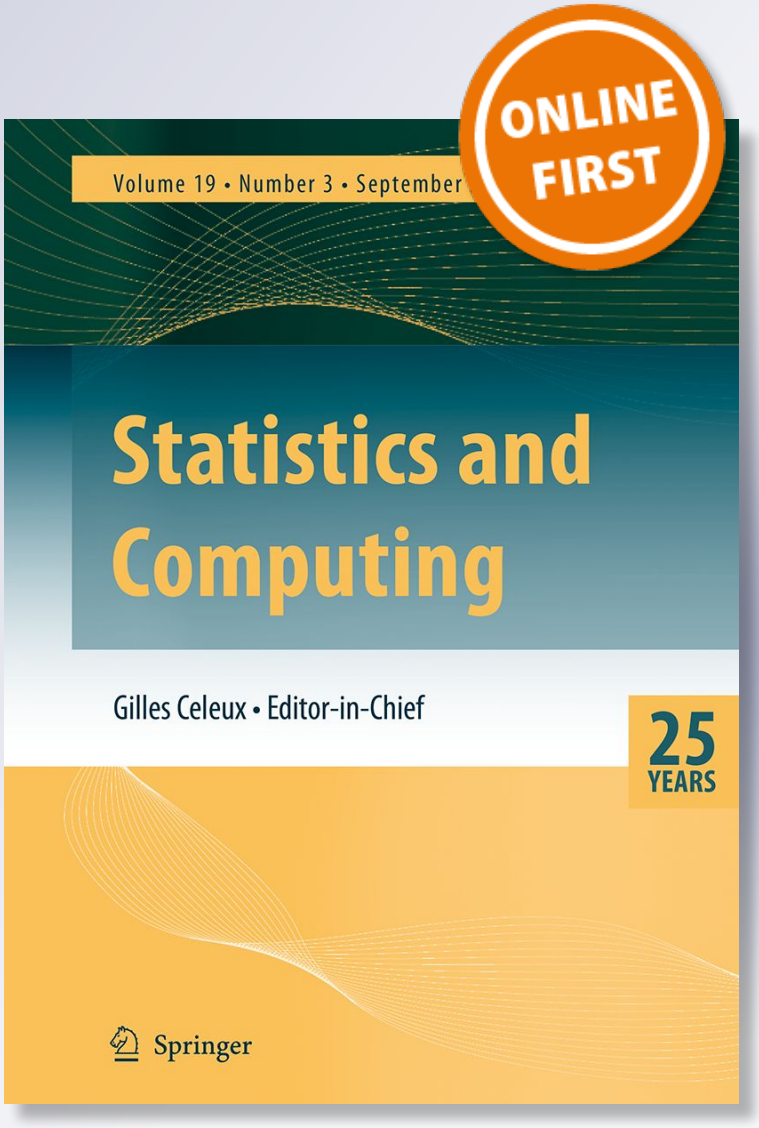
Maximum likelihood estimation for incomplete multinomial data via the weaver algorithm

Fanghu Dong & Guosheng Yin

Statistics and Computing

ISSN 0960-3174

Stat Comput
DOI 10.1007/s11222-017-9782-2



Springer

Stat Comput
DOI 10.1007/s11222-017-9782-2

Maximum likelihood estimation for incomplete multinomial data via the weaver algorithm

Fanghu Dong[1] · **Guosheng Yin**[1]

Received: 28 March 2017 / Accepted: 14 October 2017

Abstract In a multinomial model, the sample space is partitioned into a disjoint union of cells. The partition is usually immutable during sampling of the cell counts. In this paper, we extend the multinomial model to the incomplete multinomial model by relaxing the constant partition assumption to allow the cells to be variable and the counts collected from non-disjoint cells to be modeled in an integrated manner for inference on the common underlying probability. The incomplete multinomial likelihood is parameterized by the complete-cell probabilities from the most refined partition. Its sufficient statistics include the variable-cell formation observed as an indicator matrix and all cell counts. With externally imposed structures on the cell formation process, it reduces to special models including the Bradley–Terry model, the Plackett–Luce model, etc. Since the conventional method, which solves for the zeros of the score functions, is unfruitful, we develop a new approach to establishing a simpler set of estimating equations to obtain the maximum likelihood estimate (MLE), which seeks the simultaneous maximization of all multiplicative components of the likelihood by fitting each component into an inequality. As a consequence, our estimation amounts to solving a system of the equality attainment conditions to the inequalities. The resultant MLE equations are simple and immediately invite a fixed-point iteration algorithm for solution, which is referred to as the weaver algorithm. The weaver algorithm is short and amenable to parallel implementation. We also derive the asymptotic covariance of the MLE, verify main results with simulations, and compare the weaver algorithm with an MM/EM algorithm based on fitting a Plackett–Luce model to a benchmark data set.

Keywords Bradley–Terry model · Contingency table · Count data · Density estimation · Incomplete multinomial model · Plackett–Luce model · Random partition · Ranking · Weaver algorithm

Mathematics Subject Classification 62F07 · 65K10 · 62G07

1 Introduction

In this paper, we extend the multinomial model to allow the cells to be variable. In a multinomial model, the likelihood is $L(\boldsymbol{p}|\boldsymbol{a}) \propto \prod_{i=1}^{d} p_i^{a_i}$, where $\boldsymbol{p} = (p_1, \ldots, p_d)^\intercal$ are called cell probabilities and $\boldsymbol{a} = (a_1, \ldots, a_d)^\intercal$ are the corresponding counts. The sample space is the disjoint union of the cells, and we call the collection of the cells a *partition* of the sample space. Such partition cannot be changed during multinomial sampling of the counts $\boldsymbol{a}$. The relaxation of the constant partition assumption motivates the *incomplete multinomial model*, under which an implicit collection of partitions provide the observed cells which are all measured by a common probability. As a result, the cells need not be disjoint. The cell probabilities of the most refined partition, formed as the intersection of all the observable partitions, constitute the parameters of the relaxed model, held in the vector $\boldsymbol{p}$. Any cell probability $\tilde{p} = \boldsymbol{\delta}^\intercal \boldsymbol{p}$ is expressed as the sum of some elements of $\boldsymbol{p}$, via inner product with a subset indicator vector $\boldsymbol{\delta}$ consisted of 0s and 1s, which is responsible for encoding the formation of a particular variable cell and is part of the observed information. The form $\boldsymbol{\delta}^\intercal \boldsymbol{p}$, the term

✉ Fanghu Dong
jdong@hku.hk

Guosheng Yin
gyin@hku.hk

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong

Published online: 27 October 2017

probability sub-sum, and the term variable-cell probability are used interchangeably throughout the paper.

The likelihood function under the incomplete multinomial model takes the form

$$L(\boldsymbol{p}|\boldsymbol{a},\boldsymbol{b},\boldsymbol{\Delta}) \propto \prod_{i=1}^{d} p_i^{a_i} \prod_{j=1}^{q} \tilde{p}_j^{b_j} = \prod_{i=1}^{d} p_i^{a_i} \prod_{j=1}^{q} (\boldsymbol{\delta}_j^{\mathsf{T}}\boldsymbol{p})^{b_j} \quad (1)$$

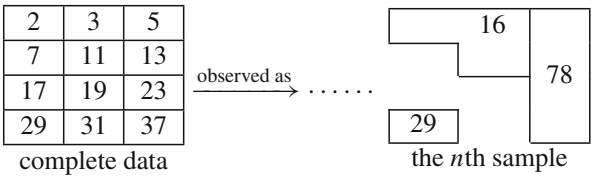
where

1. $\boldsymbol{p} = (p_1, \ldots, p_d)^{\mathsf{T}}$ holds the parameters of the model and represents the probabilities of the most refined cells.
2. $\boldsymbol{a} = (a_1, \ldots, a_d)^{\mathsf{T}}$ collects the usual multinomial counts from the most refined cells; $\boldsymbol{b} = (b_1, \ldots, b_q)^{\mathsf{T}}$ collects the variable cell counts.
3. $\boldsymbol{\delta}_j$, $j = 1, \ldots, q$, is an indicator vector containing only 0s and 1s, indicating the variable cell associated with the count b_j.
4. The observed data of this model are $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_q]$.

The expression and similar forms have appeared in Hankin (2010), Ng et al. (2011, Chapter 8), and Huang et al. (2006). Negative counts are permitted in the vector $\boldsymbol{b} = (b_1, \ldots, b_q)^{\mathsf{T}}$, which cause the corresponding probability sub-sums to appear in the denominator of the likelihood function and that enables a general way to express conditional probabilities (Loève 1977, 1978).

The variable partition property makes (1) a flexible model for probability estimation. With externally imposed structure on $\boldsymbol{\Delta}$, the incomplete multinomial model reduces to specialized models. We show in the following examples that it unifies the Bradley–Terry type multiple comparison models, the Plackett–Luce analysis of permutation models, certain contingency table models, and the general counting experiment on a random partition.

Example 1 In an experiment involving a random partition process, each multinomial sample is collected on a partition instance.



Suppose by the end of the first $n-1$ samplings, we are able to form a prior about the complete-cell probabilities expressed in a Dirichlet distribution. With the newly observed nth sample, which contains a truncation of the sample space, we can use the following incomplete multinomial likelihood to update the estimate,

$$\begin{aligned} L_{\text{obs}}(\boldsymbol{p}) = {} & p_{11}^{\alpha_{11}} p_{12}^{\alpha_{12}} p_{13}^{\alpha_{13}} p_{21}^{\alpha_{21}} p_{22}^{\alpha_{22}} p_{23}^{\alpha_{23}} p_{31}^{\alpha_{31}} p_{32}^{\alpha_{32}} p_{33}^{\alpha_{33}} p_{41}^{\alpha_{41}} \\ & \times p_{42}^{\alpha_{42}} p_{43}^{\alpha_{43}} (p_{11}+p_{12}+p_{22})^{16} \times (p_{13} \\ & + p_{23}+p_{33}+p_{43})^{78} \times p_{41}^{29} \times (p_{11}+p_{12} \\ & + p_{13}+p_{22}+p_{23}+p_{33}+p_{41}+p_{43})^{-123}. \end{aligned}$$

The likelihood is encoded into $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_3]$ as the following:

$$\begin{aligned} \boldsymbol{a}^{\mathsf{T}} &= \begin{matrix} p_{11} & p_{12} & p_{13} & p_{21} & p_{22} & p_{23} & p_{31} & p_{32} & p_{33} & p_{41} & p_{42} & p_{43} \end{matrix} \\ &= \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{41}+29 & \alpha_{42} & \alpha_{43} \end{bmatrix}, \\ \boldsymbol{b}^{\mathsf{T}} &= \begin{matrix} 1 & 2 & 3 \end{matrix} \\ &= \begin{bmatrix} 16, & 78, & -123 \end{bmatrix}, \end{aligned}$$

$$\boldsymbol{\Delta}^{\mathsf{T}} = \begin{matrix} j \\ 1 \\ 2 \\ 3 \end{matrix} \begin{matrix} p_{11} & p_{12} & p_{13} & p_{21} & p_{22} & p_{23} & p_{31} & p_{32} & p_{33} & p_{41} & p_{42} & p_{43} \end{matrix} \\ \begin{bmatrix} 1 & 1 & & & 1 & & & & & & & \\ & & 1 & & & 1 & & & 1 & & & 1 \\ 1 & 1 & 1 & & 1 & 1 & & & 1 & 1 & & 1 \end{bmatrix}.$$

The prior information and the new singleton count, 29, observed in the new sample are held in $\boldsymbol{a}^{\mathsf{T}}$. Each row in $\boldsymbol{\Delta}^{\mathsf{T}}$ indicates a sub-sum of probabilities and the corresponding exponent is collected in $\boldsymbol{b}^{\mathsf{T}}$ at the same position; for example, the first row of $\boldsymbol{\Delta}^{\mathsf{T}}$ indicates the first sub-sum $(p_{11}+p_{12}+p_{22})$ in the likelihood and $b_1 = 16$ is its exponent. The zeros have been omitted from the display. The observed truncation is encoded by the union of all untruncated cells attached with a count equal to the negative sum of all counts unionized. There are two levels of variability in this experiment: one in the multinomial counts given a partition instance; the other in the generation of the sequence of partitions. The randomness of the indicator distribution on the $\boldsymbol{\Delta}^{\mathsf{T}}$ matrix is a reflection of the underlying random partition process. In the next examples, we show more structured patterns for $\boldsymbol{\Delta}^{\mathsf{T}}$.

Example 2 This example introduces further use of negative counts based on contingency table data. Suppose that a population is classified according to the gender and age combination to understand its distribution over these two demographic factors. We are interested in the six cell probabilities.

	Young	Middle	Senior
Female	p_1	p_2	p_3
Male	p_4	p_5	p_6

Four samples of data are collected; three are incomplete to some degree.

Sample 1 of size 120 completely categorizes all six cells.

	Young	Middle	Senior
Female	21	24	18
Male	20	25	12

Sample 2 of size 40 contains only gender information, missing age.

Female	18
Male	22

Sample 3 of size 40 contains only age information, missing gender.

Young	Middle	Senior
10	20	10

Sample 4 of size 100 is collected in a primary school environment, where only the young age group is relevant.

	Young
Female	53
Male	47

The combined information can be modeled by the likelihood:

$$\begin{aligned} L(\boldsymbol{p}) = & p_1^{21} p_2^{24} p_3^{18} p_4^{20} p_5^{25} p_6^{12} \\ & \times (p_1+p_2+p_3)^{18} (p_4+p_5+p_6)^{22} \\ & \times (p_1+p_4)^{10} (p_2+p_5)^{20} (p_3+p_6)^{10} \\ & \times \left(\frac{p_1}{p_1+p_4}\right)^{53} \left(\frac{p_4}{p_1+p_4}\right)^{47}. \end{aligned} \tag{2}$$

The likelihood is encoded into $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_5]$ as the following:

$$\begin{aligned} \boldsymbol{a}^\intercal &= \overset{p_1}{[21+53}, \overset{p_2}{24}, \overset{p_3}{18}, \overset{p_4}{20+47}, \overset{p_5}{25}, \overset{p_6}{12}], \\ \boldsymbol{b}^\intercal &= \overset{1}{[18}, \overset{2}{22}, \overset{3}{10-53-47}, \overset{4}{20}, \overset{5}{10}], \end{aligned}$$

$$\boldsymbol{\Delta}^\intercal = \begin{array}{c} j \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & & & \\ & & & 1 & 1 & 1 \\ 1 & & & 1 & & \\ & 1 & & & 1 & \\ & & 1 & & & 1 \end{bmatrix} \end{array}.$$

The indicator distribution on the $\boldsymbol{\Delta}^\intercal$ matrix is less random than before. The pattern reflects once again the underlying cell formation process.

