Fast and Stable Maximum Likelihood Estimation for Incomplete Multinomial Models

Chenyang Zhang [1] **Guosheng Yin** [1]

Abstract

We propose a fixed-point iteration approach to the maximum likelihood estimation for the incomplete multinomial model, which provides a unified framework for ranking data analysis. Incomplete observations typically fall in a subset of categories, and thus cannot be distinguished as belonging to a unique category. We develop a minorization–maximization (MM) type of algorithm, which requires relatively fewer iterations and shorter time to achieve convergence. Under such a general framework, incomplete multinomial models can be reformulated to include several well-known ranking models as special cases, such as the Bradley–Terry, Plackett–Luce models and their variants. The simple form of iteratively updating equations in our algorithm involves only basic matrix operations, which makes it efficient and easy to implement with large data. Experimental results show that our algorithm runs faster than existing methods on synthetic data and real data.

1. Introduction

Multinomial modeling and inference have been widely utilized for polytomous response data analysis in many problems of statistics and machine learning. For example, some ranking models (Luce, 1959) treat ranking results as a finite sequence of multinomial samples; the cross-entropy loss built on the Kullback–Leibler divergence of two multinomial distributions can be used for the training of classification tasks (Krizhevsky et al., 2012).

In a multinomial model, the sample space Ω is partitioned into K disjoint subspaces, where K is the number of categories. However, in most real applications, observed samples are incomplete in the case of partial or conditional classification. For some observations, a subset of categories rather than a unique category is reported, or the set of possible outcomes contains only part of all categories, i.e., truncated data. The probability for an incomplete multinomial observation y_i has the form

$$\Pr(y_i \in C_i | A_i) = \frac{\sum_{j\in C_i} p_j}{\sum_{j\in A_i} p_j}, \tag{1}$$

where $C_i \subset A_i \subseteq \Omega$, $\Omega = \{1,\dots,K\}$ and $\boldsymbol{p} = (p_1,\dots,p_K)^\intercal$ is the parameter of interest. The paired sets (C_i, A_i) record the reported categories and the set of possible outcomes for y_i. An observation is complete if $|C_i| = 1$ ($|\cdot|$ represents the number of elements of a set) and $A_i = \Omega$. The probability of a reported subset can be expressed as a probability sub-sum $\tilde{p} = \boldsymbol{\delta}^\intercal \boldsymbol{p}$, where $\boldsymbol{\delta}$ is an indicator vector corresponding to the composition of each incomplete classification. Given the incomplete multinomial data, Dong & Yin (2018) proposed a weaver algorithm to maximize the likelihood function

$$L(\boldsymbol{p}|\boldsymbol{a},\boldsymbol{b},\boldsymbol{\Delta}) \propto \prod_{k=1}^{K} p_k^{a_k} \prod_{j=1}^{q} \tilde{p}_j^{b_j} = \prod_{k=1}^{K} p_k^{a_k} \prod_{j=1}^{q} (\boldsymbol{\delta}_j^\intercal \boldsymbol{p})^{b_j}, \tag{2}$$

where

- $\boldsymbol{p} = (p_1,\dots,p_K)^\intercal$ collects the probabilities of all categories, representing the parameters of the incomplete multinomial model.
- $\boldsymbol{a} = (a_1,\dots,a_K)^\intercal$ represents the counts of fully classified observations for each category, corresponding to $|C_i| = 1$.
- $\boldsymbol{b} = (b_1,\dots,b_q)^\intercal$ denotes the counts of incomplete observations, where q is the number of observed subsets. Positive terms represent the results of partial classification, and negative terms indicate the counts corresponding to truncated outcomes.
- $\boldsymbol{\Delta} = \{\Delta_{kj}\}_{K\times q} = [\boldsymbol{\delta}_1,\dots,\boldsymbol{\delta}_q]$ is the indicator matrix where each $\boldsymbol{\delta}_j$ is a column vector of indicators representing the element constituents of observed subsets, associated with the count b_j.

[1]Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong. Correspondence to: Guosheng Yin <gyin@hku.hk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

In fact, model (2) is a unified framework which includes the probability mass functions of the Bradley–Terry model (Bradley & Terry, 1952) and Plackett–Luce model (Luce, 1959; Plackett, 1975) for ranking aggregation as special cases. However, the model flexibility is restricted by constraining $\Delta_{kj} \in \{0, 1\}$, and more problems can be reformulated into incomplete multinomial models by relaxing the indicator Δ_{kj} as $\Delta_{kj} \geq 0$. We call the model allowing $\Delta_{kj} \geq 0$ as the 'generalized incomplete multinomial model'.

The rest of this paper is organized as follows. Section 2 introduces related work on incomplete multinomial inference and ranking aggregation analysis. Section 3 presents our fixed-point iteration approach for maximum likelihood estimation and its convergence properties. Simulation results on synthetic data and analysis of real data are given in Section 4.

2. Related Work

Statistical inference for multinomial models in the presence of missing data has a long history. Ireland & Kullback (1968) estimated contingency table cell probabilities when the marginal probabilities are known and fixed. Hocking & Oxspring (1971) considered the problem of maximum likelihood estimation when some of the observations are partially classified. Chen & Fienberg (1976) studied the imperfect cross-classification problem in multi-dimensional contingency tables with totally mixed-up cells. Several Bayesian approaches, such as Dickey et al. (1987) and Paulino & de Braganca Pereira (1995), were developed for solving the problem of categorical responses with censoring. Numerous studies on incomplete multinomial data considered only the partial classification problem, i.e., the counts of merged cells are positive. However, these methods, such as the expectation–maximization (EM) algorithm which splits the merged counts by current cell probabilities (Schafer, 1997), cannot be implemented on the data with negative counts induced by conditional probabilities.

Another type of missingness in multinomial data is caused by conditional classification, which has been widely discussed in ranking aggregation problems, such as the Bradley–Terry model for pairwise comparisons and the Plackett–Luce model for multiple rankings and permutations. To deal with negative terms in the log-likelihood function, Hunter (2004) proposed a minorization–maximization (MM) algorithm for iterative maximum likelihood estimation, which is applicable to a wide class of generalizations of the Bradley–Terry model. Negahban et al. (2012) introduced the ranking centrality algorithm for pairwise ranking data, which interprets ranking relationships between objects as a random walk over the comparison graph. Its generalization to multiple rankings has been discussed in Maystre & Grossglauser (2015) and Agarwal et al. (2018).

