_j) \), where

\[
\ell_k^t(\gamma_j) = \gamma_j^T \text{diag}(z_k^{(p)}) s_j(k) - \log(1 - \exp(-z_k^T \gamma_j^{(p)})),
\]

it allows us to employ the SGD approach where the \( k \)-th update for \( \gamma_j \), given by

\[
\gamma_j^{(p+1)} = \gamma_j^{(p)} - \frac{\lambda}{K} \nabla_{\gamma_j} \ell_k^t(\gamma_j) = \gamma_j^{(p)} - \frac{\lambda}{K} \nabla_{\gamma_j} \ell_k^t(\gamma_j),
\]

where \( \lambda \) is the learning rate and \( k \in [K] \). Here \( \nabla_{\gamma_j} \ell_k^t(\gamma_j) = \text{diag}(z_k^{(p)}) s_j(k) - \frac{z_k^{(p)} \exp(-\gamma_j^T z_k^{(p)})}{1 - \exp(-\gamma_j^T z_k^{(p)})} \).

The CRLBs for the partial ranking case follows from that derived in Sec. 4.3, where the normalization constant \( \psi(\theta) \) is replaced by \( \Lambda(\gamma) \) in the partial ranking scenario. We also note that \( s_{mj} \) is a geometric random variable, and hence \( \mathbb{E}[s_{mj}(k)] = (\exp(\theta_{mj}) - 1)^{-1} \).

5.3 Estimation of Top-\( t \) Set

We consider the problem of estimating the consensus top-\( t \) ranking of each cluster given by \( \sigma_{m0} \) for \( m \in [M] \). This optimization problem in the context of M4 is given by

\[
(\hat{\sigma}_m, \ldots, \hat{\sigma}_M) = \arg\min_{(\sigma_0, \ldots, \sigma_M) \in \Omega^{M}} \sum_{k=1}^K \sum_{m=1}^M z_k(m) \gamma_m^T s_m(k),
\]

where \( s_m(k) \) is a function of \( \sigma_{m0} \). Similar to Sec. 4.4, we propose a greedy approach for joint estimation of the ground truth permutation, \( z \) and \( \gamma \), along the lines of the algorithms proposed in [13]. We assume that we have as input M matrices \( R^{(1)}, \ldots, R^{(M)} \). The \( j \)-th component matrix of the \( m \)-th cluster is obtained from a homogeneous population with ground truth permutation \( \sigma_{m0} \) as follows:

\[
R^{(m)}_j = q_{mj}^T \mathbf{I} - \mathbf{P}_{mj},
\]

where \( q_{mj} = [q_{0j}], i \in \mathbb{N} \) and \( [q_{0j}] \) represents the number of times \( i \) is observed in rank \( j \). Further, \( \mathbf{P}_{mj} = [p_{ij}], i', j' \in \mathbb{N} \) is a matrix whose element \( p_{ij} \) represents the number of times \( \sigma(i) = j \) and \( \sigma(i') < j \). The sufficient statistic matrix for the \( m \)-th cluster is

\[
R^{(m)} = \sum_{j=1}^{l-1} \gamma_{mj} R^{(m)}_j
\]

Similar to the algorithm proposed in Sec. 4.4, a tree-based search algorithm is derived with the nodes representing partial orderings of \( \sigma_{m0} \) given by \( \omega_m = \{w_1, \ldots, w_{m3}\} \), i.e., each node has \( n - j \) children. The cost at node \( \omega_m \) is given by

\[
C_m(w_{m1}, \ldots, w_{mj}) = \sum_{l=1}^{j} z(l) \gamma_{mj} W_{ml}(w_{m1}, \ldots, w_{ml}),
\]

where \( W_{ml}(w_{m1}, \ldots, w_{ml}) = \sum_{i \in [l]} [w_{m1}, \ldots, w_{ml}] R^{(m)}_l \). The proposed algorithm chooses the partial ranking that leads to smallest cost at each level \( j \). Unlike the M4-SEARCHPI algorithm, we do not obtain the estimates of \( \gamma \) at every level of the tree. Instead, we obtain estimates \( \sigma_{m0} \) and \( \gamma \) separately and iterate between these steps until convergence. The estimate of \( \gamma \) is obtained using the ACS steps in (57) and (52).

The steps of the TOP-\( t \) M4-SEARCHPI algorithm are given in Algorithm 3. The computation of \( \Lambda \) is described in [12].