Example 3 The binomial conditional counts modeling is known as the Bradley–Terry Model (1952), which deals with the problem of ranking from pairwise preference scores. In this model, the relative strength of the ith subject is represented by the ith parameter p_i. The strength parameters are positive and required to sum to one. Hence, they are often treated as probabilities. The data are generated by a judge who examines a pair $(i,\ j)$ and assigns a corresponding pair of scores $(n_{ij},\ n_{ji})$ measuring the relative strengths of the pair. To link the parameters to the judge's scores, the following joint binomial likelihood is proposed:

$$L_{\mathrm{BT}}(\boldsymbol{p}) \propto \prod_{1 \le i < j \le t} \left(\frac{p_i}{p_i+p_j}\right)^{n_{ij}} \left(\frac{p_j}{p_i+p_j}\right)^{n_{ji}}.$$

A character of the Bradley–Terry model lies in that the variable cell counts vector $\boldsymbol{b}$ is consisted of all negative numbers. For $t = 5$, there are 10 pairwise comparisons, the likelihood is encoded into $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{10}]$ as the following:

$$\boldsymbol{a}^\intercal = \left[\overset{p_1}{\sum_{j:j\neq 1}} n_{1j}, \overset{p_2}{\sum_{j:j\neq 2}} n_{2j}, \overset{p_3}{\sum_{j:j\neq 3}} n_{3j}, \overset{5}{\sum_{j:j\neq 4}} n_{4j}, \overset{p_5}{\sum_{j:j\neq 5}} n_{5j} \right],$$

$$\boldsymbol{b}^\intercal = -\left[\overset{1}{n_{12}}, \overset{2}{n_{13}}, \overset{3}{n_{14}}, \overset{4}{n_{15}}, \overset{5}{n_{23}}, \overset{6}{n_{24}}, \overset{7}{n_{25}}, \overset{8}{n_{34}}, \overset{9}{n_{35}}, \overset{10}{n_{45}} \right],$$

$$\boldsymbol{\Delta}^\intercal = \begin{array}{c} j \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{ccccc} p_1 & p_2 & p_3 & p_4 & p_5 \end{array} \\ \begin{bmatrix} 1 & 1 & & & \\ 1 & & 1 & & \\ 1 & & & 1 & \\ 1 & & & & 1 \\ & 1 & 1 & & \\ & 1 & & 1 & \\ & 1 & & & 1 \\ & & 1 & 1 & \\ & & 1 & & 1 \\ & & & 1 & 1 \end{bmatrix} \end{array}.$$

Pairwise comparison data can be represented by a weighted directed graph with the t subjects as vertices. If $n_{ij} > 0$, then an edge emits from vertex i towards vertex j. If this weighted directed graph is strongly connected, then the maximum likelihood estimate (MLE) exists and is unique (Ford 1957). Hastie and Tibshirani (1998) developed a classifier using a formally equivalent likelihood and used its MLE to classify a new feature vector.

Example 4 A more sophisticated type of conditional scores known to arise from the neutral sampling process underpins models such as the Plackett–Luce model for analysis of permutations. Connor and Mosimann (1969) defined *neutrality* to characterize a type of independence arising from the sequential sampling of compositional data, in which truncation of a data segment does not alter the probability distribution on the remaining data. They proposed a likelihood for modeling the probability mass function (PMF) in such a sampling process,

$$L_{\mathrm{CM}}(\boldsymbol{p}) \propto p_n^{\beta_{n-1}-1} \prod_{i=1}^{n-1} \left\{ p_i^{\alpha_i - 1} \left(\sum_{j=i}^{n} p_j \right)^{\beta_{i-1}-(\alpha_i+\beta_i)} \right\},$$

where $\alpha_i,\ \beta_i > 1$ and $n \geq 2$. For $n = 5$, the likelihood is encoded into $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_3]$ as the following:

$$\boldsymbol{a}^{\mathsf{T}} = \begin{matrix} p_1 & p_2 & p_3 & p_4 & p_5 \end{matrix}$$
$$\left[\alpha_1 - 1,\ \alpha_2 - 1,\ \alpha_3 - 1,\ \alpha_4 - 1,\ \beta_4 - 1\right],$$

$$\boldsymbol{b}^{\mathsf{T}} = \begin{matrix} 1 & 2 & 3 \end{matrix}$$
$$\left[\beta_1 - \alpha_2 - \beta_2,\ \beta_2 - \alpha_3 - \beta_3,\ \beta_3 - \alpha_4 - \beta_4\right],$$

$$\boldsymbol{\Delta}^{\mathsf{T}} = \begin{matrix} j \\ 1 \\ 2 \\ 3 \end{matrix} \begin{matrix} p_1 & p_2 & p_3 & p_4 & p_5 \end{matrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 \\ & & 1 & 1 \end{bmatrix}.$$

Plackett (1975) considered the question of how to calculate the probability of a horse ending up in any place after a race. He started by considering the first-place probability of each horse and recursively calculated this probability on a shrinking set of horses by removing the first-place horse in the current set, until none was left. This process fits the definition of neutral sampling. The following joint likelihood is specified for the Plackett–Luce model,

$$L_{\mathrm{PL}}(\boldsymbol{p}) \propto \prod_{r=1}^{R} \prod_{i=1}^{t_r} \frac{p_{i_r}}{\sum_{j=i}^{t_r} p_{j_r}},$$

where t_r is the number of horses raced in round r, in a total of R rounds; i_r is the index of the ith ranking horse in round r.

Hunter (2004) applied this model to a car racing data with $R = 36$, $i_r = 1, \ldots, 83$, and $t_r = 43$ for all r values except in five exceptional rounds where $t_r = 42$. As a result, there are total $43 \times 36 - 5 = 1543$ terms in the full expansion of the likelihood; the first and last three of them are partially produced in the 5th column of Table 1 showing the general shape of the terms. The first four columns of the table display the number of participants for the current round, the ID of the driver, the current round number, and the final place of the driver in the current round. The encoding of $\boldsymbol{a}$ and $\boldsymbol{\Delta}$ can be programmatically produced. The vector $\boldsymbol{b}$ consists of -1 at all positions, a shortcut for the program.

A character of the Plackett–Luce likelihood is that each term's numerator contributes one count to an element of $\boldsymbol{a}$ and the same term's denominator contributes a variable cell encoded by a column indicator vector $\boldsymbol{\delta}_j$ to $\boldsymbol{\Delta}$, together with one negative count to the corresponding element of $\boldsymbol{b}$.

Table 1 Car racing data by the Plackett–Luce model

nDrv t_r	ID i_r	Rnd r	Plc i	Likelihood Component $p_{i_r} / \sum_{j=i}^{t_r} p_{j_r}$
43	83	1	1	$p_{83}/(p_{83} + p_{18} + p_{20} + \cdots)$
43	18	1	2	$p_{18}/(p_{18} + p_{20} + \cdots)$
43	20	1	3	$p_{20}/(p_{20} + \cdots)$
⋮	⋮	⋮	⋮	⋮
43	53	36	41	$p_{53}/(p_{53} + p_{38} + p_{14})$
43	38	36	42	$p_{38}/(p_{38} + p_{14})$
43	14	36	43	$p_{14}/p_{14} = 1$

Note: nDrv stands for the number of drivers in each round; ID is the driver ID; Rnd stands for round; Plc stands for place

1.1 Motivation, contribution, and structure of this paper

A challenge that follows the introduction of the unified likelihood (1) is its estimation. The theoretical side of the difficulty is that the MLE equations produced by the conventional method of setting the score function to zero (with a Lagrange multiplier term for the normalization constraint, which introduces the multiplier itself as an additional unknown) appear neither analytically nor computationally tractable. On the computational side, existing methods usually fit to a specific sub-class of the model and the burden is on the user to understand and implement the algorithm for each problem. A naive construction of an EM algorithm by splitting the probability sub-sums would fail for the Bradley–Terry, Plackett–Luce, and other models that have mostly *negative* counts associated with the variable cells. This leads to the invention of an MM algorithm by Hunter (2004) for both Bradley–Terry and Plackett–Luce MLEs, which, however, is only applicable to the likelihoods with mostly negative variable cell counts, antithetic to the EM case. Thus, a single, simple algorithm for maximizing the most general likelihood (1) is certainly desirable. In this paper, we derive a set of simple maximum likelihood estimating equations for the unified likelihood (1) as a whole and describe the weaver algorithms for solving those equations.

The rest of the paper is organized as follows. Section2 uses an inequality technique to establish the maximum likelihood estimating equations and also derives the asymptotic covariance formula for the MLE. Section 3 describes a fixed-point iteration called the weaver algorithm to solve the equations. Section 4 presents four simulation studies to evaluate the algorithms and the covariance formula as well as to demonstrate usage and estimation of the model. Section 5 compares the weaver algorithm with the MM and EM algorithms and mentions some theoretical considerations.

1.2 A note on the literature

Numerous works have been done on contingency table and incomplete categorical data; a small sample are Agresti (2003), Chen and Fienberg (1976), Dickey et al. (1987), Haberman (1977), Hartley and Hocking (1971), Turnbull (1976), and Laird (1978). One of the earliest publications related to the Bradley–Terry model for a ranking problem is Zermelo (1929). Davidson and Farquhar (1976) compiles a long list of works on the related paired comparison models. Some more recent works on Bradley–Terry model and ranking methods in general include David (1988), Dwork et al. (2001), Jech (1983), Marden (1996), Hastie and Tibshirani (1998), and Yan et al. (2012). The Plackett–Luce model is considered both as a generalization of the Bradley–Terry model and an instantiation of the choice modeling principle (Luce 1959). A few works in this space are Thurstone (1927), Suppes et al. (1971), Tversky (1972), Sattath and Tversky (1976), Luce (1977), Gormley and Murphy (2008).

2 Estimation of the incomplete multinomial model

In this section, we establish the following MLE equations for the incomplete multinomial model:

$$\begin{cases} \tau_j \left(\boldsymbol{\delta}_j^\intercal \boldsymbol{p} \right) = b_j \\ p_i \left(s - \boldsymbol{\delta}_{(i)}^\intercal \boldsymbol{\tau} \right) = a_i \\ \sum_{i=1}^{d} p_i = 1 \\ s = \sum_{i=1}^{d} a_i + \sum_{j=1}^{q} b_j \end{cases} \tag{3}$$

where $\boldsymbol{p}$ is the unknown parameter vector to be solved; $\boldsymbol{\delta}_{(i)}^\intercal$ and $\boldsymbol{\delta}_j$ are the ith row and the jth column of $\boldsymbol{\Delta}$, respectively; $\boldsymbol{\tau}$ is an auxiliary column vector of length q. The first two equations in (3) are repeated for $j = 1, \ldots, q$ and for $i = 1, \ldots, d$, respectively. The third equation is the normalization constraint on the probabilities. The last equation defines the constant s given the observed values of $\boldsymbol{a}$ and $\boldsymbol{b}$, which is used in the second equation.

2.1 Derivation of the MLE equations

The MLE equations in (3) are mainly established on two bases: the simultaneous attainment of equalities in many inequalities all adopting the form of Lemma 1 and the invariance of total counts as expressed in Lemma 2. We state Lemma 1, which may be called the multinomial-MLE inequality, as a standalone result to begin with. Its proof is given in "Appendix A."

Lemma 1 *For $x_1, \ldots, x_n > 0$ and $a_1, \ldots, a_n > 0$,*

$$\prod_{i=1}^{n} x_i^{a_i} \leqslant \frac{\prod_{i=1}^{n} a_i^{a_i}}{\left(\sum_{i=1}^{n} a_i \right)^{\sum_{i=1}^{n} a_i}} \left(\sum_{i=1}^{n} x_i \right)^{\sum_{i=1}^{n} a_i}, \tag{4}$$

where the equality is attained if and only if

$$a_i / x_i = \tau > 0, \tag{5}$$

where τ is a positive ratio same for all i.

For example, the incomplete trinomial likelihood $(p_1 + p_2)^4 p_3$ has the maximum value $4^4/5^5$:

$$\begin{aligned} (p_1 + p_2)^4 p_3 &= 4^4 \left(\frac{p_1 + p_2}{4} \right)^4 p_3 \\ &\leqslant 4^4 \left(\frac{4 \frac{p_1 + p_2}{4} + p_3}{5} \right)^5 \\ &= \frac{4^4 1^1}{5^5} (p_1 + p_2 + p_3)^5 , \end{aligned}$$

where the left-hand side is globally maximized if and only if

$$\frac{4}{p_1 + p_2} = \frac{1}{p_3}.$$

Based on the last equation alone, we can solve for $p_3 = 0.2$ and $p_1 + p_2 = 0.8$. The idea that follows the above example is to pad each sub-sum in the unified likelihood (1) such that it becomes a product of many simple multinomial likelihoods but all sharing a common MLE. One can multiply each sub-sum term $(\boldsymbol{\delta}_j^\intercal \boldsymbol{p})^{b_j}$ by complementing powers on singleton cells inversely indicated by the zeros of $\boldsymbol{\delta}_j$,

$$(\boldsymbol{\delta}_j^\intercal \boldsymbol{p})^{b_j} \mapsto (\boldsymbol{\delta}_j^\intercal \boldsymbol{p})^{b_j} \times \prod_{i:\, \delta_{ji} = 0} p_i^{c_{ji}},$$

where $c_{ji} = \tau_j p_i$ with

$$\frac{b_j}{\boldsymbol{\delta}_j^\intercal \boldsymbol{p}} = \frac{c_{ji}}{p_i} = \tau_j, \text{ for all } j \text{ such that } \delta_{ji} = 0, \tag{6}$$

according to the equality attainment condition (5) for each sub-sum term.

The complete data terms that have been multiplied need again be collectively divided to the likelihood to keep it unchanged. This amounts to a process of sorting c_{ji} into the multinomial cell counts c_i' which is then subtracted from elements of vector $\boldsymbol{a}$ to create the last multinomial component in the product:

$$\frac{\prod_{i=1}^{d} p_i^{a_i}}{\prod_{j=1}^{q}\prod_{i:\delta_{ji}=0} p_i^{c_{ji}}} = \prod_{i=1}^{d} p_i^{a_i - c_i'}.$$

Hence,

$$\sum_{i=1}^{d} c_i' = \sum_{j=1}^{q}\sum_{i:\delta_{ji}=0} c_{ji} = \sum_{j=1}^{q}\sum_{i:\delta_{ji}=0} p_i \tau_j.$$

For fixed i, we have

$$c_i' = p_i \sum_{\substack{j=1 \\ j:\delta_{ji}=0}}^{q} \tau_j. \tag{7}$$

This last multinomial component in the padded likelihood must share the same MLE with the others, which requires

$$\frac{a_i - c_i'}{p_i} = \tau_0 > 0, \text{ for all } i. \tag{8}$$

Combining (7) and (8) with rearrangement, we have

$$a_i = p_i \left(\tau_0 + \left(\mathbf{1}_q^\intercal - \boldsymbol{\delta}_{(i)}^\intercal\right)\boldsymbol{\tau}\right), \tag{9}$$

where $\boldsymbol{\delta}_{(i)}^\intercal$ is the ith row of $\boldsymbol{\Delta}$ and $\mathbf{1}_q$ is a vector of q 1s. Rearranging it further, we have

$$\left(\tau_0 + \sum_{j=1}^{q}\tau_j\right) p_i = a_i + p_i \sum_{j=1}^{q} \Delta_{ij}\tau_j,$$

where Δ_{ij} is the (i, j)th element in matrix $\boldsymbol{\Delta}$, and we have rewritten $\mathbf{1}_q^\intercal\boldsymbol{\tau}$ as the sum of all elements of $\boldsymbol{\tau}$. Summing over i on both sides, we have

$$\sum_{i=1}^{d}\left(\tau_0 + \sum_{j=1}^{q}\tau_j\right) p_i = \sum_{i=1}^{d} a_i + \sum_{i=1}^{d} p_i \sum_{j=1}^{q} \Delta_{ij}\tau_j.$$

On the other hand, from the expression of τ_j in (6) and summing over j, we have

$$\sum_{j=1}^{q} b_j = \sum_{j=1}^{q}\tau_j \sum_{i=1}^{d} \Delta_{ij} p_i.$$

Combining the last two equations with reordering of the summations on the right-hand side of the last equation, we arrive at a global invariance stated as Lemma 2, which can be understood as interpreting the auxiliary variables τ_0 and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_q)^\intercal$ as thicknesses against the probabilities as base areas so that their product is the information volume, or total counts.