Most of the existing work focus on either partial or conditional classification. When both types of missingness occur, for example, partial multiple rankings with ties, Soufiani et al. (2013) proposed to break multiple rankings into pairwise comparisons. Turnbull (1976) developed a self-consistency algorithm to solve the nonparametric estimation of the empirical distribution function with interval-censored and truncated survival data, for which the likelihood has the same form as the incomplete multinomial model. Huang et al. (2006) proposed a new method called group pairwise comparisons to deal with the pairwise comparison problems between subsets of items. Dong & Yin (2018) used an incomplete multinomial model to formulate the generalized multinomial data and derived a fixed-point iteration algorithm by solving a system of equalities satisfied on the stationary point of the likelihood function.

We consider the problem of solving the maximum likelihood estimator (MLE) with incomplete multinomial data, where partial classification and conditional probabilities co-exist. Based on the simple and interpretable incomplete multinomial modeling in (2), we explore the optimality conditions of the log-likelihood function and develop a fixed-point iteration updating rule called the stable weaver algorithm. We investigate the convergence properties of our algorithm, and demonstrate the capabilities of generalization of the incomplete multinomial model which includes many well-known ranking models as special cases.

3. Stable Weaver Algorithm

3.1. MLE from Fixed-Point Iterations

Recall the likelihood function with incomplete multinomial observations $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Delta})$ in (2), the log-likelihood function is

$$\ell(\boldsymbol{p}|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Delta}) = \sum_{k=1}^{K} a_k \log p_k + \sum_{j=1}^{q} b_j \log(\boldsymbol{\delta}_j^{\intercal} \boldsymbol{p}),$$

with the parameter space $\Theta = \{\boldsymbol{p} | p_k > 0, \sum_{k=1}^{K} p_k = 1\}$. Solving the MLE of an incomplete multinomial model can be viewed as an optimization problem,

$$\max_{\boldsymbol{p}} \ \ell(\boldsymbol{p}|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Delta}) \quad \text{subject to} \ \sum_{k=1}^{K} p_k = 1.$$

We can apply the Lagrange multipliers method,

$$\max_{\boldsymbol{p}} \sum_{k=1}^{K} a_k \log p_k + \sum_{j=1}^{q} b_j \log\left(\boldsymbol{\delta}_j^{\intercal} \boldsymbol{p}\right) - \lambda(\mathbf{1}^{\intercal} \boldsymbol{p} - 1), \quad (3)$$

where λ is the Lagrange multiplier.

Theorem 1 *Let $s = \sum_{k=1}^{K} a_k + \sum_{j=1}^{q} b_j$. If $(\boldsymbol{p}^*, \lambda^*)$ is the stationary point of (3), then $\lambda^* = s$.*

Remark 1 *A special situation is that $s = 0$. For observations $\{y_i\}_{i=1}^n$ in (1), the likelihood has the form*

$$L(\boldsymbol{p}|\{C_i, A_i\}_{i=1}^n) = \prod_{i=1}^{n} \left(\frac{\boldsymbol{\delta}_{C_i}^\intercal \boldsymbol{p}}{\boldsymbol{\delta}_{A_i}^\intercal \boldsymbol{p}} \right), \tag{4}$$

where $\boldsymbol{\delta}_{C_i}, \boldsymbol{\delta}_{A_i}$ are indicator vectors corresponding to C_i, A_i. If $s = 0$, $A_i \neq \Omega$, $\forall i$. Then for any $\boldsymbol{p}^* > \mathbf{0}$, we can find another $\boldsymbol{p}^\dagger = \alpha \boldsymbol{p}^*$ $(\alpha > 0)$, such that $L(\boldsymbol{p}^*|\{C_i, A_i\}_{i=1}^n) = L(\boldsymbol{p}^\dagger|\{C_i, A_i\}_{i=1}^n)$. As a result, we need to renormalize $\boldsymbol{p}$ in every iteration for identifiability.

With the result of Theorem 1, our goal is to maximize

$$\begin{aligned} \ell(\boldsymbol{p}|s, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Delta}) \quad = \quad & \sum_{k=1}^{K} a_k \log p_k + \sum_{j=1}^{q} b_j \log \boldsymbol{\delta}_j^\intercal \boldsymbol{p} \\ & - s \left(\sum_{k=1}^{K} p_k - 1 \right). \end{aligned} \tag{5}$$

For terms in $\boldsymbol{b}$, we divide them into positive and negative sets. Let $Q^+ = \{j \mid b_j > 0, j = 1, \ldots, q\}$ and $Q^- = \{j \mid b_j < 0, j = 1, \ldots, q\}$ be the sets of indices of positive and negative elements in $\boldsymbol{b}$ respectively. The optimality condition $\nabla \ell(\boldsymbol{p}) = 0$ implies

$$\frac{\partial \ell}{\partial p_k} = \frac{a_k}{p_k} + \sum_{j \in Q^+} \frac{|b_j| \Delta_{kj}}{\boldsymbol{\delta}_j^\intercal \boldsymbol{p}} - \sum_{j \in Q^-} \frac{|b_j| \Delta_{kj}}{\boldsymbol{\delta}_j^\intercal \boldsymbol{p}} - s = 0,$$

which is equivalent to

$$a_k + \sum_{j \in Q^+} \frac{|b_j| \Delta_{kj}}{\boldsymbol{\delta}_j^\intercal \boldsymbol{p}} p_k = \left(s + \sum_{j \in Q^-} \frac{|b_j| \Delta_{kj}}{\boldsymbol{\delta}_j^\intercal \boldsymbol{p}} \right) p_k. \tag{6}$$