6 NUMERICAL EXPERIMENTS

In this section, we numerically demonstrate the goodness-of-fit of the M4 using the popular American Psychological Association (APA) [32] and Sushi datasets [33]. We also demonstrate the parameter and permutation learning capabilities of the inference algorithms proposed in Secs. 4 and 5 on synthetic datasets.

6.1 Goodness-of-fit: APA and Sushi Dataset

In this section, we use the M4 to model the real-world APA presidential election dataset [32] and the Sushi dataset [33]. Our goal is to highlight the advantages of M4 as compared
to the homogeneous Mallows model [34] and the additive mixture Mallows model [15], [16]. For both datasets, we use the BIC and ICL [20] as measures of the goodness-of-fit. The BIC is defined as

$$\text{BIC}(G) = 2\ell_G(\mathbf{z}, \mathbf{\theta}) - v_G \log(K),$$  \hspace{1cm} (63)$$

where $\ell_G(\mathbf{z}, \mathbf{\theta})$ represents the log likelihood under the model $G$ and $v_G$ represents the number of free parameters in $G$. The ICL is used to measure the separation of the mixture components and is a popular criterion for clustering applications. The ICL is defined as

$$\text{ICL}(G) = \text{BIC}(G) - 2 \sum_{k=1}^{K} \text{Entropy}(\mathbf{z}_k),$$  \hspace{1cm} (64)$$

where $\text{Entropy}(\mathbf{z}_k) = - \sum_{m=1}^{M} z_k(m) \log(z_k(m))$. We choose to use the ICL measure to highlight the primary difference between the additive and multiplicative mixture models. In additive mixture models, every sample is represented using $M$ components, whereas in the multiplicative mixture models, every sample is represented using only 2 component mixtures. Hence, as $M$ increases, the penalty in the case of additive mixture models is larger, especially if the components have small membership weights.

### 6.1.1 APA Dataset

The 1980, the APA presidential election consisted of five candidates (A, B, C, D, E) and voters were asked to rank the candidates in their order of preference. Among the 15449 votes that were cast, 5738 voters ranked all five candidates [32]. We demonstrate the goodness-of-fit of the $M_4$ for permutation-based observations using the 5738 permutations. We also demonstrate the goodness-of-fit of the $M_4$ for top-$r$ ranking observations using all of the 15449 votes as observation samples.

We simulate the additive mixture Mallows model using the EM algorithm for comparison. Note that the number of free parameters in the $M_4$ is given by $M(n - 1) + 2M$. Using the BIC we declare the model with the largest BIC value as the best model. In the APA dataset, the candidates can be divided into one of 3 types (candidates A and C are research psychologists, D and E are clinical psychologists and candidate B is a community psychologist), and hence we set $M = 2$ and $M = 3$ for the purpose of this study. The nature of preference of voters is shown in Fig. 1.

Using the M-M based inference algorithm, we fit the $M_4$ to this dataset, and computed the per-sample weights and the parameters, for $M = 2$ and $M = 3$. We also ran the EM algorithm to learn the additive mixture Mallows model [15], [16] and the homogeneous Mallows model [12] (with $M = 1$). We used the initialization parameters as suggested in [15], with 20 iterations in every realization. In the E-step, the membership weights for each cluster is computed, and the parameters are estimated in the M-step. It can be seen from Tables 2 and 3 that for both $M = 2$ and $M = 3$, $M_4$ provides a better fit compared to the additive mixture model in both the complete and partial rankings scenarios.

When $M = 3$, we obtain a lower value of BIC and ICL indicating that a 3 component model is better suited for this dataset. This corroborates the fact that the voting criterion is based on whether the voters prefer research psychologists, clinical psychologists or community psychologists, and the model of the overall mixture distribution tends to have one of the 3 types of candidates as the most preferred choice.

These trends also follow from Fig. 1. From this experiment, we see that the permutations that rank community psychologists or research psychologists in the first position have the maximum number of votes. Hence, the permutations that
have $B$, $A$ or $C$ in the first position tend to be the ground truth permutations as listed in Table 1.

6.1.2 Sushi Dataset

We now fit the M4 to the popular Sushi dataset. This dataset compares 10 types of sushi namely shrimp, sea eel, tuna, squid, sea urchin, salmon roe, egg, fatty tuna, tuna roll, and cucumber roll. The data was collected by surveying 5000 individuals living in Japan about their preferences regarding the sushi variants in the form of ranks assigned to the sushi variants in the order of their preference [33].