Lemma 2 *For the auxiliary variables τ_0 and $\boldsymbol{\tau}$, the observed counts vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, and the probabilities $\boldsymbol{p}$, it holds that*

$$\left(\tau_0 + \sum_{j=1}^{q}\tau_j\right)\left(\sum_{i=1}^{d} p_i\right) = \sum_{i=1}^{d} a_i + \sum_{j=1}^{q} b_j.$$

With the normalization constraint

$$\sum_{i=1}^{d} p_i = 1, \tag{10}$$

Lemma 2 is reduced to

$$\tau_0 + \sum_{j=1}^{q}\tau_j = \sum_{i=1}^{d} a_i + \sum_{j=1}^{q} b_j.$$

Hence, (9) can be simplified to

$$a_i = p_i\left(s - \boldsymbol{\delta}_{(i)}^\intercal\boldsymbol{\tau}\right) \tag{11}$$

where

$$s = \sum_{i=1}^{d} a_i + \sum_{j=1}^{q} b_j \tag{12}$$

is a global constant. We have now completed the derivation of all four MLE equations in (3) respectively as (6), (11), (10), and (12).

2.2 The score function and the observed information matrix

The incomplete multinomial likelihood function (1) has the score function,

$$\nabla_{d-1}\ell(\boldsymbol{p}) = \underbrace{\begin{bmatrix} 1 & \cdots & 0 & -1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & -1 \end{bmatrix}}_{(d-1)\times d} \times \left\{ \begin{bmatrix} \frac{a_1}{p_1} \\ \frac{a_2}{p_2} \\ \vdots \\ \frac{a_d}{p_d} \end{bmatrix} + \begin{bmatrix} \Delta_{11} & \Delta_{12} & \cdots & \Delta_{1q} \\ \Delta_{21} & \Delta_{22} & \cdots & \Delta_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{d1} & \Delta_{d2} & \cdots & \Delta_{dq} \end{bmatrix} \begin{bmatrix} \frac{b_1}{\boldsymbol{\delta}_1^\intercal\boldsymbol{p}} \\ \frac{b_2}{\boldsymbol{\delta}_2^\intercal\boldsymbol{p}} \\ \vdots \\ \frac{b_q}{\boldsymbol{\delta}_q^\intercal\boldsymbol{p}} \end{bmatrix} \right\},$$

and the observed information matrix,

$$-\nabla^2_{d-1}\ell\,(\boldsymbol{p}) = \operatorname{diag}\left(\frac{a_1}{p_1^2},\cdots,\frac{a_{d-1}}{p_{d-1}^2}\right) + \frac{a_d}{p_d^2}\underbrace{\begin{bmatrix}1&\cdots&1\\ \vdots&\ddots&\vdots\\ 1&\cdots&1\end{bmatrix}}_{(d-1)\times(d-1)} + \begin{bmatrix}\psi_{11}&\cdots&\psi_{1,(d-1)}\\ \vdots&\ddots&\vdots\\ \psi_{(d-1),1}&\cdots&\psi_{(n-1),(d-1)}\end{bmatrix} \tag{13}$$

where $\ell(\boldsymbol{p})$ represents the log-likelihood and

$$\psi_{ik} = \sum_{j=1}^{q}\frac{b_j\left(\Delta_{ij}-\Delta_{dj}\right)\left(\Delta_{kj}-\Delta_{dj}\right)}{\left(\boldsymbol{\delta}_j^{\mathsf{T}}\boldsymbol{p}\right)^2}.$$

The inverse of the observed information matrix (13) evaluated at the MLE is used to approximate the asymptotic variance–covariance matrix of the MLE of the first $d-1$ parameters. For the last element of the MLE, $\hat{p}_d$,

$$\begin{aligned}\operatorname{var}\left(\hat{p}_d\right) &= \operatorname{var}\left(\mathbf{1}_{d-1}^{\mathsf{T}}\hat{\boldsymbol{p}}_{[1..(d-1)]}\right)\\ &= \mathbf{1}_{d-1}^{\mathsf{T}}\operatorname{var}\left(\hat{\boldsymbol{p}}_{[1..(d-1)]}\right)\mathbf{1}_{d-1},\end{aligned} \tag{14}$$

$$\begin{aligned}\operatorname{cov}\left(\hat{p}_d,\hat{p}_i\right) &= \operatorname{cov}\left(\hat{p}_i, 1-\sum_{k=1}^{d-1}\hat{p}_k\right)\\ &= -\sum_{k=1}^{d-1}\operatorname{cov}\left(\hat{p}_i,\hat{p}_k\right),\end{aligned} \tag{15}$$

where the subscript $[1..(d-1)]$ denotes the sub-vector of the first $d-1$ elements.

3 Algorithms

The first two equations in (3) suggest a fixed-point iteration on the data structure illustrated in Fig. 1. We refer to Algorithm 1 as the weaver algorithm. Like the mechanical weaving machine, its operations are highly parallelizable. Prior incorporation to the weaver algorithm simply means adjusting the two counts vectors and adding new rows to the matrix $\boldsymbol{\Delta}^{\mathsf{T}}$ for new sub-sums carried by the prior. This ability to easily accommodate Bayesian modeling mitigates the MLE's "over-fitting" tendency (Guiver and Snelson 2009; Caron and Doucet 2012). It also means that weaver is an online algorithm.

3.1 Convergence measure

Convergence in the log-likelihood sequence becomes difficult to discern after it reaches a plateau. A better view can

Algorithm 1 Weaver

0. Initialize $\boldsymbol{p} = (1/d, ..., 1/d)$ and define the scalar $s = \operatorname{sum}(\boldsymbol{a}) + \operatorname{sum}(\boldsymbol{b})$.
1. Compute $\boldsymbol{\tau} = \boldsymbol{b}/(\boldsymbol{\Delta}^{\mathsf{T}}\boldsymbol{p})$ (element-wise division).
2. Update $\boldsymbol{p} = \boldsymbol{a}/(s\mathbf{1}_d - \boldsymbol{\Delta}\boldsymbol{\tau})$ (element-wise division and subtraction).
3. Normalize $\boldsymbol{p} = \boldsymbol{p}/\operatorname{sum}(\boldsymbol{p})$.
4. Repeat 1–3 till convergence.

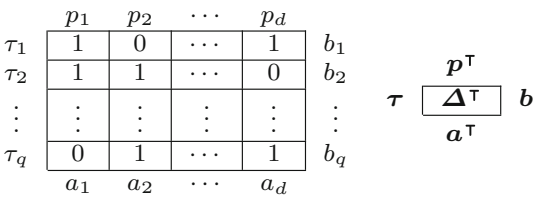


Fig. 1 Data structure for the weaver algorithm

be obtained if we switch to the space of the exponents $\boldsymbol{a}$ and $\boldsymbol{b}$. The kth iteration $\hat{p}_i^{(k)}$ can reconstruct the corresponding complete exponent $\hat{a}_i^{(k)}$ by

$$\hat{a}_i^{(k)} = \hat{p}_i^{(k)}\left(s - \sum_{j:\Delta_{ij}=1}\tau_j^{(k)}\right).$$

Similarly, $\boldsymbol{b}$'s elements are reconstructed by

$$\hat{b}_j^{(k)} = \tau_j^{(k)}\left(\sum_{i:\Delta_{ij}=1}\hat{p}_i^{(k)}\right).$$

We therefore define the sum of squared errors (SSE),

$$\mathrm{SSE} = \sum_{i=1}^{d}\left(\hat{a}_i^{(k)} - a_i\right)^2 + \sum_{j=1}^{q}\left(\hat{b}_j^{(k)} - b_j\right)^2$$

for the kth iteration. The advantage of this definition is that the new sequence have a known limit $(\boldsymbol{a}, \boldsymbol{b})$. For the likelihood of Example 2, nine iterations reach $\sqrt{\mathrm{SSE}} = 10^{-3}$; another six more iterations reach $\sqrt{\mathrm{SSE}} = 10^{-6}$; even further seven iterations reach $\sqrt{\mathrm{SSE}} = 10^{-9}$. The weaver algorithm converges at a linear rate; it is also ascent at every step. A proof of both statements is given in "Appendix C." Figure 2 plots the convergence of the weaver algorithm applied to the likelihood of Example 2. In practice, we believe 30 steps of the weaver iterations are adequate to satisfy the relatively low precision requirement of most problems.

3.2 Relation with the Ford (1957) algorithm

Ford (1957) proved a sufficient condition formulated in graph-theoretical terms for the existence and uniqueness

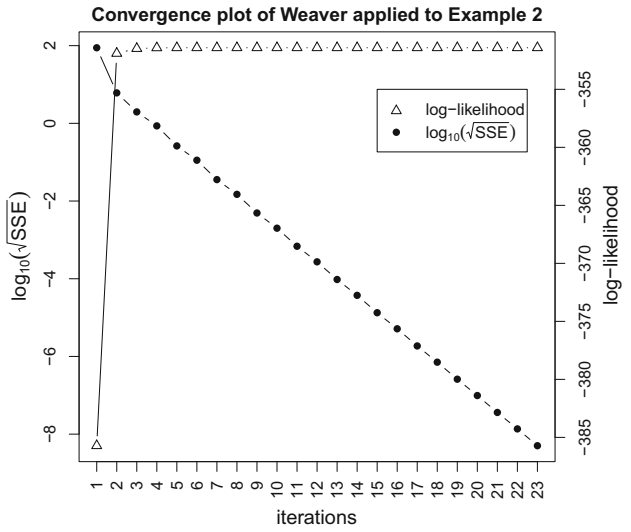


Fig. 2 Convergence plot of the Algorithm 1 (Weaver) being applied to the likelihood of Example 2. The iteration path shows that the log-likelihood is increased at every step and the rate of convergence is linear (c.f. the proof in "Appendix C")

of the Bradley–Terry MLE solution, essentially connecting the paired comparison model to a weighted directed graph. In fact, every edge in the typical graph connects only two vertices. An adjacency matrix can be used to represent a Bradley–Terry likelihood. The following zero diagonal, asymmetric square matrix of nonnegative elements is an example in Ford (1957),

$$\boldsymbol{A} = (a_{ij}) = \begin{bmatrix} 0 & 15 & 15 & 0 \\ 11 & 0 & 10 & 20 \\ 11 & 10 & 0 & 20 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

where each a_{ij} represents the number of wins of i over j. The following algorithm is used there for the MLE:

$$p_i \leftarrow \left(\sum_{j=1}^{d} a_{ij}\right) \Bigg/ \left(\sum_{j=1}^{d} \frac{a_{ij} + a_{ji}}{p_i + p_j}\right), \tag{16}$$

where a_{ij} is the (i, j)th element in the adjacency matrix. The ith row sum $\sum_{j=1}^{d} a_{ij}$ corresponds to a_i, the anti-diagonal pair sum $a_{ij} + a_{ji}$ corresponds to the opposite sign of one element of the variable cell counts vector $\boldsymbol{b}$, and the probability sub-sum $p_i + p_j$ corresponds to one element of $\boldsymbol{\Delta}^{\mathsf{T}}\boldsymbol{p}$ in the Algorithm 1 notation. The Ford algorithm can be rewritten as

$$p_i \leftarrow \frac{a_i}{\boldsymbol{\delta}_{(i)}^{\mathsf{T}}(-\boldsymbol{b}/\boldsymbol{\Delta}^{\mathsf{T}}\boldsymbol{p})}$$

which corresponds to Algorithm 1 with $s = 0$.

In any adjacency matrix representation, the sum of all row sums must equal to that of all anti-diagonal pairwise sums. Since, as mentioned above, each row sum corresponds to an element of $\boldsymbol{a}$ and each anti-diagonal pairwise sum corresponds to -1 times an element of $\boldsymbol{b}$, the right-hand side of Lemma 2 equals zero for any Bradley–Terry likelihood. It is understood now as the exact condition that permits the original derivation of (16) by unconstrained maximization. The Plackett–Luce model also has the zero sum property. The sum of all exponents represents the total volume of the information in the joint likelihood, which should be nonnegative. For a typical example of the pathological case, $p_1 p_2 p_3 (p_1 + p_2)^{-4}$ has this sum equal to -1 and diverges near the vertex $(0, 0, 1)$.

3.3 The Bayesian weaver

When a p_i is very small relative to the other probabilities, we say the signal is weak at p_i. For MLE containing a weak signal, the solution is very close to the boundary of the simplex. There is a chance for the weaver path to cross that boundary before reaching the solution. In such cases, thickening the complete counts vector $\boldsymbol{a}$ can help enhance the stability of the iteration. The Bayesian weaver implements this idea using a Dirichlet prior. For any incomplete multinomial likelihood, if the prior mode and the posterior mode are the same, then the MLE equals to them. One can then develop a new algorithm by making the prior mode converging to the posterior mode. Computationally, this means to run a two-layer iteration. The master sequence converges to the true maximizer/mode while for each term in the master sequence there is a weaver iteration converging to the term. Specifically, we add some positive counts to the complete data by an amount equal to the rescaled current MLE solution to produce a posterior. A weaver is then launched to find the posterior mode which determines a new prior. The iteration repeats to produce a posterior mode sequence whose limit is the MLE. Algorithm 2 describes the steps of the Bayesian weaver. Our experience suggests setting the prior thickness $\gamma = \sum_{j=1}^{q} |b_j|$ or slightly more than that.

Algorithm 2 Bayesian weaver

0. Initialize $p = (1/d, ..., 1/d)$ and choose the prior thickness γ.
1. Multiply the Dirichlet prior $\prod_{i=1}^{d} p_i^{\gamma p_i}$ to the original incomplete multinomial likelihood to produce a posterior incomplete multinomial F_{new}.
2. Find the MLE of F_{new} and update $\boldsymbol{p}$ to this value.
3. Repeat 1–2 till convergence.