Based on (6), we formulate a fixed-point iteration approach to deriving the stationary point of the likelihood function,

$$p_k^{(t+1)} = \frac{a_k + \sum_{j \in Q^+} \frac{|b_j| \Delta_{kj}}{\boldsymbol{\delta}_j^\intercal \boldsymbol{p}^{(t)}} p_k^{(t)}}{s + \sum_{j \in Q^-} \frac{|b_j| \Delta_{kj}}{\boldsymbol{\delta}_j^\intercal \boldsymbol{p}^{(t)}}}, \ k = 1, \ldots, K. \tag{7}$$

We rearrange (7) into a matrix format as shown in Algorithm 1, which can be viewed as a stable version of the weaver algorithm in Dong & Yin (2018). The weaver algorithm updates the parameter by $\boldsymbol{p} = \boldsymbol{a}/(s\mathbf{1} - \boldsymbol{\Delta}\boldsymbol{\tau})$, and its iteration might collapse due to the existence of negative terms in the denominator $(s\mathbf{1} - \boldsymbol{\Delta}\boldsymbol{\tau})$ and zeros in $\boldsymbol{a}$, leading to nonpositive $\boldsymbol{p}^{(t+1)}$. To overcome this defect, Bayesian weaver with a two-layer iteration structure was proposed, which thickens the complete counts $\boldsymbol{a}$ with Dirichlet priors to enlarge s so that $(s\mathbf{1} - \boldsymbol{\Delta}\boldsymbol{\tau})$ remains positive during updates (Dong & Yin, 2018). However, Bayesian weaver is time-consuming due to the inner–outer iteration structure and the selection of the thickening parameter is difficult. Our stable weaver algorithm does not require the Bayesian outer iterations, and thus uses fewer iterations and less time to reach the MLE with a stable updating path.

Algorithm 1 Stable Weaver

Input: Observations $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Delta})$
Initialize: $\boldsymbol{p}^{(0)} = (1/K, \ldots, 1/K)^\intercal$
repeat
  $\boldsymbol{\tau} = \boldsymbol{b}/\boldsymbol{\Delta}^\intercal \boldsymbol{p}^{(t)}$ (element-wise division)
  $\boldsymbol{\tau}^+ = \max(\boldsymbol{\tau}, \mathbf{0})$, $\boldsymbol{\tau}^- = \min(\boldsymbol{\tau}, \mathbf{0})$
  $\boldsymbol{p}^{(t+1)} = \left[\boldsymbol{a} + (\boldsymbol{\Delta}\boldsymbol{\tau}^+) \circ \boldsymbol{p}^{(t)}\right]/(s\mathbf{1} - \boldsymbol{\Delta}\boldsymbol{\tau}^-)$
  ($\circ$ represents element-wise product)
  $\boldsymbol{p}^{(t+1)} = \boldsymbol{p}^{(t+1)}/\mathrm{sum}(\boldsymbol{p}^{(t+1)})$
until convergence

3.2. Properties of the Stable Weaver Algorithm

Our stable weaver (Algorithm 1) is inspired by the fixed-point iteration method which reaches convergence at the stationary point of the log-likelihood function. In fact, under reasonable assumptions, the stable weaver algorithm has the form of an MM algorithm and thus strictly increases $\ell(\boldsymbol{p}|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Delta})$ in every updating step.

Assumption 1 *For any $k \in \{1, \ldots, K\}$, the observations $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Delta})$ satisfy the following two conditions:*

(i) Either $s > 0$ or $\sum_{j=1}^{q} I\{b_j < 0, \Delta_{kj} > 0\} > 0$;

(ii) Either $a_k > 0$ or $\sum_{j=1}^{q} I\{b_j > 0, \Delta_{kj} > 0\} > 0$.

Assumption 1 requires that each category should appear in both positive count cells and negative count cells for at least once. Note that s is the number of observations with the whole sample space as possible outcomes, and $\boldsymbol{a}$ records the counts corresponding to $|C_i| = 1$.

Theorem 2 *At the t-th iteration, if (i) Assumption 1 is satisfied, (ii) $\boldsymbol{p}^{(t)} > \mathbf{0}$ and (iii) $\exists k$, $\partial \ell(\boldsymbol{p}^{(t)})/\partial p_k \neq 0$, then under the updating rule of the stable weaver algorithm, it holds that*

$$\ell(\boldsymbol{p}^{(t+1)}) > \ell(\boldsymbol{p}^{(t)}).$$

Theorem 2 guarantees a strict increase of the log-likelihood in each iteration, which indicates the effectiveness of our updating rule. If the log-likelihood function (5) contains only one stationary point which satisfies (6), then this stationary point would be the global maximum, i.e., the MLE.

Theorem 3 *Under Assumption 1, $\ell(\boldsymbol{p})$ has a unique stationary point if $Q^- = \varnothing$ or $Q^+ = \varnothing$.*

However, the condition $Q^- = \varnothing$ or $Q^+ = \varnothing$ in Theorem 3 may not be satisfied for incomplete multinomial data if both partial and conditional classifications exist, therefore the stationary point of (5) might not be the MLE.

Following the strategy for proving convergence of the MM algorithm (Hunter, 2004), we can derive the convergence

properties of the stable weaver algorithm under some mild conditions.

Lemma 1 *(Lyapunov's Theorem) (Lange, 1995) Suppose that*

(i) $F : \Theta \mapsto \Theta$ *is continuous;*

(ii) $\ell : \Theta \mapsto \mathbb{R}$ *is differentiable;*

(iii) $\forall \boldsymbol{p} \in \Theta$, $\ell(F(\boldsymbol{p})) \geq \ell(\boldsymbol{p})$, *with equality held if and only if* $\boldsymbol{p}$ *is a stationary point of* $\ell(\cdot)$*;*

then for an arbitrary $\boldsymbol{p}^{(0)} \in \Theta$, *any limit points of the sequence* $\boldsymbol{p}^{(t+1)} = F(\boldsymbol{p}^{(t)})$ *is a stationary point of* $\ell(\boldsymbol{p})$.