It is observed from Fig. 2 that Fatty tuna is a common favorite and hence tends to be ranked highest, while cucumber roll is the least liked. However, the voters seem to have a divided opinion about sea urchin as 15-20% of the voters rank it as their most favourite or their least favourite item, hinting that this dataset is suited for mixture modelling. Furthermore, it is possible to obtain a Condorcet ranking with fatty tuna in the first position and cucumber roll in the last position [35].

The global search for the candidate permutation that leads to the smallest value of BIC is infeasible as the number of possible permutation choices is too large. We pick a small subset of permutations such that it consists of the Condorcet permutation, other permutations with the Kendall-Tau distance of one to the Condorcet permutation, and some permutations that capture the voters’ divided opinion on sea urchin. We obtain the BIC values for the subset of permutations as given in Table 4. In the case of the Sushi dataset, the EM algorithm for additive mixture density necessitates a weighted sum over $10^7$ permutations, which is infeasible unless we make changes to the original approach [16]. However, we are able to model the heterogeneous population using M4, hence substantiating our claim that the inference framework for the M4 is indeed computationally simple.

6.2 Parameter and Ground Truth Estimation Performance on Synthetic Datasets

In this section, we demonstrate the MSE and the permutation recovery performance of the proposed inference algorithms. We generated a synthetic dataset with $n = 8$ using Gibbs sampling. The ground truth permutation was chosen randomly, ensuring that the distance between permutation is sufficiently large [14]. The experiments are repeated over 100 trials. Note that we study the performance of the algorithms in scenarios where the number of components $M$ in the mixture distribution is known a priori and hence, we do not consider the model fitting errors that may arise due to uncertainty in the number of clusters.

6.2.1 Permutation-Based Ranking Observations

Here, we numerically study the M4 and the ground truth recovery performance of the inference algorithms with observations being permutations. We set $\theta_{1m} \geq \theta_{2m} \geq \ldots \theta_{(n-1)m} > 0$ for each $m$. That is, $\theta_i$ has comparatively smaller values at large $j$, signifying that the most important stages of ranking are the highly ranked elements of $\pi_{n0}$.

In Fig. 3, we demonstrate the MSE performance of the proposed algorithm for $M = 1, 2$ and $4$ for varying number of observation samples. In our experiments, we noted that the M-M algorithm is sensitive to initialization and learning rate w.r.t. $\theta$, namely $\eta_\theta$. In all our experiments, we set $\eta_\theta \propto \beta^{-\gamma}$, where $i$ is the iteration number of the SGD algorithm and $\beta = 2$. We used random shuffling of observations before applying the SGD algorithm [30]. We see that MSE increases with $M$ for a given sample size, as the number of parameters grows with $M$. However, we note that the MSE decreases with increasing sample size $K$. In fact, it can be observed empirically that $\text{MSE} \propto \frac{1}{\sqrt{K}}$ (MSE is halved as $K$ quadruples), as expected from the central limit theorem.

In Fig. 4, we depict the efficacy of the M4-SEARCHPI approach for ground truth permutation estimation. We display the success rates of the proposed algorithm, where success rate refers to the average number of times the algorithm recovers the ground truth permutation perfectly ($\hat{\pi}_{n0} = \pi_{n0}$) for $M = 1, 2, 4$. We observe that as $K$ increases, the success rate also increases. In addition, we observe that the deviation of $\hat{\theta}$ as compared to $\theta$ (as seen from the MSE performance of the M-M approach in Fig. 3) is not reflected in the performance of the M4-SEARCHPI algorithm in terms of estimating the true permutations $\pi_{n0}$.

6.2.2 Top-t Ranking Observations

Here, we demonstrate the M4 and the ground truth recovery performance of the inference algorithms with observations being the top-t rankings, where $t = 5$ and $n = 8$. Hence we are only concerned with the identities of the top 5 out of 8 items. We set $\gamma_{1m} \geq \gamma_{2m} \geq \ldots \gamma_{(t-1)m} > 0$ for each...
Figure 2. Position preferences of the 10 Sushi types in the Sushi dataset.

Figure 3. Plot of the MSE of $\hat{\theta}_{m_j}(v)$ as a function of the number of samples.

Figure 4. Plot of the success rates of the $M4$-SEARCHPI algorithm.

In Fig. 5 we demonstrate the MSE performance of the proposed algorithm for $M = 1, 2$ and 3 for varying number of observation samples, when the observations are top-5 ranking orders. We use the same algorithmic setting as in the previous sections. Overall, the MSE obtained is lower than the full permutation case, as expected. Further, the CRLB for the top-$t$ ranking scenario is lower than the CRLB in the permutation scenario, indicating that one may expect lower MSE in the top-$t$ ranking scenarios.