Table 2 Simulation results for MLE's large-sample property under Example 2

Simulation results of Sect. 4.1.1.					
True $\boldsymbol{p}$					
0.1654	0.2024	0.1444	0.1532	0.2301	0.1046
Theoretical asymptotic covariances Σ *of MLE* $\hat{\boldsymbol{p}}$ $(\times 10^{-4})$					
5.33	−2.20	−1.56	1.63	−2.20	−1.00
−2.20	11.52	−2.32	−1.69	−4.51	−0.80
−1.56	−2.32	8.35	−1.23	−1.35	−1.89
1.63	−1.69	−1.23	4.98	−2.54	−1.15
−2.20	−4.51	−1.35	−2.54	12.84	−2.24
−1.00	−0.80	−1.89	−1.15	−2.24	7.08
Theoretical asymptotic $\mathtt{SE}(\hat{\boldsymbol{p}})$ *as* $\sqrt{\mathtt{diag}\,\Sigma}$					
0.0231	0.0339	0.0289	0.0223	0.0358	0.0266
Sample mean of $\hat{\boldsymbol{p}}$*'s computed for every simulation*					
0.1653	0.2026	0.1445	0.1532	0.2299	0.1045
Sample covariance of the $\hat{\boldsymbol{p}}$*'s* $(\times 10^{-4})$					
5.45	−2.31	−1.63	1.59	−2.15	−0.95
−2.31	11.43	−2.45	−1.74	−4.11	−0.80
−1.63	−2.45	8.58	−1.26	−1.32	−1.92
1.59	−1.74	−1.26	5.00	−2.47	−1.12
−2.15	−4.11	−1.32	−2.47	12.17	−2.11
−0.95	−0.80	−1.92	−1.12	−2.11	6.91
Mean of the covariances computed for every simulation $(\times 10^{-4})$					
5.44	−2.28	−1.62	1.55	−2.13	−0.96
−2.28	11.36	−2.42	−1.74	−4.09	−0.83
−1.62	−2.42	8.52	−1.25	−1.38	−1.85
1.55	−1.74	−1.25	4.93	−2.40	−1.10
−2.13	−4.09	−1.38	−2.40	12.05	−2.05
−0.96	−0.83	−1.85	−1.10	−2.05	6.79
Mean of the $\mathtt{SE}(\hat{\boldsymbol{p}})$*'s computed for every simulation*					
0.0233	0.0336	0.0291	0.0221	0.0347	0.0259

4 Simulations

Simulations are devised in this section to demonstrate usage of the model, the weaver algorithms to obtain the MLE, and the covariance formulae.

4.1 Simulation for large-sample properties of the MLE based on Example 2

We simulate data based on the same configuration as Example 2, which involves one complete sample, two marginal samples, and one conditional sample. The total counts of the samples are fixed at 120, 40, 40, and 100, respectively. The true probabilities are set to the MLE of the likelihood (2). Note that the marginal samples and the conditional sample all follow multinomial distributions with probabilities adapted to marginal spaces or normalized to a restricted space. Each simulation generates a new set of four samples by multinomially resampling the original four, then puts them into an incomplete multinomial likelihood, and finally computes an MLE and an observed information matrix. After N simulations, we obtain N pairs of the MLE and the observed information matrix.

4.1.1 Large-sample mean and covariance of the MLE

We performed $N = 120,000$ simulations. The mean vector and sample covariance matrix of the N MLEs, the mean matrix of the estimated asymptotic covariance matrices of the MLEs, and the mean vector of the standard errors of the MLEs are reported in Table 2 with comparison to the true probability, the theoretical covariance derived from (13), (14), and (15). The results show that the estimates are very close to their true values.

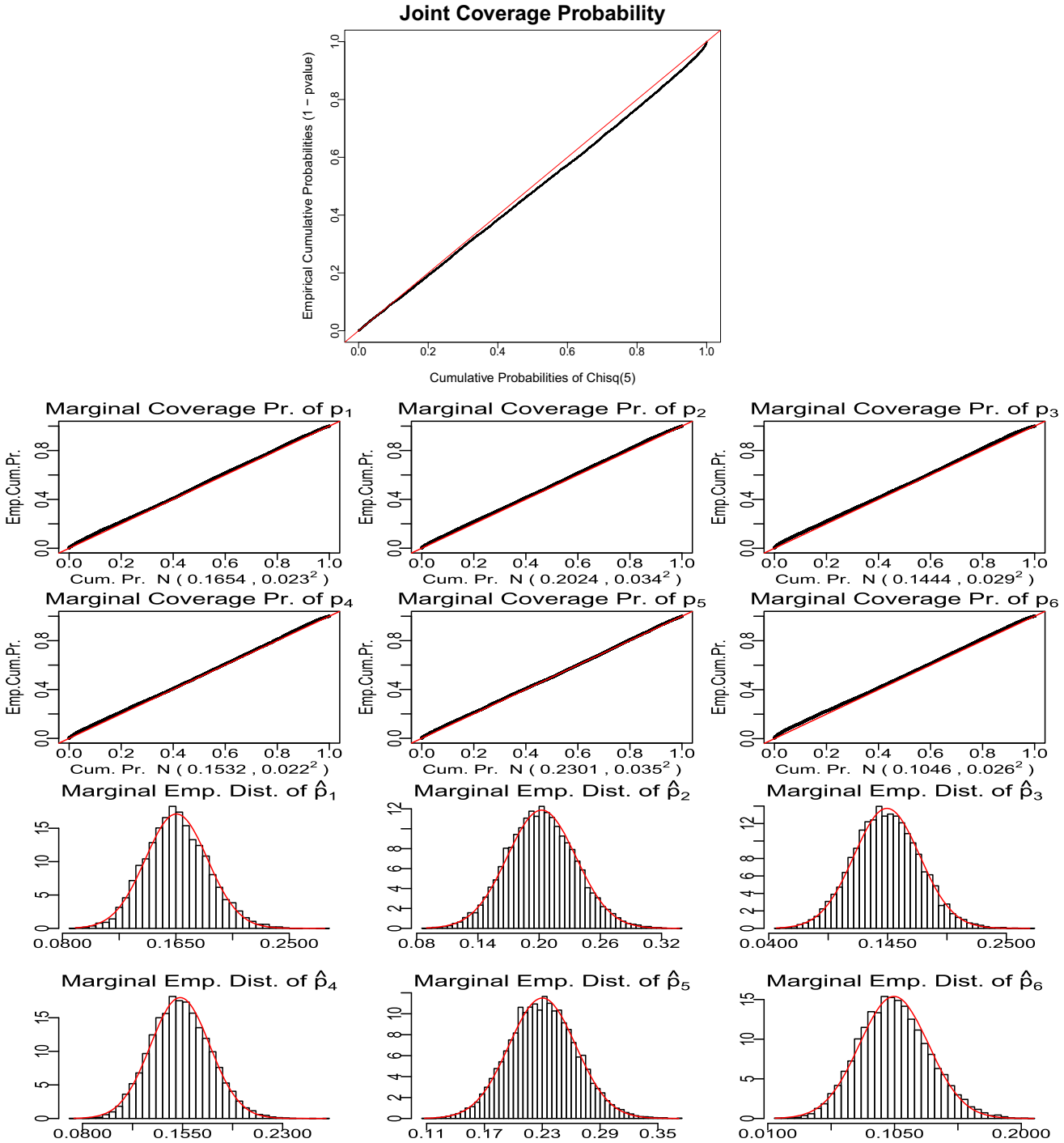


Fig. 3 Simulation based on Example 2: Joint (upper) and marginal (lower) coverage probabilities of the MLE confidence region and intervals

4.1.2 *Coverage probability*

To assess coverage of the true probability vector by the joint confidence region, we use the χ^2-distributed squared Mahalanobis distance between the true $\boldsymbol{p}$ and the MLE,

$$D^2\left(\boldsymbol{p}_{[1..5]}, \hat{\boldsymbol{p}}^{(i)}_{[1..5]}\right) = \left(\boldsymbol{p}_{[1..5]} - \hat{\boldsymbol{p}}^{(i)}_{[1..5]}\right)^{\mathsf{T}} \left(\boldsymbol{S}^{(i)}\right)^{-1} \left(\boldsymbol{p}_{[1..5]} - \hat{\boldsymbol{p}}^{(i)}_{[1..5]}\right) \overset{\text{iid}}{\sim} \chi^2 \quad (5)$$

where $\boldsymbol{S}^{(i)}$ is the asymptotic covariance of the ith MLE $\hat{\boldsymbol{p}}^{(i)}$, and the [1..5] subscript denotes the sub-vector of the first five elements. The top panel in Fig. 3 is a plot of the sampling frequency of accepting the null hypothesis,

$$\frac{1}{N} \times \#\left\{i \in \{1, \ldots, N\} : D^2\left(\boldsymbol{p}_{[1..5]}, \hat{\boldsymbol{p}}^{(i)}_{[1..5]}\right) \leqslant \chi^2_q(5)\right\}$$

versus the quantity $q \in \left\{\frac{1}{N}, \frac{2}{N}, \ldots, 1\right\}$ which is the confidence level and regulates the size of the elliptical confidence region. The joint coverage probability of the true parameter by the region of confidence level q should converge to q when sample size increases to infinity, as will be verified in the simulation of Sect. 4.3.

The marginal coverage probability reduces to the standard normal case and is computed for every $p_1, \ldots, p_6$ at all confidence levels. At the 95% confidence level, these empirical coverage probabilities based on the $N = 120,000$ simulations are estimated as 94.35, 94.38, 94.16, 94.41, 94.32, 93.68%, respectively. Also included in Fig. 3 are histograms of the marginal distributions of every $\hat{p}_i$ superposed with the respective normal density curves. The joint and marginal coverage probability computations, together with the normality checks, verify the correctness of both the MLE algorithm and the asymptotic covariance formula for the incomplete multinomial model.

4.2 Parameter identification based on the modified Example 2

If we estimate the PMF using only the incomplete samples 2–4, the parameters p_2, p_3, p_5, and p_6 as a group is only identifiable to two ratios $(p_2 + p_3) : (p_5 + p_6)$, from sample 2, and $(p_2 + p_5) : (p_3 + p_6)$, from sample 3. The following table reports the mean MLE and the mean standard error (SE) of MLE based on $N = 120,000$ simulated data from the same true $\boldsymbol{p}$.

mean $\hat{\boldsymbol{p}}$			mean SE($\hat{\boldsymbol{p}}$)		
0.1654	0.2204	0.1261	0.0410	**5805**	**5805**
0.1533	0.2119	0.1229	0.0384	**5805**	**5805**

All p_2, p_3, p_5, and p_6 have a very large standard error which indicates their unidentifiability. Note that the group does succeed to identify the two ratios: $(p_2+p_3) : (p_5+p_6)$ is estimated at 1.035 compared to the true value 1.036, and $(p_2 + p_5) : (p_3 + p_6)$ estimated at 1.733 compared to the true 1.737.

We now simulate for a modified case constructed to determine the missing ratio: $(p_2 + p_6) : (p_3 + p_5)$. Toward this goal, a 5th conditional binomial sample of size 40 is added. The two categories in this sample are (i) the union of middle-aged female and senior male, and (ii) the union of senior female and middle-aged male. Figure 4 illustrates the modified design.

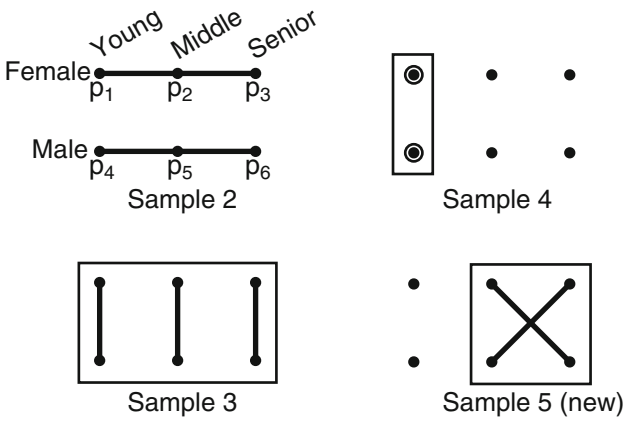


Fig. 4 Simulation study of parameter identification: Modified design to determine the ratio $(p_2 + p_6) : (p_3 + p_5)$

Middle-aged female or senior male	$p_2 + p_6$
Senior female or middle-aged male	$p_3 + p_5$

The joint likelihood now adopts the form

$$\begin{aligned} L_*(\boldsymbol{p}) = {} & (p_1 + p_2 + p_3)^{b_1}(p_4 + p_5 + p_6)^{b_2} \\ & \times (p_1 + p_4)^{b_3}(p_2 + p_5)^{b_4}(p_3 + p_6)^{b_5} \\ & \times p_1^{a_1} p_4^{a_4}(p_1 + p_4)^{b_6} \\ & \times (p_2 + p_6)^{b_7}(p_3 + p_5)^{b_8}(p_2 + p_3 + p_5 + p_6)^{b_9}, \end{aligned}$$

where $b_6 = -(a_1 + a_4) < 0$ and $b_9 = -(b_7 + b_8) < 0$. As a result, the mean MLE and mean SE of MLE from the modified $N = 120,000$ simulations are given by

mean $\hat{\boldsymbol{p}}$			mean SE($\hat{\boldsymbol{p}}$)		
0.1653	0.2023	0.1441	0.0410	0.0613	0.0593
0.1531	0.2298	0.1053	0.0384	0.0621	0.0582

Note that all SE's are estimated at reasonable levels and all elements in the MLE vector are close to the true parameters.

4.3 Simulations of general partition patterns

A true $\boldsymbol{p} \in T^\circ_{d-1}$ is specified to generate n incomplete multinomial likelihoods $L^{(i)}(\boldsymbol{p})$, $i = 1, \ldots, n$. For each likelihood $L^{(i)}(\boldsymbol{p})$, the weaver algorithm is applied to find the MLE, $\hat{\boldsymbol{p}}^{(i)}$. The empirical distribution of the MLEs is studied to confirm the reduction of variance as sample size

increases. The following describes the two-stage scheme of the simulation,

$$\boldsymbol{p} \to \left(\boldsymbol{a}^{(i)}, \boldsymbol{b}^{(i)}, \boldsymbol{\Delta}^{(i)}\right) \to \hat{\boldsymbol{p}}^{(i)}, \text{ for } i = 1, \ldots, n. \tag{17}$$

The first stage simulates the ith incomplete multinomial likelihood represented by the tuple $(\boldsymbol{a}^{(i)}, \boldsymbol{b}^{(i)}, \boldsymbol{\Delta}^{(i)})$, and the second stage employs the weaver algorithm to find the MLE of the ith likelihood. The simulation of the tuple $(\boldsymbol{a}^{(i)}, \boldsymbol{b}^{(i)}, \boldsymbol{\Delta}^{(i)})$ consists of simulating R multinomial likelihood slices which are subsequently multiplied into a single joint likelihood.

4.3.1 Random partition

The simulation of every multinomial slice involves a random partition process of the sample space described as follows. Denote by $(\Omega, \mathcal{A}, \mathbb{P})$ the probability space implied by the PMF, $\boldsymbol{p}$, and the most basic events in $\mathcal{A}$ by $e_1, \ldots, e_d$. Denote the $\mathbb{P}$-measurable outcome partition by π, defined as $\pi = \{S_0, \ldots, S_k\} \subset \mathcal{A}$ such that $2 \le k \le d$, $\bigcup_{i=1}^{k} S_i = \Omega$, $S_i \cap S_j = \emptyset$ for any $i \ne j$, and $S_i \ne \emptyset$ for any $i \ge 1$. Denote the random partition process that produces π by $\mathcal{P}$, which consists of the following two random draws.