The mapping F in Lemma 1 is taken to be one iteration of the stable weaver algorithm, and its continuity is easy to verify. Obviously, the log-likelihood function (5) is differentiable on Θ. The condition $\ell(F(\boldsymbol{p})) \geq \ell(\boldsymbol{p})$ is guaranteed by Theorem 2 when $\boldsymbol{p}$ is not a stationary point of $\ell(\boldsymbol{p})$. The proof of Theorem 2 in the Supplementary Material shows that $\ell(F(\boldsymbol{p})) = \ell(\boldsymbol{p})$ for $\boldsymbol{p} \in \Theta$ if and only if $F(\boldsymbol{p}) = \boldsymbol{p}$, and thus $\boldsymbol{p}$ satisfies (6), i.e., $\boldsymbol{p}$ is a stationary point of $\ell(\cdot)$. It remains to show that the sequence $\boldsymbol{p}^{(t+1)} = F(\boldsymbol{p}^{(t)})$ has limit points for any starting point $\boldsymbol{p}^{(0)} \in \Theta$.

During the updates of the stable weaver algorithm, we can obtain a bounded increasing sequence $\ell(\boldsymbol{p}^{(t)})$, where $\ell(\boldsymbol{p}^{(0)}) \leq \ell(\boldsymbol{p}^{(t)}) \leq \ell(\hat{\boldsymbol{p}})$ and $\hat{\boldsymbol{p}}$ is the MLE of $\ell(\boldsymbol{p})$. The parameter space Θ is not a compact set. However, given any starting point $\boldsymbol{p}^{(0)}$, the updating path of the stable weaver algorithm must be included in a subset of the parameter space $\{\boldsymbol{p} \in \Theta : \ \ell(\boldsymbol{p}) \geq \ell(\boldsymbol{p}^{(0)})\}$. Since a compact set implies the existence of at least one limit point, we need to give sufficient conditions for the compactness of the set $\{\boldsymbol{p} \in \Theta : \ \ell(\boldsymbol{p}) \geq c\}, \forall c \in \mathbb{R}$, i.e., the upper compactness of the log-likelihood function $\ell(\cdot)$.

Assumption 2 *Given the observations* $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Delta})$, *there exists a reformulation of the likelihood function as (4), such that* $\forall k, j \in \{1, \ldots, K\}$, $k \neq j$, $\exists i \in \{1, \ldots, n\}$, *(i)* $\{k\} = C_i$*; and (ii)* $j \in A_i \backslash C_i$.

Assumption 2 has a form similar to those in Hunter (2004) and Huang et al. (2006), indicating that k 'beats' at least one set containing j. Assumption 2 might be too strong while these conditions can be fulfilled by adding an extra term $\gamma \sum_{k=1}^{K} \log p_k$ to (5). Usually a relatively small γ is used to alleviate the induced bias, e.g., we set $\gamma = 10^{-6}$ in all scenarios.

Under Assumption 2, we can claim the upper compactness of $\ell(\boldsymbol{p})$ and derive convergence properties of the stable weaver algorithm.

Theorem 4 *If Assumptions 1 and 2 hold, then* $\ell(\boldsymbol{p})$ *is upper compact, and there exists at least one limit point in the sequence* $\boldsymbol{p}^{(t+1)} = F(\boldsymbol{p}^{(t)})$. *According to Lemma 1, any convergence point of the stable weaver algorithm is a stationary point of (5).*

On the other hand, the sequence of $\ell(\boldsymbol{p}^{(t)})$ is increasing and bounded above. The monotone convergence theorem indicates the convergence of this sequence, which means that our updates have at least one limit point. Further, we can draw the conclusion that the stable weaver algorithm converges to a stationary point of the log-likelihood function (5). If the conditions in Theorem 3 are satisfied, the existence of at most one stationary point is guaranteed, and thus we can obtain the unique stationary point as the MLE.

Our stable weaver algorithm is a simple fixed-point iteration method with only element-wise matrix operations and matrix products. The complexity of one stable weaver iteration is $O(K \times q)$, where $K \times q$ is the dimension of $\boldsymbol{\Delta}$. This renders our algorithm great time efficiency and capability of handling large datasets.

3.3. Applications of Incomplete Multinomial Models

3.3.1. Polytomous Response Data with Underlying Categories

For some problems, we are not interested in the observed, explicit categories, but the underlying, implicit categories. One such example is the phenotype expressions on blood types according to genotypes as shown in Table 1. With the observed counts of the blood types $(n_A,\ n_B,\ n_O,\ n_{AB})$, we can derive the likelihood function,

$$\begin{aligned} L(p_A, p_B, p_O) &= (p_A^2 + 2p_A p_O)^{n_A}(p_B^2 + 2p_B p_O)^{n_B} \\ &\quad \times (2p_A p_B)^{n_{AB}}(p_O^2)^{n_O}, \end{aligned}$$

which can be rearranged into an incomplete multinomial model with

$$\begin{aligned} \boldsymbol{a} &= (n_A + n_{AB}, n_B + n_{AB}, 2n_O)^\intercal, \\ \boldsymbol{b} &= (n_A, n_B)^\intercal, \\ \boldsymbol{\Delta}^\intercal &= \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 2 \end{bmatrix}. \end{aligned}$$

Table 1. Relationship between the blood types and genotypes

Phenotype	Genotype	Probability
A	AA	p_A^2
	AO	$2p_A p_O$
B	BB	p_B^2
	BO	$2p_B p_O$
O	OO	p_O^2
AB	AB	$2p_A p_B$

Similar models are also applicable in recommendation systems. For example, in online video recommendation problems, each individual has a personalized probability vector which reflects his/her interest on different types of videos, and each video often contains multiple labels, e.g., 'travel', 'cooking' and 'sports'. The selection of a video can be viewed as a choice from all available videos on the web page. Assuming video A with labels $(1,3)$, video B with labels $(2,3)$, video C with labels $(1,3,4)$, if video A is selected from (A,B,C), the probability of the selection has the form of a multinomial slice, $(p_1+p_3)/[(p_1+p_3)+(p_2+p_3)+(p_1+p_3+p_4)] = (p_1+p_3)/(2p_1+p_2+3p_3+p_4)$. An incomplete multinomial model can be built upon individual browsing history, and the levels of preference on different video labels can be estimated to improve customized recommendation.

