Holonomic Gradient Method for Two Way Contingency Tables

Tachibana, Yoshihito
Goto, Yoshiaki
Koyama, Tamio
Takayama, Nobuki

(Citation)
Algebraic Statistics,11(2):125-153

(Issue Date)
2020-12-28

(Resource Type)
journal article

(Version)
Version of Record

(URL)
https://hdl.handle.net/20.500.14094/90009308

11:2 2020

Algebraic Statistics

HOLONOMIC GRADIENT METHOD FOR TWO-WAY CONTINGENCY TABLES

YOSHIHITO TACHIBANA, YOSHIAKI GOTO, TAMIO KOYAMA AND NOBUKI TAKAYAMA

msp

Algebraic Statistics
Vol. 11, No. 2, 2020
https://doi.org/10.2140/astat.2020.11.125

HOLONOMIC GRADIENT METHOD FOR TWO-WAY CONTINGENCY TABLES

YOSHIHITO TACHIBANA, YOSHIAKI GOTO, TAMIO KOYAMA AND NOBUKI TAKAYAMA

The holonomic gradient method gives an algorithm to efficiently and accurately evaluate normalizing constants and their derivatives. We apply the holonomic gradient method in the case of the conditional Poisson or multinomial distribution on two-way contingency tables. We utilize the modular method in computer algebra or some other tricks for an efficient and exact evaluation, and we compare them and discuss on their implementation. We also discuss on a theoretical aspect of the distribution from the viewpoint of the conditional maximum likelihood estimation. We decompose parameters of interest and nuisance parameters in terms of sigma algebras for general two-way contingency tables with arbitrary zero cell patterns.

1. Introduction

The holonomic gradient method (HGM) proposed in [17] provides an algorithm to efficiently and accurately evaluate normalizing constants and their derivatives. This algorithm utilizes holonomic differential equations or holonomic difference equations. Y. Goto and K. Matsumoto [8] determined a system of difference equations for the hypergeometric system of type (k, n). The normalizing constant of the conditional Poisson or multinomial distribution on two-way contingency tables is a polynomial solution of this hypergeometric system. Thus, we can apply these difference equations to exactly evaluate the normalizing constant and its derivatives by HGM. However, there is a difficulty: numerical evaluation errors, incurred by repeatedly applying these difference equations or recurrence relations, increase rapidly if we use floating point number arithmetic. Accordingly, we evaluate the normalizing constant by exact rational arithmetic. However, in general, exact evaluation is slow. The modular method in computer algebra (see, e.g., [18], [25]) has been used for efficient and exact evaluation over the field of rational numbers. We apply the modular method or some other tricks to our evaluation procedure. We compare these methods and explore implementation of these algorithms in Sections 4 and 5.

We then turn from computation to a theoretical question before presenting statistical applications. An interesting application of the evaluation of the normalizing constant is the conditional maximum likelihood estimation (CMLE) of parameters of interest with fixed marginal sums. Broadly speaking, the parameters of interest in this case are (generalized) odds ratios. However, we could not identify a rigorous formulation on parameters of interest for contingency tables with zero cells in the literature. In Sections 7 and 8, we introduce $\mathcal{A}$-distributions as a conditional distribution. The conditional Poisson

MSC2010: 33C90, 65Q10, 62B05, 62H17.
Keywords: holonomic gradient method, two-way contingency tables, modular method, conditional maximum likelihood estimation.

or multinomial distribution on contingency tables with fixed marginal sums is a special and important case of $\mathcal{A}$-distributions. We will decompose parameters of interest and nuisance parameters in terms of σ-algebras. We note that the conditional distribution of a statistic given the occurrence of a sufficient statistic of a nuisance parameter does not depend on the value of the nuisance parameter. Hence, by the conditional distribution, we can estimate the parameter of interest without being affected by the nuisance parameter.

Finally, we apply our method to a CMLE problem for contingency tables. This problem is discussed in [20] for the case of $2 \times n$ contingency tables and the work presented here generalizes this to two-way contingency tables of any size and with any pattern of zero cells.

2. Two-way contingency tables

We introduce our notation for contingency tables and review how the normalizing constant for a conditional distribution is expressed by a hypergeometric polynomial of type (k, n). There are several salient references on contingency tables. Among them, we will refer to [1] and [10, Chap 4] herein.

2.1. $r_1 \times r_2$ *contingency table.*

Definition 1 ($r_1 \times r_2$ (two-way) contingency table). An $r_1 \times r_2$ matrix with nonnegative integer entries is called an $r_1 \times r_2$ *contingency table*. For a contingency table $u = (u_{ij})$, we define the *row sum vector* by $\beta^r = \left(\sum_j u_{1j}, \cdots, \sum_j u_{r_1 j}\right)^T$, and the *column sum vector* by $\beta^c = \left(\sum_i u_{i1}, \cdots, \sum_i u_{ir_2}\right)^T$. A contingency table u is also written as a column vector of length $r_1 \times r_2$, denoted by u^f. The column vector obtained by joining β^r and β^c is denoted by β, which is called the *row column sum vector* or the *marginal sum vector*.