1. Draw $\bar{k} \sim U[2, d]$ as the upper bound of k, the number of subsets in the partition.
2. Draw $\gamma_1, \ldots, \gamma_d \overset{\text{iid}}{\sim} U\left[0, \bar{k}\right]$ the group labels for each of the d elements and resolve $k = \#\{i : \gamma_i > 0\}$ as the number of unique group labels except for the label 0. Let $\left\{\zeta_1, \ldots, \zeta_k : 1 \le \zeta_i \le \bar{k}\right\}$ be the set of unique positive values realized by $\gamma_1, \ldots, \gamma_d$ and define $S_i = \bigcup_{j:\gamma_j = \zeta_i} e_j$. Then $\pi^+ = \{S_i : i = 1, \ldots, k\}$ is the outcome partition before conditioning. The group label 0 is included to enable the simulation of conditioning patterns. When $\left\{j : \gamma_j = 0\right\} \ne \emptyset$, we let $S_0 = \bigcup_{j:\gamma_j = 0} e_j$ and use its complement $S_0^c = \Omega \backslash S_0$ as the conditioning set. In the case of $\{j : \gamma_j = 0\} = \emptyset$, we let $S_0 = \emptyset$. The conditioning set S_0^c is incorporated in the final likelihood with a negative exponent (i.e., on the denominator) being equal to the minus sum of all exponents of the positive indexed subsets in the current draw. For example, suppose $d = 5$ and the current draw of $(\gamma_1, \ldots, \gamma_5)$ realizes to $(0, 0, 5, 3, 3)$, then $S_0 = \{e_1, e_2\}$, $S_1 = \{e_3\}$, $S_2 = \{e_4, e_5\}$ and the structure of the corresponding likelihood component is

$$\frac{{p_3}^{\alpha}(p_4 + p_5)^{\beta}}{(1 - p_1 - p_2)^{\alpha+\beta}} = \frac{{p_3}^{\alpha}(p_4 + p_5)^{\beta}}{(p_3 + p_4 + p_5)^{\alpha+\beta}}$$

for any exponents (α, β) subsequently drawn based on the partition from a multinomial distribution. The final outcome partition is $\pi = \pi^+ \cup \{S_0\}$.

Let R be the number of multinomial slices whose product is the final incomplete multinomial likelihood. Let m be a prespecified constant regulating the level of the total counts in a single multinomial slice. Each of the R multinomial likelihood slices is mapped to the random tuple

$$(\boldsymbol{y}, \pi, \nu) \sim U(\nu|m)\mathcal{P}(\pi|\boldsymbol{p})M\left(\boldsymbol{y}|\nu, \boldsymbol{p}^{\pi}\right),$$

where the multinomial parameter $\nu \sim U[md, 2md]$ as the total sum of exponents is drawn from the uniform distribution, the vector $\boldsymbol{y} \in \mathbb{Z}^d$ is the multinomial outcome drawn from $M(\boldsymbol{y}|\nu, \boldsymbol{p}^{\pi})$. The R slices are then transformed into the standard incomplete multinomial parameterization for the ith joint likelihood

$$\left(\boldsymbol{a}^{(i)}, \boldsymbol{b}^{(i)}, \boldsymbol{\Delta}^{(i)}\right) = \mathrm{g}\left(\boldsymbol{y}^{(i_s)}, \pi^{(i_s)}, \nu^{(i_s)}; s = 1, \ldots, R\right),$$

where the complete exponent vector $\boldsymbol{a}^{(i)}$ is increased by $c\boldsymbol{\alpha}^{(i)}$ with $c = 1$ and $\boldsymbol{\alpha}^{(i)} \sim M(\cdot|\sum_{s=1}^{R} \nu^{(i_s)}, \boldsymbol{p})$. Here, c regulates the proportion of the complete-data part in the likelihood; for example, in Example 2, c rescales the counts of Sample 1. The transformation algorithm g splits $\boldsymbol{y}$ into $\boldsymbol{a}$ and $\boldsymbol{b}$ by separating the columns of the binary matrix π, which uses binary vectors to represent the sets S_i, between columns that have only a single 1 (whose $\boldsymbol{y}$-counts go to $\boldsymbol{a}$) and have multiple 1s (whose $\boldsymbol{y}$-counts go to $\boldsymbol{b}$). The multiple-1 columns of π make $\boldsymbol{\Delta}$.

4.3.2 Intrinsic sample size of a single incomplete multinomial likelihood

For the ith likelihood, the intrinsic sample size is controlled by m, for the level of total exponents in multinomial sampling, and R, for the *number* of multinomial samples contained in a single incomplete multinomial likelihood. In a sense, a single sample from the incomplete multinomial distribution encapsulates an intrinsic sample size of the order $O(m \times R)$. Implementing the scheme (17) explained in the previous section, we choose a true vector of parameters,

$$\boldsymbol{p} = \frac{1}{5050}(1, 2, \ldots, 100)$$

and simulate an incomplete multinomial likelihood for every pair (m, R) from the grid $\{1, 10, 100, 1000\} \times \{2, 4, 8, 16\}$, and compute its MLE, $\hat{\boldsymbol{p}}$, using the weaver algorithm. The comparison of $\hat{\boldsymbol{p}}$ vs. $\boldsymbol{p}$ is reported in Fig. 5 using PP–plots. The convergence of $\hat{\boldsymbol{p}} \to \boldsymbol{p}$ improves as m and/or R increases, as shown in the PP–plots array.

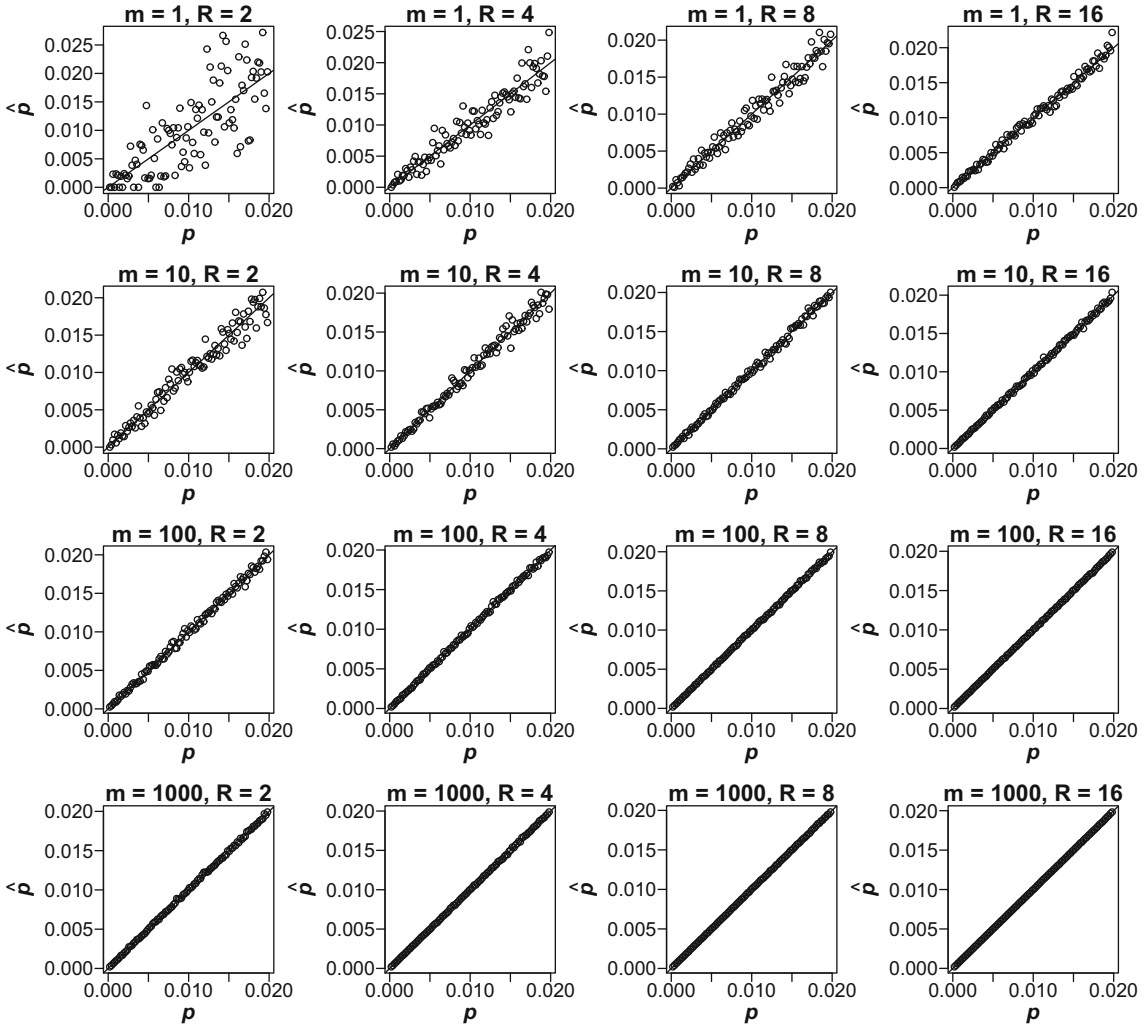


Fig. 5 Simulating counts on a random partition: PP–plots of $\hat{\boldsymbol{p}}$ vs. $\boldsymbol{p}$ for each pair of m and R

To evaluate the variance of the MLE, we obtain an empirical distribution of the MLE based on 600 simulations, with one of the pair $(m,\ R)$ fixed at the stressed level of 2 and the other taking values from 2 to a large value (2^{10} for m and 20 for R). Without loss of generality, we report results only for the smallest and the largest parameters $\hat{p}_1$ and $\hat{p}_{100}$ by making a box plot for every pair of $(m,\ R)$ simulated. The box plots in Fig. 6 show steady reduction of the variances as either m or R increases. Note that there are two anomalies for the sample median of $\hat{p}_1$ occurring at $(m, R) = (2, 2)$ and $(4, 2)$, where the medians drop to zero. This is immediately understood as a large number of the 600 simulated likelihoods have realized zero count on the first cell with true probability $p_1 = 1/5050$ due to the small intrinsic sample size. In other words, the signal is very weak for the first cell and the intrinsic sample size is too small to capture any trace of it. Despite this, the means of $\hat{p}_1$ and $\hat{p}_{100}$ over the 600 simulated values still estimate the respective parameters with small biases.

4.4 Estimation of weak signals

In this simulation, we first study the weak signal performance using the following true PMF,

$$\boldsymbol{p} = (p_1, \ldots, p_{100}) = \frac{1}{9901}(1, 100, \ldots, 100),$$

where all elements of $\boldsymbol{p}$ are the same, except for p_1, here set to be one-100th of the rest. We use the weak signal (ws) likelihood $L_{\mathrm{ws}}(\boldsymbol{p})$ constructed to have the MLE exactly equal to the true $\boldsymbol{p}$,

$$\begin{aligned} L_{\mathrm{ws}}(\boldsymbol{p}) = {} & p_1 p_2^{100} \cdots p_{100}^{100} \\ & \times (p_1 + p_2)^{101} (p_3 + p_4)^{200} \cdots (p_{99} + p_{100})^{200} \\ & \times p_1 (p_2 + p_3)^{200} \cdots (p_{96} + p_{97})^{200} \\ & (p_{98} + p_{99} + p_{100})^{300}. \end{aligned}$$

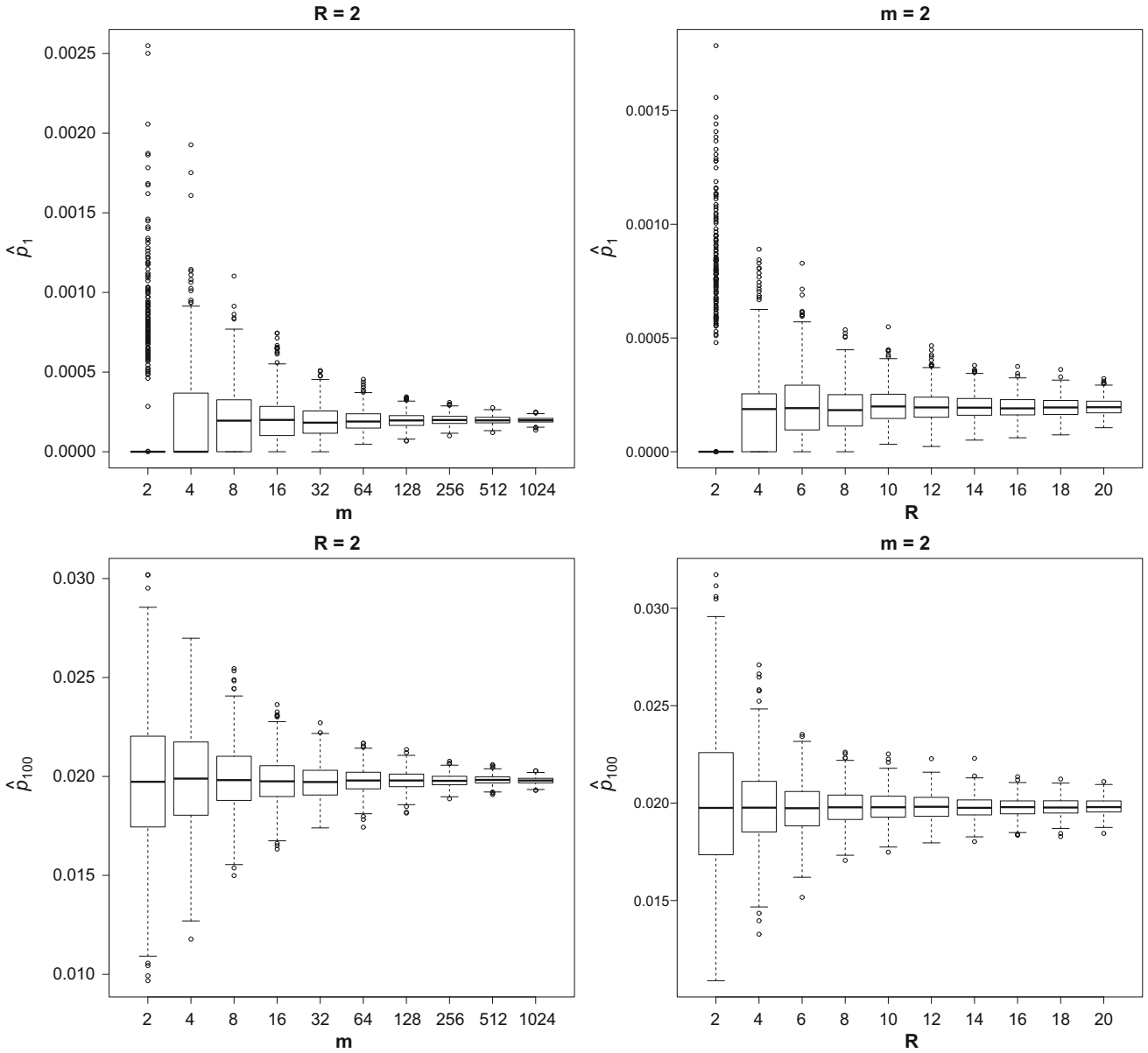


Fig. 6 Simulating counts on a random partition: empirical distributions of $\hat{p}_1$ and $\hat{p}_{100}$ based on 600 simulations for every combinations of R and m with one of the pair set at the small value of 2 while the other increases from 2 to a large value

Figure 7 plots the iterations of a Bayesian weaver solving $\hat{p}_1 = 0.0001009998990001$, achieving correct 16 places after the decimal point.

The above shows under an artificial noiseless condition, a single weak signal of 1/100 intensity can be recovered almost arbitrarily well. Next, we simulate under a more realistic and stressed condition with sampling noise and multiple much weaker signals. The unnormalized true PMF has length 60 and consists of three segments of equal sizes and drastically increasing magnitudes:

- 20 draws from $U(0, 1)$,
- 20 draws from $U(100, 1000)$, and
- 20 draws from $U(10000, 100000)$.