3.3.2. Pairwise Comparison with Home Advantage and Ties

With the generalized settings of $\boldsymbol{\Delta}$, incomplete multinomial models can suit more ranking problems. For example, Agresti (1990) considered the pairwise comparisons with home-field advantage,

$$\Pr(i \succ j) = \begin{cases} \theta p_i/(\theta p_i + p_j), & \text{if } i \text{ is home,} \\ p_i/(p_i + \theta p_j), & \text{if } j \text{ is home,} \end{cases}$$

where θ is the advantage parameter. For the pair (i,j), let $\omega_{ij|h}$ be the count of i wining j if $h \in \{i,j\}$ is home. The probabilities can then be encoded into the expressions of an incomplete multinomial model for which

$$\boldsymbol{b}^\intercal = \begin{matrix} \cdots & t_{ij} & t_{ij}+1 & \cdots \end{matrix} \\ \begin{bmatrix} \cdots, & -\omega_{ij|i}, & -\omega_{ij|j}, & \cdots \end{bmatrix},$$

$$\boldsymbol{\Delta}^\intercal = \begin{matrix} \\ \vdots \\ t_{ij} \\ t_{ij}+1 \\ \vdots \end{matrix} \begin{matrix} \begin{matrix} \cdots & p_i & \cdots & p_j & \cdots \end{matrix} \\ \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \theta & \mathbf{0} & 1 & \mathbf{0} \\ \mathbf{0} & 1 & \mathbf{0} & \theta & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \end{matrix}, \quad (8)$$

where $\mathbf{0}$ represents a row vector of 0s.

Another important extension is to handle ties in pairwise comparisons. Rao & Kupper (1967) established that

$$\begin{aligned} \Pr(i \succ j) &= p_i/(p_i + \theta p_j), \\ \Pr(i \prec j) &= p_j/(\theta p_i + p_j), \\ \Pr(i = j) &= (\theta^2 - 1)p_i p_j/[(p_i + \theta p_j)(p_j + \theta p_i)], \end{aligned}$$

where $\theta > 1$ is the parameter associated with ties. For the above model, we can derive an incomplete multinomial expression for $\boldsymbol{\Delta}$ similar to (8).

3.3.3. Tied Rankings in the Plackett–Luce Model

The Plackett–Luce model formulates the probability of ordering sequences, e.g., $1 \succ 4 \succ 3 \succ 5 \succ 2$. However, partial rankings or ties often appear in multiple ranking data. If rankings are known up to several subsets of items, such as $\{1,4\} \succ \{3,2\} \succ 5$, the likelihood has the form

$$\frac{p_1+p_4}{\sum_{k=1}^5 p_k} \times \frac{p_2+p_3}{p_2+p_3+p_5},$$

which can be encoded into the tuple $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Delta})$.

Ties often appear in sports and gaming (e.g., gymnastics, diving) where the matches are evaluated by referees. If a survey collects people's opinions on the best k items (unordered) among all objects, this also belongs to ranking models with ties, e.g., election voting. A discrete rating system is another type of Plackett–Luce model with ties, which treats objects with the same rating as a subset and generates rankings among the subsets.

4. Experimental Results

In this section, we study large-sample properties and numerical performances of our algorithm on synthetic data and real data. Although our experiments only consider the standard incomplete multinomial model with $\Delta_{kj} \in \{0,1\}$, Section 3 implies that our algorithm is also applicable to data with $\Delta_{kj} \geq 0$. All the methods are coded in the form of matrix operations for fair comparisons on time efficiency.

The convergence criteria of all implemented algorithms are set to be

$$\| \boldsymbol{p}^{(t+1)} - \boldsymbol{p}^{(t)} \|_{L_1} < \epsilon,$$

where ϵ is the tolerance. We set $\epsilon = 10^{-9}$ for the simulation studies in Section 4.1 and $\epsilon = 10^{-6}$ for the remaining experiments.

4.1. Large-Sample Properties

We generate data based on a contingency table whose population is cross-classified into six categories according to gender and age. The contingency table includes one complete sample, two partially classified samples, and one conditionally classified sample with corresponding sample sizes of $\{120, 40, 40, 100\}$. Details of the contingency table are given in the Supplementary Material. In each simulation, four multinomial samples are created with true probabilities $\boldsymbol{p}$, and then they are encoded into incomplete multinomial observations, which means that we have the same $\boldsymbol{\Delta}$ but distinct $\boldsymbol{a}$ and $\boldsymbol{b}$. We run $N = 100000$ simulations to examine the asymptotic properties of our estimator.

The simulation results are reported in Table 2. The MLE of (5) satisfies consistency and asymptotic normality, for which

Table 2. Simulation results under the gender–age contingency table averaged over 100000 simulations

	TRUE $\boldsymbol{p}$	MLE $\hat{p}$	SD	SE
p_1	0.1654	0.1654	0.0233	0.0233
p_2	0.2024	0.2022	0.0338	0.0337
p_3	0.1444	0.1444	0.0293	0.0292
p_4	0.1532	0.1533	0.0222	0.0222
p_5	0.2301	0.2302	0.0348	0.0347
p_6	0.1046	0.1046	0.0264	0.0261

the asymptotic covariance corresponds to the inverse of the observed information matrix. The 'SE' column is computed as the square root of the diagonal of the averaged asymptotic covariance matrix. The parameter estimator of the stable weaver algorithm is close to the true $\boldsymbol{p}$, and the sample standard deviation and standard error match well, which indicates the accuracy of the MLE obtained by the stable weaver algorithm. To further examine asymptotic properties of the stable weaver estimation, we consider the coverage probability of true $\boldsymbol{p}$. We define the squared Mahalanobis distance between the estimator and true $\boldsymbol{p}$,

$$D^{(i)} = (\hat{\boldsymbol{p}}^{(i)}_{[1-5]} - \boldsymbol{p}_{[1-5]})^\intercal {\boldsymbol{S}^{(i)}}^{-1} (\hat{\boldsymbol{p}}^{(i)}_{[1-5]} - \boldsymbol{p}_{[1-5]}),$$

where $\boldsymbol{p}_{[1-5]} = (p_1, \ldots, p_5)^\intercal$ is the vector of the first five elements of $\boldsymbol{p}$ and $\boldsymbol{S}^{(i)}$ is the asymptotic covariance matrix in the i-th simulation. The asymptotic normality of the MLE implies that $D^{(i)}$ follows a $\chi^2(5)$ distribution when sample size is large. We consider the proportion of $D^{(i)}$ that leads to acceptance of the null hypothesis under the significance level q, i.e., $\sum_{i=1}^{N} I\{D^{(i)} \le \chi^2_{1-q}(5)\}/N$, where N is the number of simulations, $\chi^2_{1-q}(5)$ is the $(1-q)$-th quantile of $\chi^2(5)$ distribution.