In Fig. 6, we show the efficacy of the TOP-t $M4$-SEARCHPI approach for top-$t$ ranking estimation. We display the success rates of the proposed algorithm for $M = 1, 2, 3$. We observe that as $K$ increases, the success rate also increases. Also as expected, the sample size $K$ required to attain the perfect top-$t$ ranking is smaller compared to obtaining the full ground truth permutations as we are only concerned with a subset of the items. Finally, as $M$ increases,
the success rate decreases as there are more parameters to estimate.

To prove that the functions \( f(\alpha_{k,j}) \) and \( g(\alpha_{k,j}) \) are convex, we derive their Hessians and verify that it is positive semidefinite. The Hessian of \( f(\alpha_{k,j}) \) is given by

\[
\nabla^2_{\alpha_{k,j}} f(\alpha_{k,j}) = c'_{k,j} 1^T 1,
\]

where

\[
c'_{k,j} = \frac{\exp[- \sum_{m=1}^{M} \alpha_{k,j}(m)]}{1 - \exp[- \sum_{m=1}^{M} \alpha_{k,j}(m)]} + \left( \frac{\exp[- \sum_{m=1}^{M} \alpha_{k,j}(m)]}{1 - \exp[- \sum_{m=1}^{M} \alpha_{k,j}(m)]} \right)^2.
\]

Similarly, for \( g(\alpha_{k,j}) \), we have

\[
c''_{k,j} = (n + j + 1)^2 \frac{\exp[-(n-j+1) \sum_{m=1}^{M} \alpha_{k,j}(m)]}{1 - \exp[-(n-j+1) \sum_{m=1}^{M} \alpha_{k,j}(m)]} + \left( \frac{\exp[-(n-j+1) \sum_{m=1}^{M} \alpha_{k,j}(m)]}{1 - \exp[-(n-j+1) \sum_{m=1}^{M} \alpha_{k,j}(m)]} \right)^2.
\]

Since \( \sum_{m=1}^{M} \alpha_{k,j}(m) \) is positive for all \( k, j \), we conclude that \( h^T \nabla^2_{\alpha} f(\alpha_{k,j}) h \geq 0 \) and \( h^T \nabla^2_{\alpha} g(\alpha_{k,j}) h \geq 0 \) for all \( h \in \mathbb{R}^M \).

Using the second-order Hessian condition for convexity [36], we conclude that \( f(\alpha_{k,j}) \) and \( g(\alpha_{k,j}) \) are convex. This completes the proof.

### Appendix B

**Proof of Proposition 2**

To prove that \( Q(\alpha_{k,j}; \alpha_{k,j}^{(p)}) \) is bi-convex in \( z_k \) and \( \theta_j \), we need to prove that the function \( Q(\alpha_{k,j}; \alpha_{k,j}^{(p)}) \) is convex in \( \theta_j \) when \( z_k \) is held constant, and convex in \( z_k \) when \( \theta_j \) is held constant.

In order to prove convexity, we derive the Hessian of \( Q(\alpha_{k,j}; \alpha_{k,j}^{(p)}) \) w.r.t. \( z_k \in Z \) and \( \theta \in \Theta \). The Hessian w.r.t. \( z_k \) and \( \theta \) is given by

\[
\nabla^2_{\alpha_{k,j}} Q(\alpha_{k,j}; \alpha_{k,j}^{(p)}) = \frac{\nabla^2_{\alpha_{k,j}} \log(1 - \exp(-1^T \alpha_{k,j}^{(p)}))}{\nabla^2_{\alpha_{k,j}} \log(1 - \exp(-1^T \alpha_{k,j}))} \left( \frac{\exp(-1^T \alpha_{k,j}^{(p)})}{\nabla^2_{\alpha_{k,j}} \log(1 - \exp(-1^T \alpha_{k,j}))} \right)^2.
\]

Further, the Hessian w.r.t. \( \theta_j \) is given by

\[
\nabla^2_{\theta_j} Q(\alpha_{k,j}; \alpha_{k,j}^{(p)}) = z_k z_k^T \left[ \frac{\exp(-1^T \alpha_{k,j}^{(p)})}{\nabla^2_{\theta_j} \log(1 - \exp(-1^T \alpha_{k,j}))} + \left( \frac{\exp(-1^T \alpha_{k,j})}{\nabla^2_{\theta_j} \log(1 - \exp(-1^T \alpha_{k,j}))} \right)^2 \right].
\]