Thus, the largest element of the PMF can be more than 1 million times the smallest. Having multiple very weak signals in the true PMF means the solution is very close to a 20-dimensional hyper-plane boundary of the 60-dimensional simplex; and locally to that hyper-plane, the solution is close to a lower dimensional boundary. The purpose is to show the algorithms developed in this paper is capable of finding the solution when it falls in that region. The generation of the likelihood functions uses the same random partition

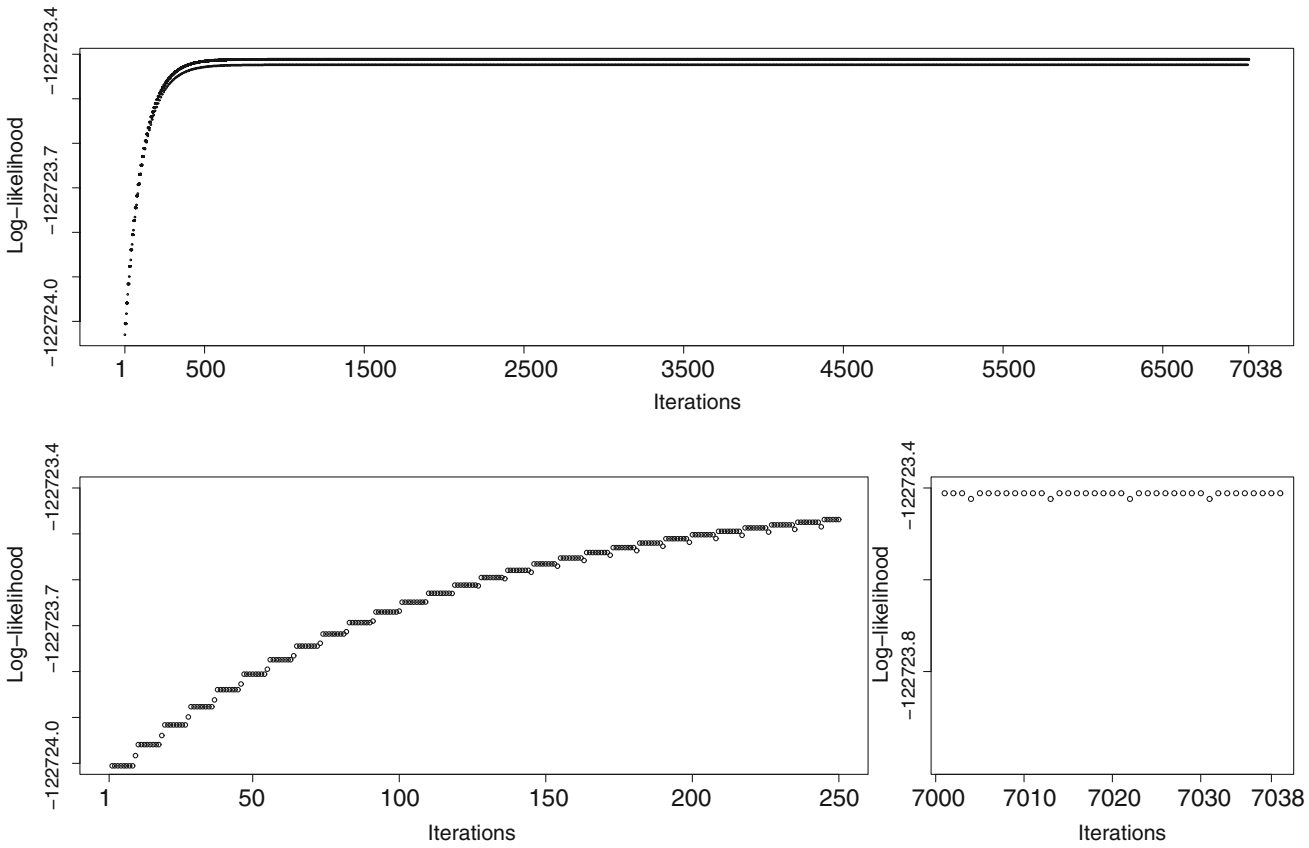


Fig. 7 Estimating a weak signal p_1 using the Bayesian weaver. The MLE solution is accurate to the 16th place after decimal

algorithm of Sect. 4.3.1. To add even more burden, we used $c = 0.001$ to reduce the complete-data part, where c is described in Sect. 4.3.1 near the description of g. We generate three likelihoods under increasing (m, R)=(100, 8), (1000, 32), and (10000, 128). Figure 8 plots the estimated $\hat{\boldsymbol{p}}$ vs. the true $\boldsymbol{p}$ generating the likelihood. As (m, R) increase, the estimation of the very weak signals (first column in Fig. 8) improves.

5 Discussion

5.1 Comparison with the EM and MM algorithms

The EM algorithm (Dempster et al. 1977; Wu 1983; McLachlan and Krishnan 2008) is an elegant technique to compute the MLE for a broad class of likelihood functions. Its implementation depends on a creative success in augmenting the likelihood function of the observed data into a likelihood function of the complete data under the constraint that the expectation of the latter equals to the former, functionwise. An implicit goal is that the maximization of the complete-data likelihood should be much simplified as a result of its construction. The maximization is called the M-step; the mean function constraint is called the E-step; the construction of the complete-data likelihood is called the imputation step. A common procedure for constructing a complete-data likelihood is to first identify a sufficient statistic, T, in the observed-data likelihood, then make T a (measurable) function of the latent variables, and finally replace it by that function of the latent variables. The complete-data likelihood constructed by this procedure automatically satisfies the mean function constraint. The MM algorithm (Lange et al. 2000; Hunter and Lange 2004; Lange 2013) is based on a similar idea but a more general construction. Like EM, instead of directly dealing with the objective function, the MM algorithm constructs a maximization-friendly surrogate function and imposes the so-called minorization conditions that the surrogate should be dominated by the objective function and that the two functions have a single touch point which is the current iteration of the parameter.

The Bradley–Terry and Plackett–Luce type models, both adopting the form of the incomplete multinomial likelihood, have been historically difficult to compute because (i) the parameter vector is typically high dimensional, causing many generic convex optimization methods to perform inefficiently; (ii) a naive construction of the EM algorithm by simply splitting out the probability sub-sums fails to work

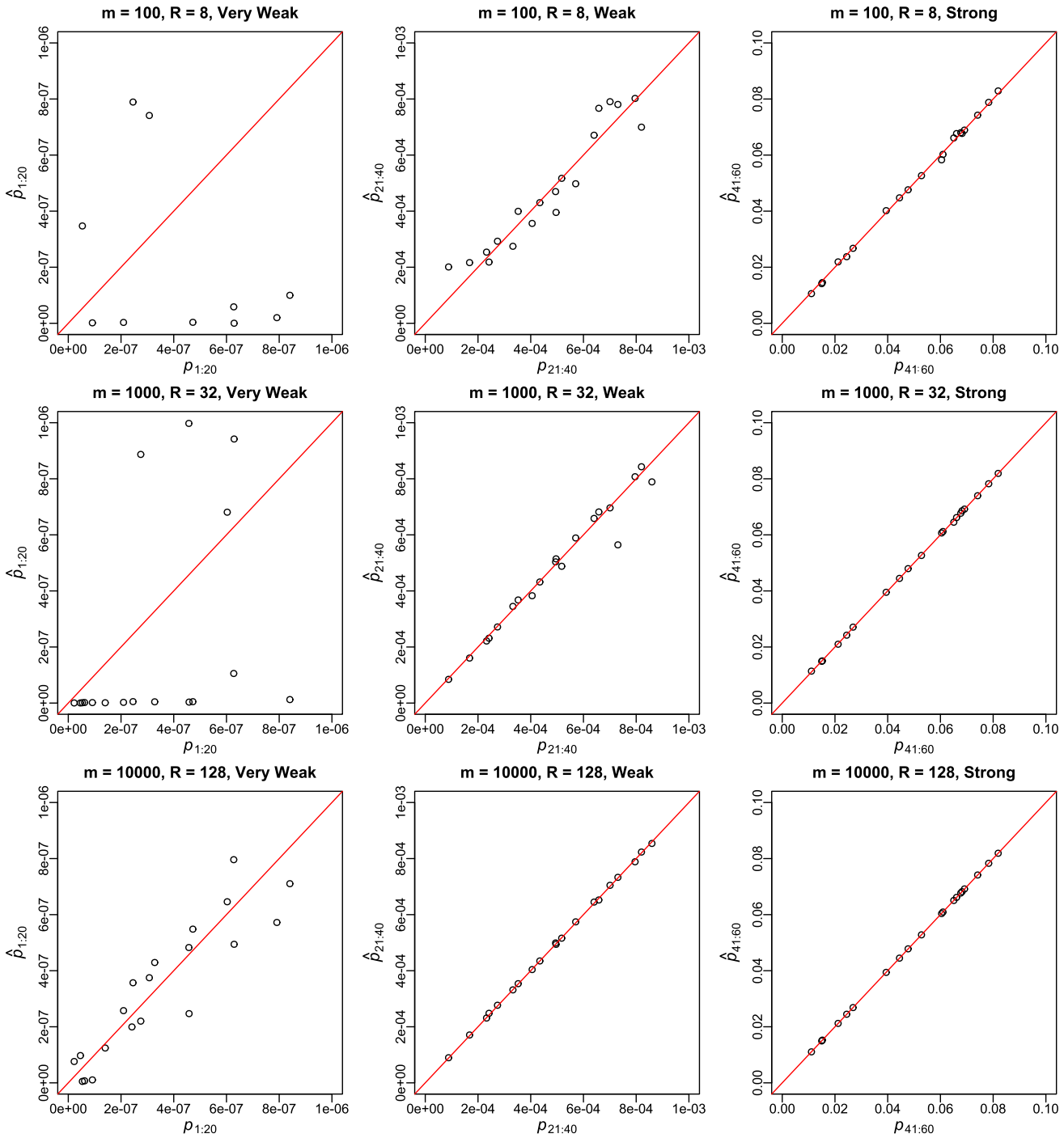


Fig. 8 Estimating weak and very weak signals by three likelihoods (top to bottom) generated from increasing (m, R). Horizontal axis: the true PMF for data generation; vertical axis: MLE of the PMF solved from the generated likelihoods. The PMF is of length 60, equally split to 20 very weak (left column), 20 weak (middle column), and 20 strong (right column) signals. The strongest signal is more than 1 million times the weakest. Some of the very weak signals are outside the frame of the top-left panel

because the sub-sums all appear in the denominator of the two likelihoods. This is fatal to such construction as each iteration tends to decrease the complete-data likelihood and move away from the solution.

The NASCAR2002 car racing data set is a benchmark for computational methods developed for the Bradley–Terry type model. The data set can be captured by a 1543-term

Table 3 Comparison of computing time for the MLE of a Plackett–Luce likelihood on the car racing data in the top panel; and comparison of the elapsed time for the serial and parallel weavers applied to three large incomplete multinomial likelihood functions with $d = 1000$ and $q = 1000, 10,000, 100,000$ in the bottom panel. The likelihood functions are generated by the random partition algorithm described in Sect. 4.3

Time elapsed for 100 calls	EM/MM	Serial weaver
	0.359 sec	0.189 sec

$d \equiv 1000$, q	Serial weaver	Parallel weaver
1000	0.184 sec	0.070 sec
10000	3.424 sec	0.463 sec
100000	38.047 sec	4.061 sec

Plackett–Luce likelihood

$$L_{\mathrm{PL}}(\boldsymbol{p}) = \frac{p_{83}}{p_{83}+p_{18}+p_{20}+\cdots} \times \frac{p_{18}}{p_{18}+p_{20}+\cdots} \times \frac{p_{20}}{p_{20}+\cdots} \times \cdots\cdots \times \frac{p_{53}}{p_{53}+p_{38}+p_{14}} \times \frac{p_{38}}{p_{38}+p_{14}} \times \frac{p_{14}}{p_{14}},$$

which is the product of the last column in Table 1. Hunter (2004) developed an MM algorithm to compute its MLE. The MM algorithm uses a simple but effective minorizing surrogate based on the convexity of the minus logarithm function. Caron and Doucet (2012) used the minimum order statistic of two exponential variables as the latent variable variable [c.f. Gordon (1983)] to construct an EM algorithm, which, by coincidence, has the same final form as the above MM algorithm.

Compared with EM and MM, which are algorithm templates, the weaver algorithm is stated in the final form, has relatively simple code, and is easy for a parallel implementation. We report in Table 3 that the serial weaver implementation finishes faster, according to the MATLAB profiler (MathWorks 2017), on the benchmark NASCAR2002 data set than the MM/EM algorithm.

In "Appendix D," we list the full PMF solution, along with a log-odds column for direct comparison with Hunter (2004)'s Table 2. Table 4 in "Appendix D" also reports a complete 87-driver ranking using the gained points (recorded in the raw data per driver per race) modeled by a Bradley–Terry likelihood,

$$L_{\mathrm{BT}}(\boldsymbol{p}) = \frac{p_{87}^{180} p_{19}^{170} \cdots p_{86}^{34}}{(p_{87}+p_{19}+\cdots+p_{86})^{180+170+\cdots+34}} \times \cdots \times \frac{p_{51}^{180} p_{42}^{180} \cdots p_{15}^{34}}{(p_{51}+p_{42}+\cdots+p_{15})^{180+180+\cdots+34}},$$

which has 123 terms. The two models' ranking results are generally similar.

Parallelization is critical for the weaver algorithm to perform in high dimensions with long $\boldsymbol{p}$ and large $\boldsymbol{\Delta}$. We report in Table 3 a speed comparison between a serial weaver implemented in C (Kernighan and Ritchie 1988) and a parallel weaver implemented in CUDA C (NVIDIA 2017). The elapsed times are measured for both implementations running on the same three large likelihood functions generated by the random partition algorithm of Sect. 4.3. The PMFs all have length $d = 1000$. The vector $\boldsymbol{b}$'s length ranges from 1000 to 100000.

Some additional references on relevant optimization methods are Heiser (1995), Tanner (1996), Lange and Zhou (2014), Nelder and Mead (1965), and Lagarias et al. (1998).

5.2 Eigenstructure and some algebraic considerations

In the multinomial model, the parameter–data–estimator relationship exhibits what might be called an eigenstructure:

$$\hat{\theta}(x) = \lambda x,$$

where $\hat{\theta}$ is the estimator, x is the data, and λ is a nonzero scalar. Such estimator characteristic brings a great computational benefit. In the incomplete multinomial model, many such eigenstructures with different λs are integrated into a single likelihood. Here, we have basically taken an inequality approach to unlocking the same computational benefit for the incomplete multinomial model and, through it, for all its sub-class models. Figure 9 illustrates the picture of such thought.

The variable cell pattern can influence the SEs of the point estimates through the observed information matrix formula derived in Sect. 2.2, where it shows the SEs of the point estimates depend on the $\boldsymbol{\Delta}$ matrix. For the point estimates, the complexity of the variable cell pattern has limited influence on the effectiveness of the estimating equations from our experience. This is probably because only very basic arithmetic operations are required in the estimating equations (3). In particular, they do not involve matrix inversion. The most complex operation in equations (3) is the matrix-vector multiplication.

The solution of the MLE equations (3) is not computationally limited to Algorithms 1 and 2. Given the simplicity of the equations, symbolic computation is also accessible, which has the advantage of arbitrary precision. The speed lost in symbolic manipulation can be regained from parallelization. Parallelizability is a main gift from equations (3). Some references on algebraic methods are Diaconis (1988), Pistone et al. (2000) and the text book Cox et al. (2007).