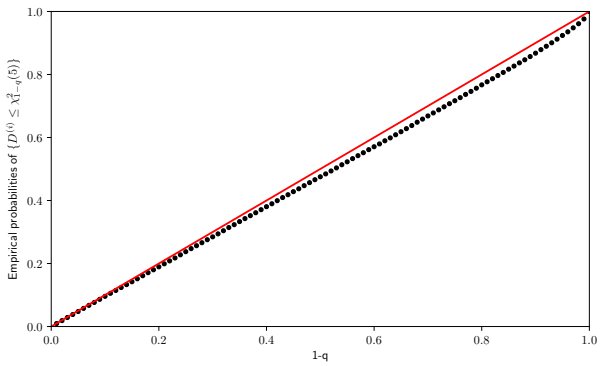

Figure 1. Probability-probability plot of $\chi^2(5)$ and $D^{(i)}$

Figure 1 shows that the empirical distribution of $D^{(i)}$ matches well with the $\chi^2(5)$ distribution. For each p_k, the asymptotic normality of $\hat{p}_k$ allows us to verify marginal coverage probabilities. The empirical marginal coverage probabilities based on simulation results are $(0.944, 0.944, 0.942, 0.945, 0.945, 0.937)$ under the 95% confidence level. The evidence from the joint and marginal coverage probabilities confirms the asymptotic properties of the MLE from the stable weaver estimation.

Table 3. Number of iterations needed for convergence

ITERATION	STABLE WEAVER	WEAVER	BAYESIAN WEAVER	SELF CONSISTENCY
MEAN	20.32	16.25	439.23	47.65
SD	2.72	3.26	42.25	5.51

Table 3 presents the average number of iterations and standard deviation for the convergence of three weaver-type algorithms and the self-consistency approach (Turnbull, 1976). The stable weaver algorithm takes slightly more iterations than the basic weaver due to the simple data structure of the contingency table. While under some more complex simulation settings, e.g., experiments in Section 4.3, or on real datasets, the weaver algorithm often collapses during updates, and our stable weaver algorithm has significant advantages over other methods.

4.2. Estimation with Weak Signal and A Large Number of Categories K

To investigate the convergence performance of our algorithm with incomplete multinomial data, we consider the simulation under weak signal and a large number of categories $K = 20000$, where

$$\begin{aligned} \boldsymbol{p} &= (p_1, \ldots, p_{20000})^\intercal, \\ p_i &= \begin{cases} \frac{1}{200\times 9901}, & i = 1, 101, 201 \ldots, 19901, \\ \frac{100}{200\times 9901}, & \text{otherwise.} \end{cases} \end{aligned} \tag{9}$$

Here all the p_i's are the same except p_{1+100j}, $j = 0, \ldots, 199$, which are one percent of the rest elements. We construct the weak signal likelihood function $L_{\text{ws}}(\boldsymbol{p})$ whose MLE exactly equals to the true $\boldsymbol{p}$ in (9),

$$L_{\text{ws}}(\boldsymbol{p}) = \prod_{j=1}^{200} L_j(\boldsymbol{p}_{[j]}), \tag{10}$$

where $\boldsymbol{p}_{[j]} = \left(p_{1+100(j-1)}, \ldots, p_{100j}\right)^\intercal$ and for example if $j = 1$,

$$\begin{aligned} L_1(\boldsymbol{p}_{[1]}) =& p_1 p_2^{100} \cdots p_{100}^{100} \\ & \times (p_1 + p_2)^{101} (p_3 + p_4)^{200} \cdots (p_{99} + p_{100})^{200} \\ & \times \frac{(p_1 + p_3)^{101} (p_2 + p_4)^{200}}{(p_1 + p_2 + p_3 + p_4)^{301}} \\ & \times \frac{(p_5 + p_7)^{200} (p_6 + p_8)^{200}}{(p_5 + p_6 + p_7 + p_8)^{400}} \\ & \times \cdots \times \frac{(p_{97} + p_{99})^{200} (p_{98} + p_{100})^{200}}{(p_{97} + p_{98} + p_{99} + p_{100})^{400}}. \end{aligned}$$

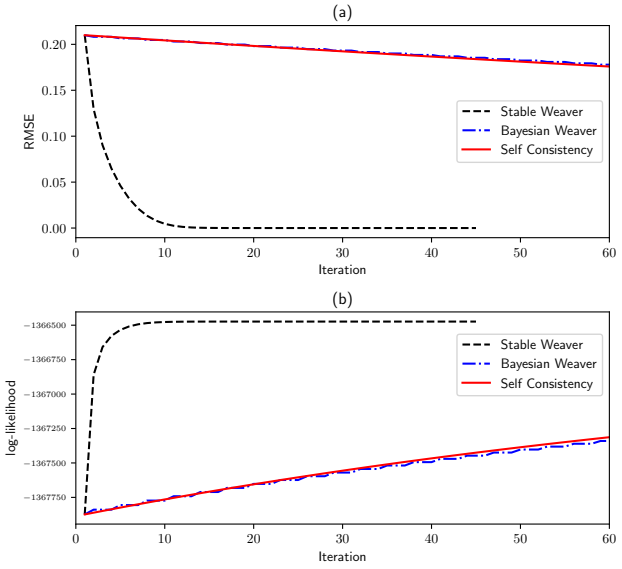

Figure 2. Convergence plots of the stable weaver, Bayesian weaver, self-consistency algorithms based on the likelihood (10), (a) RMSE (b) log-likelihood.