Since \( \theta \in \mathbb{R}^M \), \( z_k \in \mathbb{R}^M \), \( \theta_{k,j}, z_k(m) \geq 0 \) for all \( m, k \) resulting in \( h^T \nabla^2_{\theta} Q(\alpha_{k,j}) h \geq 0 \) and \( h^T \nabla^2_{\theta} Q(\alpha_{k,j}) h \geq 0 \) for all \( h \in \mathbb{R}^M \), we conclude that \( Q(\alpha_{k,j}; \alpha_{k,j}^{(p)}) \) is bi-convex in \( z_k \) and \( \theta_j \). This completes the proof.

### Appendix C

**Proof of Theorem 1**

To obtain the CRLB, we verify that the underlying Mallows model satisfies the requisite regularity conditions [31]. For \( M = 1 \), the expectation of \( v_j \) is

\[
\mathbb{E}[v_j] = \sum_{\psi_j = 0}^{n-j} \frac{v_j \exp(-\theta_j v_j)}{\psi(\theta_j)} = -\psi'(\theta_j) \psi(\theta_j).
\]
and hence, the regularity conditions are satisfied [31]. The first order derivatives of the log-likelihood function in (18) w.r.t. the unknown parameters $\theta$ and $z$ are given by

$$
\frac{\partial \log p(\mathbf{v} | \theta, \mathbf{z})}{\partial \theta_{pj}} = -\sum_{k=1}^{K} z_k(p) \left( \psi_{k,j}(k) + \psi_{j,k}(\xi_{k,j}) \right),
$$

(72)

$$
\frac{\partial \log p(\mathbf{v} | \theta, \mathbf{z})}{\partial z_k(p)} = -\sum_{j=1}^{n-1} \left( \theta_{pj} \psi_{k,j}(k) + \theta_{jp} \psi_{j,k}(\xi_{k,j}) \right),
$$

(73)

where $\xi_{k,j} = \sum_{m=1}^{M} \theta_{mj} z_k(m)$. Differentiating (72) w.r.t. $\theta_j$, and taking expectation w.r.t. the mixture distribution, we obtain

$$
\mathbf{I}^\theta_j(p, q) = -\sum_{k=1}^{K} z_k(p) z_k(q) \left( \left( \frac{\psi_{k,j}(\xi_{k,j})}{\psi_{j,k}(\xi_{k,j})} \right)^2 - \frac{\psi_{j,k}(\xi_{k,j})}{\psi_{j,k}(\xi_{k,j})} \right).
$$

(74)

Note that the off-block-diagonal entries of $\mathbf{I}^\theta$ are zero since the mixture distribution in (17) factorizes in $j$. Further, differentiating (73) w.r.t. $z_k(q)$, and taking expectation w.r.t. the mixture distribution, we obtain

$$
\mathbf{I}^2_j(p, q) = -\sum_{j=1}^{n-1} \theta_{pj} \theta_{jq} \left( \frac{\psi_{j,k}(\xi_{k,j})}{\psi_{j,k}(\xi_{k,j})} \right)^2 - \frac{\psi_{j,k}(\xi_{k,j})}{\psi_{j,k}(\xi_{k,j})} \delta(p - q).
$$

(75)

Further, differentiating (72) w.r.t. $z_k(q)$, we obtain

$$
\frac{\partial^2 \log p(\mathbf{v} | \theta, \mathbf{z})}{\partial \theta_{pj} \partial z_k(q)} = -\sum_{j=1}^{n-1} \theta_{pj} \left( \frac{\psi_{j,k}(\xi_{k,j})}{\psi_{j,k}(\xi_{k,j})} \right)^2 - \frac{\psi_{j,k}(\xi_{k,j})}{\psi_{j,k}(\xi_{k,j})} \delta(p - q).
$$

(76)

Taking expectation of the above expression w.r.t. the mixture distribution, we obtain

$$
\mathbf{I}^{\theta, \ast}_{j,k}(p, q) = \mu_{pj} \delta(p - q)
$$

(77)

where $\mu_{pj} = \mathbb{E}[v_{pj}(k)]$ is the expectation of $v_{pj}(k)$ w.r.t. the mixture distribution in (17). The off-block-diagonal matrices of $\mathbf{I}^\ast$ and $\mathbf{I}^{\theta, \ast}$ are zero since the observation samples are i.i.d.

**References**