One can define an equivalence relation on the class of all *concave* incomplete multinomial likelihood functions via

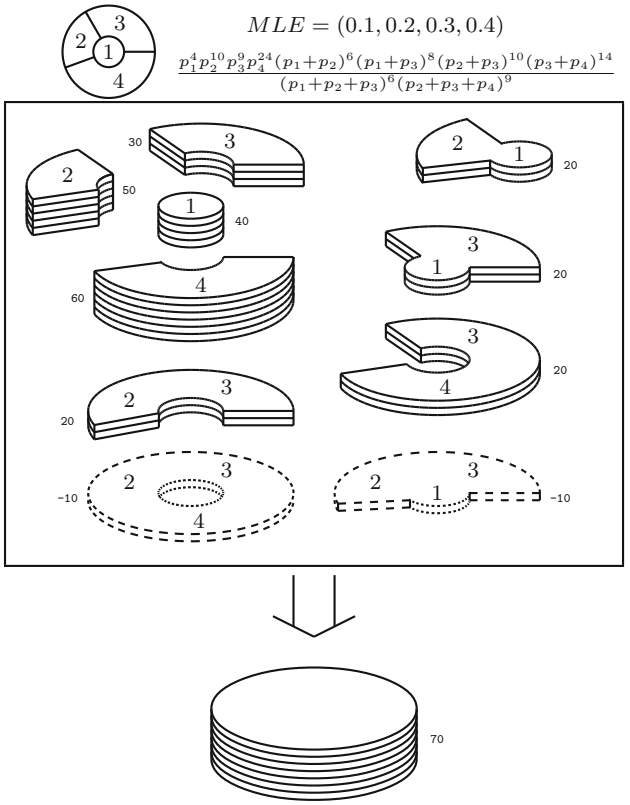


Fig. 9 *Picture of the eigenstructure.* Each piece represents a component in the likelihood. The volume and base area of each piece equals the (variable) cell count and the (variable) cell probability. Terms with negative exponents are drawn as dashed lines. The estimation process amounts to varying the base areas until all pieces can be assembled seamlessly into the big cylinder of volume 70 with a base area of 1

equality of the MLE. Within each equivalent class, the variability comes only from the partition process. Across the equivalent classes, variability lies in the multinomial sampling. The equivalence class is closed under multiplication, namely the product likelihood of two MLE-sharing and concave likelihoods has the same MLE and is still concave. In forming the product likelihood by natural experiments, each component is observed without being required to satisfy the MLE-sharing property, but they do usually satisfy concavity, even with conditional observations. This ensures the product likelihood to be concave.

Acknowledgements The authors are grateful to the two referees, Associate Editor, and Editor for their insightful comments that have significantly improved the article. Yin's research was supported in part by a grant (17326316) from the Research Grants Council of Hong Kong.

Appendix A: Proof of Lemma 1

Proof (Work with x_i/a_i and connect to the *weighted AM–GM inequality*, with its equality condition). Rewrite the target inequality as

$$\prod_{i=1}^n a_i^{a_i} \prod_{i=1}^n \left(\frac{x_i}{a_i}\right)^{a_i} \leqslant \frac{\prod_{i=1}^n a_i^{a_i}}{\left(\sum_{i=1}^n a_i\right)^{\sum_{i=1}^n a_i}} a_i^{\sum_{i=1}^n a_i} \left(\sum_{i=1}^n \frac{x_i}{a_i}\right)^{\sum_{i=1}^n a_i},$$

By substituting y_i for x_i/a_i and taking the $\left(\sum_{i=1}^n a_i\right)$-th root on both sides, we have

$$\prod_{i=1}^n y_i^{\frac{a_i}{\sum_{i=1}^n a_i}} \leqslant \sum_{i=1}^n \frac{a_i}{\sum_{i=1}^n a_i} y_i.$$

After a further substitution of $w_i = a_i / \sum_{i=1}^n a_i$, we arrive at

$$\prod_{i=1}^n y_i^{w_i} \leqslant \sum_{l=1}^n w_i y_i,$$

which is the *weighted AM-GM inequality*. It is crucial that we now check and confirm that all equalities can hold jointly if and only if $x_i/a_i = \tau$ for all i, given the existence of such a uniform constant τ which must be positive. □

Appendix B: Examples and Corollaries of Lemma 1

Example 5 $(x_1 + x_2)^5 \geqslant \frac{5^5}{3^3 2^2} x_1^3 x_2^2$. This inequality holds because

$$\begin{aligned} x_1^3 x_2^2 &= \frac{x_1}{3}\frac{x_1}{3}\frac{x_1}{3}\frac{x_2}{2}\frac{x_2}{2} 3^3 2^2 \\ &\leqslant 3^3 2^2 \left(\frac{3\frac{x_1}{3} + 2\frac{x_2}{2}}{3+2}\right)^{3+2} \\ &= 3^3 2^2 \left(\frac{x_1 + x_2}{5}\right)^5, \end{aligned}$$

where the equality is attained if and only if (x_1, x_2) is colinear with $(3, 2)$.

Example 6 $(x_1 + x_2)^7 x_3^3 x_4^5 \leqslant \frac{3^3 5^5 7^7}{15^{15}} (x_1 + x_2 + x_3 + x_4)^{15}$. This inequality holds because

$$(x_1 + x_2)^7 x_3^3 x_4^5 \leqslant 7^7 3^3 5^5 \left(\frac{7\frac{x_1+x_2}{7} + 3\frac{x_3}{3} + 5\frac{x_4}{5}}{7+3+5}\right)^{7+3+5},$$

where the equality is attained if and only if $(x_1 + x_2,\ x_3,\ x_4)$ is colinear with $(7, 3, 5)$. More importantly, together with the inequality in the previous example, the two equalities are

jointly attained if and only if $(x_1,\ x_2,\ x_3,\ x_4)$ is colinear with $(21, 14, 15, 25)$.

Corollary 1 *If we require* $\sum_{i=1}^n x_i = \sum_{i=1}^n a_i = 1$ *in Lemma 1, then*

$$\prod_{i=1}^n x_i^{a_i} \leqslant \prod_{i=1}^n a_i^{a_i},$$
$$\sum_{i=1}^n a_i \ln x_i \leqslant \sum_{i=1}^n a_i \ln a_i, \tag{18}$$

and the equalities are attained if and only if $x_i = a_i$ *for* $i = 1, \ldots, n$.

Corollary 2 *Let* $\boldsymbol{x} \in (0, +\infty)^n$ *be a vector of n positive reals. Let* $\boldsymbol{\delta} \in \{0, 1\}^n$ *be a vector of n bits. Let* $\boldsymbol{\beta} \in [0, +\infty)^n$ *be a nonzero vector of n nonnegative reals such that* $\beta_j = 0$ *if* $\delta_j = 0$. *Let* $b = \sum_{i=1}^n \beta_i > 0$. *Define* $0^0 = 1$. *Then*

$$\left(\boldsymbol{\delta}^\intercal \boldsymbol{x}\right)^b \geqslant \frac{b^b}{\prod_{i=1}^n \beta_i^{\beta_i}} \prod_{i=1}^n x_i^{\beta_i},$$

where the equality is attained if and only if there exists a positive k such that $x_i/\beta_i = k$ *for each of the i's having* $\delta_i = 1$.

Example 7 Let $n = 5, \boldsymbol{\delta} = (1, 0, 1, 0, 1)^\intercal, \boldsymbol{\beta} = (3, 0, 4, 0, 6)^\intercal$, $b = 3 + 0 + 4 + 0 + 6 = 13$. Then $\forall \boldsymbol{x} \in (0, +\infty)^n$, we have

$$(1x_1 + 0x_2 + 1x_3 + 0x_4 + 1x_5)^{13} \geqslant \frac{13^{13}}{3^3 0^0 4^4 0^0 6^6} x_1^3 x_2^0 x_3^4 x_4^0 x_5^6,$$

which attains the equality if and only if $x_1 : x_3 : x_5 = 3 : 4 : 6$.

Corollary 3 *If we rescale each* x_i *by an independent positive constant* c_i, *then we have the a seemingly more general but rather equivalent formulation of Lemma 1,*

$$\prod_{i=1}^n x_i^{a_i} \leqslant \frac{\prod_{i=1}^n a_i^{a_i}}{\prod_{i=1}^n c_i^{a_i} \left(\sum_{i=1}^n a_i\right)^{\sum_{i=1}^n a_i}} \left(\sum_{i=1}^n c_i x_i\right)^{\sum_{i=1}^n a_i},$$

which attains the equality if and only if there exists some positive constant k such that $c_i x_i / a_i = k$ *for all i.*

Example 8 Let $n = 3$, $a = (1, 2, 3)$, $c = (4, 5, 6)$, then we have

$$(4x_1)\,(5x_2)^2 (6x_3)^3 \leqslant \left(\frac{4x_1 + \frac{5x_2}{2} + \frac{5x_2}{2} + \frac{6x_3}{3} + \frac{6x_3}{3} + \frac{6x_3}{3}}{6}\right)^6.$$

Therefore,

$$x_1 x_2^2 x_3^3 \leqslant \frac{1}{4^1 5^2 6^3} \frac{1^1 2^2 3^3}{6^6} (4x_1 + 5x_2 + 6x_3)^6,$$

which attains equality if and only if $4x_1 = 5x_2/2 = 6x_3/3$ or $x_1 : x_2 : x_3 = 5 : 8 : 10$.

Corollary 4 *Generalizing Corollary 3 to a linear transform* $\boldsymbol{U}$ *on vector* $\boldsymbol{x}$,

$$\prod_{i=1}^n \left(\boldsymbol{u}_i^\intercal \boldsymbol{x}\right)^{a_i} \leq \left\{\prod_{i=1}^n \left(\frac{a_i}{\theta_i}\right)^{a_i}\right\} \left(\frac{\boldsymbol{\theta}^\intercal \boldsymbol{U} \boldsymbol{x}}{\sum_{i=1}^n a_i}\right)^{\sum_{i=1}^n a_i},$$

which attains the equality if and only if

$$\begin{bmatrix} \frac{\theta_1}{a_1} & & 0 \\ & \ddots & \\ 0 & & \frac{\theta_n}{a_n} \end{bmatrix} \boldsymbol{U} \boldsymbol{x} = k \mathbf{1}_n,$$

where k is a constant and can be solved explicitly under an extra constraint such as an affine constraint on $\boldsymbol{x}$.

Example 9 Let $x_1 = 2y_1 + y_2$ and $x_2 = y_1 + 2y_2$ in the first case of Example 5, we have

$$(2y_1 + y_2)^3 (y_1 + 2y_2)^2 \leq \frac{2^2 3^8}{5^5} (y_1 + y_2)^5,$$

which attains equality if and only if $y_1 = 4y_2$. By requiring the constraint $y_1 + y_2 = 1$ on the solution, it follows

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix},$$

and the unique maximum of $(2y_1 + y_2)^3 (y_1 + 2y_2)^2$ attained is $2^2 3^8 / 5^5 = 8.398$.

We recursively apply the inequality to the objective, as this inequality transforms the maximization problem into a set of equality attainment conditions, which becomes a system of simple equations.

Appendix C: Proof of the ascent property and the linear rate of convergence of the weaver algorithm when s is sufficiently large

We instead maximize the log-likelihood with a Lagrange multiplier term to incorporate the equality constraint,

$$\ell(\boldsymbol{p}) = \boldsymbol{a}^\intercal \ln \boldsymbol{p} + \boldsymbol{b}^\intercal \ln \boldsymbol{\Delta}^\intercal \boldsymbol{p} - s\left(\mathbf{1}^\intercal \boldsymbol{p} - 1\right),$$

where the Lagrange multiplier is the known constant

$$s = \mathbf{1}^\intercal \boldsymbol{a} + \mathbf{1}^\intercal \boldsymbol{b},$$

not adding an extra unknown.
The derivative of $\ell(\boldsymbol{p})$ with respect to p_i at iteration k is given by

$$\frac{\partial \ell(\boldsymbol{p})}{\partial p_i^{(k)}} = \frac{a_i}{p_i^{(k)}} + \sum_{j=1}^{q} \frac{\Delta_{ij} b_j}{\sum_{h=1}^{d} \Delta_{hj} p_h^{(k)}} - s.$$

Combining the weaver steps 1 and 2, $p_i^{(k)}$ is updated according to

$$p_i^{(k+1)} = \frac{a_i}{s - \sum_{j=1}^{q} \frac{\Delta_{ij} b_j}{\sum_{h=1}^{d} \Delta_{hj} p_h^{(k)}}}.$$

We seek to establish the positivity of the quantity

$$\begin{aligned}\left(p_i^{(k+1)} - p_i^{(k)}\right) \frac{\partial \ell(\boldsymbol{p})}{\partial p_i^{(k)}} &= \left\{ \frac{a_i}{p_i^{(k)}} + \sum_{j=1}^{q} \frac{\Delta_{ij} b_j}{\sum_{h=1}^{d} \Delta_{hj} p_h^{(k)}} - s \right\} \\ &\quad \times \left\{ \frac{a_i}{s - \sum_{j=1}^{q} \frac{\Delta_{ij} b_j}{\sum_{h=1}^{d} \Delta_{hj} p_h^{(k)}}} - p_i^{(k)} \right\} \\ &= \frac{\left(a_i - p_i^{(k)} v^{(k)}\right)^2}{p_i^{(k)} v^{(k)}},\end{aligned}$$

where

$$v^{(k)} \equiv s - \sum_{j=1}^{q} \frac{\Delta_{ij} b_j}{\sum_{h=1}^{d} \Delta_{hj} p_h^{(k)}}.$$

It is now clear the condition for the last quantity to be positive is $v^{(k)} > 0$. Then, under this condition, every step of the iteration increases $\ell(\boldsymbol{p})$. Since $\ell(\boldsymbol{p})$ is clearly bounded from above, the iteration converges.

Next, we show the rate of convergence is linear. We denote the ith component of the solution as $p_i^{(*)}$ and use the simpler symbol g to denote the derivative function $g(p_i) \equiv \frac{\partial \ell(\boldsymbol{p})}{\partial p_i}$, hence $g\left(p_i^{(*)}\right) = 0$. We assume $\ell(\boldsymbol{p})$ is locally concave at $\boldsymbol{p}^{(*)}$ and assume g to be Lipschitz continuous, viz. there exists a positive constant L such that, for all pairs of (p, q) in the domain, $|g(p) - g(q)| \le L|p - q|$. Then, we have

$$\begin{aligned}p_i^{(k+1)} - p_i^{(*)} &= \frac{a_i}{\frac{a_i}{p_i^{(k)}} - g\left(p_i^{(k)}\right)} - p_i^{(*)} \\ &= \frac{a_i p_i^{(k)}}{a_i - p_i^{(k)} g\left(p_i^{(k)}\right)} - p_i^{(*)} \\ &= \frac{a_i\left(p_i^{(k)} - p_i^{(*)}\right) + p_i^{(*)} p_i^{(k)} g\left(p_i^{(k)}\right)}{a_i - p_i^{(k)} g\left(p_i^{(k)}\right)},\end{aligned}$$

and further,

$$\begin{aligned}\left|\frac{p_i^{(k+1)} - p_i^{(*)}}{p_i^{(k)} - p_i^{(*)}}\right| &= \left|\frac{a_i + p_i^{(*)} p_i^{(k)} \frac{g\left(p_i^{(k)}\right)}{p_i^{(k)} - p_i^{(*)}}}{a_i - p_i^{(k)} g\left(p_i^{(k)}\right)}\right| \\ &= \left|\frac{a_i - p_i^{(k)} \frac{g\left(p_i^{(k)}\right)}{1 - p_i^{(k)} \Big/ p_i^{(*)}}}{a_i - p_i^{(k)} g\left(p_i^{(k)}\right)}\right|\end{aligned}$$

If $p_i^{(k)} < p_i^{(*)}$, then $g\left(p_i^{(k)}\right) > 0$ and $1 - p_i^{(k)} \Big/ p_i^{(*)} > 0$. Therefore,

$$\frac{g\left(p_i^{(k)}\right)}{1 - p_i^{(k)} \Big/ p_i^{(*)}} > g\left(p_i^{(k)}\right).$$

If $p_i^{(k)} > p_i^{(*)}$, then $g\left(p_i^{(k)}\right) < 0$. Therefore, $\frac{g\left(p_i^{(k)}\right)}{p_i^{(k)} - p_i^{(*)}} < 0$ and

$$a_i + p_i^{(*)} p_i^{(k)} \frac{g\left(p_i^{(k)}\right)}{p_i^{(k)} - p_i^{(*)}} < a_i < a_i - p_i^{(k)} g\left(p_i^{(k)}\right).$$

In both cases, the numerator is smaller than the denominator, hence $\left|\frac{p_i^{(k+1)} - p_i^{(*)}}{p_i^{(k)} - p_i^{(*)}}\right| < 1$ and the rate of convergence is linear.