Figure 2 displays the covergence plots of the stable weaver algorithm compared with two existing approaches, the Bayesian weaver and self-consistency algorithms in terms of the log-likelihood and root mean square error (RMSE) as defined in Maystre & Grossglauser (2015). We observe that both RMSEs and log-likelihoods of our stable weaver algorithm reach the plateaus in a few updates (achieving final convergence after 45 iterations). Although both metrics keep improving for the Bayesian weaver and self-consistency methods, their improvement rates are quite low, which eventually achieve convergence after 2229 and 3438 iterations respectively under the same stopping criterion.

4.3. Simulations Under Random Partition

To further evaluate statistical efficiency of the stable weaver algorithm, we consider the random partition procedure for data generation. Each incomplete multinomial expression consists of R multinomial slices. In a multinomial slice, a random partition process (Dong & Yin, 2018) is used to split the sample space into several subspaces, then the samples are drawn from corresponding multinomial distributions. The number of observations in each multinomial slice is controlled by another parameter m. The sample size of a single sample has the order $O(R \times m)$. We take $K = 32$, true $\boldsymbol{p} = (1, \ldots, 32)/528$, and vary R and m. For each pair of (R, m), we replicate 2000 simulations.

The boxplots in Figure 3 show the distribution of the estimated probability of the first class $\hat{p}_1$. The stable weaver

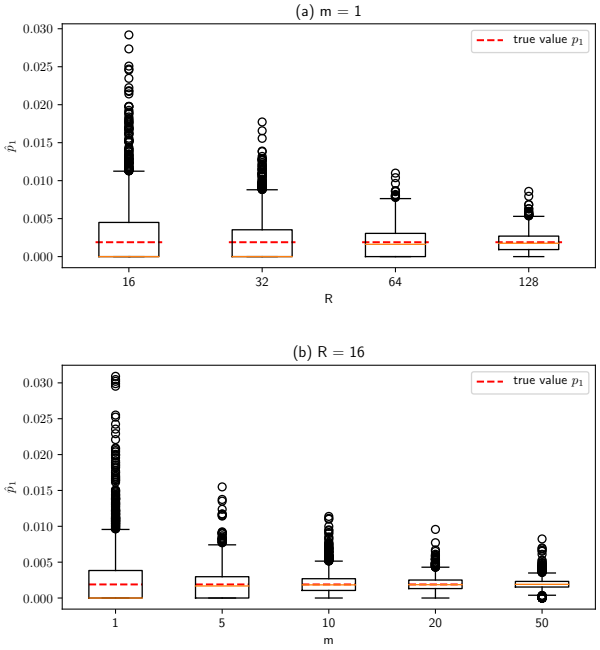

Figure 3. Boxplots of $\hat{p}_1$ (a) with $R \in \{16, 32, 64, 128\}$ and $m = 1$ (b) with $R = 16$ and $m \in \{1, 5, 10, 20, 50\}$

leads to remarkable reduction of the variance when R or m increases. As the true probability of the first class is $p_1 = 1/528$, which is very small so that abnormal estimation might occur with several relatively small pairs of (R, m), for which the sample medians of $\hat{p}_1$ are close to zero. Such abnormality of estimation disappears when more samples are included.

Table 4. Time efficiency (seconds per simulation) under random partition simulation settings.

K	R	STABLE WEAVER	WEAVER	GBTML	SELF CONSISTENCY
32	16	0.0105	1.7804	0.0137	0.0175
	32	0.0077	1.5591	0.0112	0.0116
	64	0.0054	1.5271	0.0095	0.0087
	128	0.0041	1.7566	0.0090	0.0080
64	32	0.0160	3.0583	0.0331	0.0301
	64	0.0124	3.3152	0.0317	0.0253
	128	0.0101	4.0837	0.0318	0.0265
	256	0.0117	6.2097	0.0469	0.0541
128	64	0.1001	11.5110	1.6094	1.2040
	128	0.7439	16.1314	2.2453	1.2659
	256	0.3426	21.0529	1.1494	1.7715
	512	0.5723	155.7229	1.9377	1.9291

GBTML: Generalized Bradley–Terry model using maximum likelihood.

In terms of time efficiency, we compare the stable weaver algorithm with three methods, the Bayesian weaver (Dong & Yin, 2018) (the basic weaver algorithm fails for all simulations in this section), the generalized Bradley–Terry model

using maximum likelihood (GBTML) (Huang et al., 2006) and self-consistency algorithm (Turnbull, 1976). As shown in Table 4, the stable weaver algorithm consistently achieves the best time efficiency under different configurations of K and R, where we set $m = 1$ for all scenarios.

4.4. Experiments on Real Data

We further evaluate the performances of the stable weaver and existing algorithms on three real datasets. The first dataset NASCAR (Hunter, 2004) contains complete ranking results for 43 car races. To investigate performances of the algorithms under partial rankings with ties, we manually aggregate some adjacent rankings in each race and build a partial ranking dataset NASCAR (with ties) upon the aggregated results. The rules of aggregation take NASCAR points scoring systems 1975–2003 as reference and the details are given in the Supplementary Material. The second dataset contains results of 1561 horse races in the Hong Kong Jockey Club (HKJC) from 2014–2016 called HKJC1416, in which 1741 horses participated at least one race. Five existing algorithms including the MM (Hunter, 2004), Bayesian weaver (Dong & Yin, 2018), iterative Luce spectral ranking (ILSR) (Maystre & Grossglauser, 2015), self-consistency (Turnbull, 1976) and trust region constrained algorithm (Byrd et al., 1999) are compared with the stable weaver. As the MM and ILSR cannot deal with multiple rankings with ties, we consider a smaller dataset by removing the races with ties.

Table 5. Comparison of six algorithms on real datasets.