Appendix D: Ranking results of the car racing data

See Table 4.

Table 4 NASCAR2002 car racing data: complete ranking results using the Placket–Luce and Bradley–Terry models

Driver name	PL rank	PL PMF est.	i	PL $\beta_i = \text{Log}(\hat{P}_i/\hat{P}_1)$	BT rank	BT PMF est.
P.J. Jones	1	0.186404564	58	4.15	1	0.023463204
Scott Pruett	2	0.109555541	68	3.62	2	0.021996754
Mike Bliss	3	0.027419058	54	2.23	13	0.017077210
Mark Martin	4	0.023488563	51	2.08	4	0.018477977
Rusty Wallace	5	0.023046199	66	2.06	10	0.017748481
Jimmie Johnson	6	0.020492992	37	1.94	7	0.017849368
Tony Stewart	7	0.018402700	82	1.83	3	0.018625428
Jeff Gordon	8	0.016795380	32	1.74	6	0.017876530
Sterling Marlin	9	0.016694860	72	1.73	8	0.017840921
Ricky Rudd	10	0.016142733	61	1.70	14	0.016774526
Jeff Burton	11	0.015354723	31	1.65	16	0.016526187
Kurt Busch	12	0.015311248	48	1.65	5	0.018008461
Matt Kenseth	13	0.014842682	52	1.62	11	0.017197478
Dale Jarrett	14	0.014637976	13	1.60	12	0.017131513
Robert Pressley	15	0.014616780	63	1.60	28	0.013494644
Tom Hubert	16	0.014178605	80	1.57	29	0.013344697
Dale Earnhardt, Jr.	17	0.013790301	12	1.54	15	0.016568870
Bill Elliott	18	0.013451800	2	1.52	17	0.016134277
Ryan Newman	19	0.013283018	67	1.51	9	0.017822206
Dave Blanev	20	0.012682449	14	1.46	23	0.014240692
Ricky Craven	21	0.012603025	60	1.45	20	0.015086597
Ron Fellows	22	0.012601384	64	1.45	38	0.012385872
Michael Waltrip	23	0.012331733	53	1.43	19	0.015462985
Jeff Green	24	0.011798589	33	1.39	22	0.014372622
Robby Gordon	25	0.011530124	62	1.36	25	0.014093240
Bobby Labonte	26	0.011485129	4	1.36	21	0.014783933
Ted Musgrave	27	0.011249832	76	1.34	36	0.012713267
Kyle Petty	28	0.010888701	49	1.31	27	0.013584921
Terry Labonte	29	0.010183684	77	1.24	31	0.013258976
Jamie McMurray	30	0.009939150	27	1.22	18	0.015764967
Johnny Benson, Jr.	31	0.009861159	42	1.21	24	0.014132291
Jimmy Spencer	32	0.009689244	38	1.19	32	0.013077287
Kevin Harvick	33	0.009446842	45	1.17	26	0.013980069
Kenny Wallace	34	0.009260041	44	1.15	37	0.012464773
Jeremy Mayfield	35	0.009205626	34	1.14	34	0.012839904
Bobby Hamilton	36	0.009034232	3	1.12	35	0.012768050
Greg Biffle	37	0.008382600	21	1.05	44	0.011395510
Elliott Sadler	38	0.008232034	18	1.03	30	0.013262857
Jim Inglebright	39	0.008050031	36	1.01	58	0.009430152
Lance Hooper	40	0.008010967	50	1.00	56	0.009833190
John Andretti	41	0.007624819	41	0.95	39	0.012265620
Steve Park	42	0.007284639	74	0.91	41	0.011757295
Mike Skinner	43	0.007252152	55	0.90	49	0.011198539
Ken Schrader	44	0.007217817	43	0.90	43	0.011462399
Jerry Nadeau	45	0.006968941	35	0.86	48	0.011220590
Hut Stricklin	46	0.006952570	25	0.86	46	0.011292480
Hank Parker, Jr.	47	0.006781568	22	0.83	59	0.009034514

Table 4 continued

Driver name	PL rank	PL PMF est.	i	PL $\beta_i = \text{Log}(\hat{P}_i/\hat{P}_1)$	BT rank	BT PMF est.
Chad Little	48	0.006500003	10	0.79	63	0.008861344
Buckshot Jones	49	0.006473392	7	0.79	50	0.011110529
Boris Said	50	0.006320236	5	0.76	40	0.011778176
Jack Sprague	51	0.006228830	26	0.75	64	0.008820770
Jason Leffler	52	0.006223202	28	0.75	61	0.008897284
Brett Bodine	53	0.006214035	6	0.75	55	0.009944951
Steve Grissom	54	0.006137621	73	0.73	51	0.010710452
Casey Atwood	55	0.006082382	9	0.73	52	0.010466079
Ward Burton	56	0.005885004	83	0.69	33	0.013045560
Todd Bodine	57	0.005796735	79	0.68	42	0.011606318
Rick Mast	58	0.005692005	59	0.66	60	0.008900801
Joe Nemechek	59	0.005677445	39	0.66	45	0.011363619
Tim Sauter	60	0.005489072	78	0.62	66	0.007899625
Hermie Sadler	61	0.005314426	23	0.59	57	0.009594072
Stacy Compton	62	0.005277739	71	0.58	54	0.010126959
Ron Hornaday, Jr.	63	0.005239631	65	0.58	65	0.008134088
Geoffrey Bodine	64	0.005156242	20	0.56	47	0.011260146
Mike Wallace	65	0.004944446	56	0.52	53	0.010336122
Derrike Cope	66	0.003883784	16	0.28	69	0.007161137
Dave Marcis	67	0.003022438	15	0.03	82	0.005147442
Austin Cameron	68	0.002945440	1	0.00	80	0.005207696
Shawna Robinson	69	0.002940897	70	0.00	68	0.007197775
Scott Wimmer	70	0.002777679	69	−0.06	62	0.008889281
Joe Varde	71	0.002547493	40	−0.15	78	0.005425866
Frank Kimmel	72	0.002206856	19	−0.29	70	0.006730990
Tony Raines	73	0.002191885	81	−0.30	67	0.007565824
Dick Trickle	74	0.002157488	17	−0.31	81	0.005150774
Carl Long	75	0.002139651	8	−0.32	87	0.003205604
Kirk Shelmerdine	76	0.002131900	47	−0.32	79	0.005332018
Christian Fittipaldi	77	0.001893864	11	−0.44	77	0.005491400
Morgan Shepherd	78	0.001877493	57	−0.45	74	0.005615740
Kevin Lepage	79	0.001853367	46	−0.46	71	0.006090557
Jay Sauter	80	0.001724303	30	−0.54	73	0.005933925
Jason Small	81	0.001722761	29	−0.54	76	0.005510013
Stuart Kirby	82	0.001496052	75	−0.68	72	0.005989078
Hideo Fukuyama	83	0.001375393	24	−0.76	75	0.005590129
Andy Hillenburg	#N/A	#N/A	#N/A	#N/A	84	0.004695920
Gary Bradberry	#N/A	#N/A	#N/A	#N/A	85	0.004683511
Jason Hedlesky	#N/A	#N/A	#N/A	#N/A	83	0.004735487
Randy Renfrow	#N/A	#N/A	#N/A	#N/A	86	0.004670611

References

Agresti, A.: Categorical Data Analysis, 2nd edn. Wiley, New York (2003)

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

Caron, F., Doucet, A.: Efficient Bayesian inference for generalized Bradley–Terry models. J. Comput. Graph. Stat. **21**(1), 174–196 (2012)

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

Connor, R.J., Mosimann, J.E.: Concepts of independence for proportions with a generalization of the Dirichlet distribution. J. Am. Stat. Assoc. **64**(325), 194–206 (1969)
Cox, D.A., Little, J., O'Shea, D.: Ideals, Varieties, and Algorithm: An Introduction to Computational Algebraic Geometry and Commutative Algebra, 3rd edn. Springer, New York (2007)
David, H.A.: The Method of Paired Comparisons, 2nd edn. Oxford University Press, Oxford (1988)
Davidson, R., Farquhar, P.: A bibliography on the method of paired comparisons. Biometrics **32**, 241–252 (1976)
Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodol.) **39**(1), 1–38 (1977)
Diaconis, P.: In: Gupta, S.S. (ed.) Group Representations in Probability and Statistics, Lecture Notes-Monograph Series, vol. 11. Institute of Mathematical Statistics Hayward, CA. https://projecteuclid.org/euclid.lnms/1215467407 (1988)
Dickey, J.M., Jiang, J.M., Kadane, J.B.: Bayesian methods for censored categorical data. J. Am. Stat. Assoc. **82**(399), 773–781 (1987)
Dwork, C., Kumar, R., Naor, M., Sivakumar, D.: Rank aggregation methods for the web. In: Proceedings of the 10th International Conference on World Wide Web, pp. 613–622. ACM (2001)
Ford, L.R.J.: Solution of a ranking problem from binary comparisons. Am. Math. Mon. **64**(8), 28–33 (1957)
Gordon, L.: Successive sampling in large finite populations. Ann. Stat. **11**(2), 702–706 (1983)
Gormley, I.C., Murphy, T.B.: Exploring voting blocs within the irish electorate: a mixture modeling approach. J. Am. Stat. Assoc. **103**(483), 1014–1027 (2008)
Guiver, J., Snelson, E.: Bayesian inference for Plackett-Luce ranking models. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 377–384. ACM, Pittsburgh (2009)
Haberman, S.J.: Product models for frequency tables involving indirect observation. Ann. Stat. **5**(6), 1124–1147 (1977)
Hankin, R.K.S.: A generalization of the Dirichlet distribution. J. Stat. Softw. **33**(11), 1–18 (2010)
Hartley, H.O., Hocking, R.R.: The analysis of incomplete data. Biometrics **27**(4), 783–823 (1971)
Hastie, T., Tibshirani, R.: Classification by pairwise coupling. Ann. Stat. **26**(2), 451–471 (1998)
Heiser, W.J.: Convergent computing by iterative majorization: theory and applications in multidimensional data analysis. In: Krzanowski, W.J. (ed.) Recent Advances in Descriptive Multivariate Analysis, pp. 157–189. Clarendon Press, Oxford (1995)
Huang, T.K., Weng, R.C., Lin, C.J.: Generalized Bradley–Terry models and multi-class probability estimates. J. Mach. Learn. Res. **7**, 85–115 (2006)
Hunter, D.R.: MM algorithms for generalized Bradley–Terry models. Ann. Stat. **32**(1), 384–406 (2004)
Hunter, D.R., Lange, K.: A tutorial on MM algorithms. Am. Stat. **58**(1), 30–37 (2004)
Jech, T.: The ranking of incomplete tournaments: a mathematician's guide to popular sports. Am. Math. Mon. **90**(4), 246–266 (1983)
Kernighan, B.W., Ritchie, D.M.: In: Ritchie, D.M. (ed.) The C Programming Language, 2nd edn. Prentice Hall Professional Technical Reference, Upper Saddle River (1988)
Lagarias, J., Reeds, J., Wright, M., Wright, P.: Convergence properties of the Nelder–Mead simplex method in low dimensions. SIAM J. Optim. **9**(1), 112–147 (1998)
Laird, N.: Nonparametric maximum likelihood estimation of a mixing distribution. J. Am. Stat. Assoc. **73**(364), 805–811 (1978)
Lange, K.: Optimization, 2nd edn. Springer, New York (2013)
Lange, K., Zhou, H.: MM algorithms for geometric and signomial programming. Math. Program. **143**(1–2), 339–356 (2014)
Lange, K., Hunter, D.R., Yang, I.: Optimization transfer using surrogate objective functions. J. Comput. Graph. Stat. **9**(1), 1–59 (2000)
Loève, M.: Probability Theory I, 4th edn. Springer, New York (1977)
Loève, M.: Probability Theory II, 4th edn. Springer, New York (1978)
Luce, R.D.: Individual Choice Behavior: A Theoretical Analysis. Wiley, New York (1959)
Luce, R.D.: The choice axiom after twenty years. J. Math. Psychol. **15**, 215–223 (1977)
Marden, J.I.: Analyzing and Modeling Rank Data. Chapman & Hall/CRC, Boca Raton (1996)
MathWorks: Matlab documentation. URL https://www.mathworks.com/help/matlab/ref/profile.html (2017)
McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley, New York (2008)
Nelder, J.A., Mead, R.: A simplex method for function minimization. Comput. J. **7**(4), 308–313 (1965)
Ng, K.W., Tian, G.L., Tang, M.L.: Dirichlet and Related Distributions: Theory, Methods and Applications. Wiley, New York (2011)
NVIDIA: CUDA Toolkit Documentation v8.0. URL http://docs.nvidia.com/cuda/index.html (2017)
Pistone, G., Riccomagno, E., Wynn, H.P.: Algebraic Statistics: Computational Commutative Algebra in Statistics. Chapman & Hall/CRC, Boca Raton (2000)
Plackett, R.L.: The analysis of permutations. Appl. Stat. **24**, 193–202 (1975)
Sattath, S., Tversky, A.: Unite and conquer: a multiplicative inequality for choice probabilities. Econometrica **44**(1), 79–89 (1976)
Suppes, P., Krantz, D.H., Luce, R.D., Tversky, A.: Foundations of Measurement: Geometrical, Threshold, and Probabilistic Representations. Academic Press, New York (1971)
Tanner, M.A.: Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. Springer, New York (1996)
Thurstone, L.L.: Psychophysical analysis. Am. J. Psychol. **38**(3), 368–389 (1927)
Turnbull, B.W.: The empirical distribution function with arbitrarily grouped, censored and truncated data. J. R. Stat. Soc. Ser. B (Methodol.) **38**(3), 290–295 (1976)
Tversky, A.: Elimination by aspects: a theory of choice. Psychol. Rev. **79**, 281–299 (1972)
Wu, C.F.J.: On the convergence properties of the EM algorithm. Ann. Stat. **11**(1), 95–103 (1983)
Yan, T., Yang, Y., Xu, J.: Sparse paired comparisons in the Bradley–Terry model. Statistica Sinica **22**(3), 1305–1318 (2012)
Zermelo, E.: Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. Mathematische Zeitschrift **29**(1), 436–460 (1929)