Algorithm		NASCAR (w/o ties)	NASCAR (w/ ties)	HKJC1416 (w/o ties)	HKJC1416 (w/ ties)
Stable Weaver	Iteration	22	459	40.4K	27.2K
	Time (s)	<0.01	0.03	38.46	86.40
Bayesian Weaver	Iteration	128K	263K	>1M	>1M
	Time (s)	25.27	50.12	>5000	>5000
MM	Iteration	22	–	40.4K	–
	Time (s)	<0.01	–	375.79	–
Trust Region*	Iteration	1937	5048	636†	649†
	Time (s)	74.31	125.68	1139.14	1835.37
ILSR	Iteration	12	–	4056	–
	Time (s)	0.06	–	1166.97	–
Self Consistency	Iteration	36798	11282	–‡	–
	Time (s)	11.61	2.08	–	–

* The number of iterations for the trust region constrained algorithm refers to the number of the objective function evaluations.
† We use the approximated Hessian matrix when fitting the trust region constrained algorithm to the HKJC1416 data because its calculation is too time-consuming.
‡ For the HKJC1416 data, the self-consistency approach converges to a wrong solution.

As shown in Table 5, in terms of running time, the stable weaver algorithm outperforms all the five existing methods significantly. For the complete ranking data, although ILSR requires the smallest number of iterations, formulation of the $K \times K$ transition matrix and calculation of the left leading eigenvector take too much time in each iteration, resulting in worse performance on the total execution time.

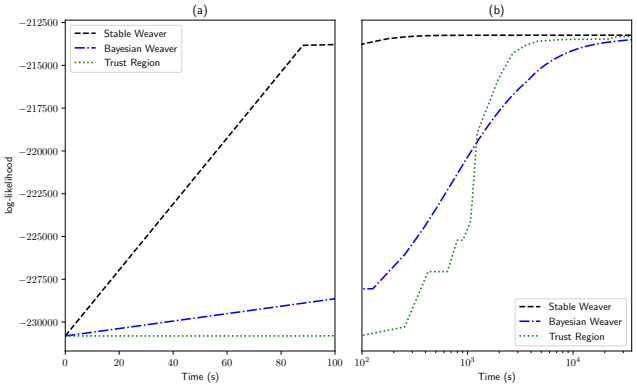

Figure 4. Convergence plot of the stable weaver algorithm compared with existing methods on the dataset HKJC9916 against running time (a) $t \in [0, 100]$ and (b) $t \in [100, 36000]$ (s).

Considering a larger dataset, we collect the HKJC horse racing results from 1999 to 2016 (HKJC9916), which includes 13223 races and 7878 horses. Figure 4 presents the log-likelihood of the stable weaver, Bayesian weaver and trust region constrained approaches against execution time. It is shown that the log-likelihood of the stable weaver arrives at the plateau in 100 seconds, while those of the Bayesian weaver and trust region constrained algorithms climb at much lower rates. The performance of the stable weaver remains the best on a larger real dataset, while all other methods could not achieve convergence even after ten hours.

5. Conclusion

This paper considers the problem of statistical estimation and inference for incomplete multinomial models. A fixed-point iteration approach is proposed for obtaining the MLEs of the model parameters. Under some mild assumptions, we prove the convergence of our stable weaver algorithm to a stationary point of the likelihood function. The extended framework of modeling incomplete multinomial data demonstrates its great capabilities of generalization, which unifies several well-known models for categorical response problems. Simulation studies show that the estimator from the stable weaver has the properties of the MLE, and experiments on real data also illustrate the advantages of our algorithm in terms of computational efficiency compared with existing methods. The simple updating rule in our algorithm only involves basic matrix operations, which makes it run fast and easy to implement with large data.

Acknowledgements

We thank the four anonymous reviewers for insightful suggestions that have significantly improved the paper. This research is supported by the Research Grants Council of Hong Kong (17326316) and TCL Corporate Research (Hong Kong).

References

Agarwal, A., Patil, P., and Agarwal, S. Accelerated spectral ranking. In *International Conference on Machine Learning*, pp. 70–79, 2018.

Agresti, A. *Categorical Data Analysis*. John Wiley & Sons, 1990.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Byrd, R. H., Hribar, M. E., and Nocedal, J. An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4):877–900, 1999.

Chen, T. and Fienberg, S. E. The analysis of contingency tables with incompletely classified data. *Biometrics*, 32 (1):133–144, 1976.

Dickey, J. M., Jiang, J. M., and Kadane, J. B. Bayesian methods for censored categorical data. *Journal of the American Statistical Association*, 82(399):773–781, 1987.

Dong, F. and Yin, G. Maximum likelihood estimation for incomplete multinomial data via the weaver algorithm. *Statistics and Computing*, 28:1095–1117, 2018.

Hocking, R. R. and Oxspring, H. Maximum likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, 66(333):65–70, 1971.

Huang, T. K., Weng, R. C., and Lin, C. J. Generalized Bradley–Terry models and multi-class probability estimates. *Journal of Machine Learning Research*, 7(Jan): 85–115, 2006.

Hunter, D. R. MM algorithms for generalized Bradley–Terry models. *The Annals of Statistics*, 32(1):384–406, 2004.

Ireland, C. T. and Kullback, S. Contingency tables with given marginals. *Biometrika*, 55(1):179–188, 1968.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Lange, K. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 57(2):425–437, 1995.

Luce, R. D. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons, 1959.

Maystre, L. and Grossglauser, M. Fast and accurate inference of Plackett–Luce models. In *Advances in Neural Information Processing Systems*, pp. 172–180, 2015.

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

Paulino, C. D. M. and de Braganca Pereira, C. A. Bayesian methods for categorical data under informative general censoring. *Biometrika*, 82(2):439–446, 1995.

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

Rao, P. and Kupper, L. L. Ties in paired-comparison experiments: A generalization of the Bradley–Terry model. *Journal of the American Statistical Association*, 62(317): 194–204, 1967.

Schafer, J. L. *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC, 1997.

Soufiani, H. A., Chen, W., Parkes, D. C., and Xia, L. Generalized method-of-moments for rank aggregation. In *Advances in Neural Information Processing Systems*, pp. 2706–2714, 2013.

Turnbull, B. W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B*, 38(3):290–295, 1976.