Example 1 (2×3 contingency table and the row sum and the column sum). For the 2×3 contingency table $u = \left(\begin{smallmatrix} 5 & 3 & 6 \\ 7 & 2 & 4 \end{smallmatrix}\right)$ the row sum vector and the column sum vector are

$$\beta^r = \begin{pmatrix} 5+3+6=14 \\ 7+2+4=13 \end{pmatrix}, \quad \beta^c = \begin{pmatrix} 5+7=12 \\ 3+2=5 \\ 6+4=10 \end{pmatrix}.$$

The corresponding vector expressions of u^f and β are

$$u^f = \begin{pmatrix} 5 & 3 & 6 & 7 & 2 & 4 \end{pmatrix}^T, \quad \beta = \begin{pmatrix} 14 & 13 & 12 & 5 & 10 \end{pmatrix}^T.$$

We fix $p = (p_{ij}) \in \mathbb{R}_{>0}^{r_1 \times r_2}$, $N \in \mathbb{N}_0$ and consider the multinomial distribution

$$\frac{N! p^u}{u! |p|^N}, \quad p^u = \prod_{i,j} p_{ij}^{u_{ij}}, \quad u! = \prod_{i,j} u_{ij}!$$

on contingency tables satisfying $|u| = \sum_{i,j} u_{ij} = N$. The conditional distribution obtained by fixing the

row sum vector β^r and the column sum vector β^c is

$$\frac{p^u}{u!Z(\beta;\, p)}, \qquad Z(\beta;\, p) = \sum_{Au^f=\beta,\ u\in\mathbb{N}_0^{r_1\times r_2}} \frac{p^u}{u!}. \tag{1}$$

Here, the polynomial $Z(\beta;\, p)$ is the normalizing constant of this conditional distribution. The matrix A satisfies the following conditions: (1) entries are 0 or 1; (2) Au^f is the marginal sum vector (see Example 2). The expectation of the u-value at $(i,\, j)$ of this conditional distribution is equal to

$$E[U_{ij}] = p_{ij}\frac{\partial \log Z}{\partial p_{ij}}. \tag{2}$$

Exact evaluation of the conditional probability of getting a contingency table u and evaluation of the expectation is reduced to the evaluation of the normalizing constant Z and its derivatives. For given rational numbers p_{ij}, we provide an efficient and exact method to evaluate Z and its derivatives.

Example 2 (example of A). When $u^f = \begin{pmatrix} 5 & 3 & 6 & 7 & 2 & 4 \end{pmatrix}^T$, the matrix A is

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

and we have

$$Au^f = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ 3 \\ 6 \\ 7 \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 14 \\ 13 \\ 12 \\ 5 \\ 10 \end{pmatrix} = \beta.$$

Example 3. We consider 2×2 contingency tables with the marginal sum vector $\beta = \begin{pmatrix} 5 & 7 & 8 & 4 \end{pmatrix}^T$. All contingency tables u satisfying $Au^f = \beta$ are

$$\begin{pmatrix} 5 & 0 \\ 3 & 4 \end{pmatrix}, \begin{pmatrix} 4 & 1 \\ 4 & 3 \end{pmatrix}, \begin{pmatrix} 3 & 2 \\ 5 & 2 \end{pmatrix}, \begin{pmatrix} 2 & 3 \\ 6 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 4 \\ 7 & 0 \end{pmatrix}.$$

These u are written as

$$\begin{pmatrix} 5 & 0 \\ 3 & 4 \end{pmatrix} + i \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \quad (i = 0, 1, 2, 3, 4).$$

3. The normalizing constant of 2 × 2 tables

It is known that the normalizing constant for the conditional distribution for $r_1 \times r_2$ tables is A-hypergeometric polynomial (see, e.g., [10, Section 6.13]). We will illustrate this correspondence for 2×2 contingency tables.

Consider the marginal sum vector $\beta = (u_{11}, u_{21}+u_{22}, u_{11}+u_{21}, u_{22})$ with $u_{ij} \geq 0$. The 2×2 contingency tables with the marginal sum vector β are

$$u = \begin{pmatrix} u_{11} & 0 \\ u_{21} & u_{22} \end{pmatrix} + i \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \quad (i = 0, 1, 2, \cdots, n).$$

Here, we have $n = \min\{u_{11}, u_{22}\}$. The normalizing constant is

$$Z(\beta; p) = \sum_{i=0}^{n} \frac{p_{11}^{u_{11}-i} p_{12}^{i} p_{21}^{u_{21}+i} p_{22}^{u_{22}-i}}{(u_{11}-i)!\,(i)!\,(u_{21}+i)!\,(u_{22}-i)!} = \frac{p_{11}^{u_{11}} p_{21}^{u_{21}} p_{22}^{u_{22}}}{u_{11}!u_{21}!u_{22}!} \sum_{i=0}^{n} \frac{(-u_{11})_i (-u_{22})_i}{(u_{21}+1)_i (1)_i} \left(\frac{p_{12}p_{21}}{p_{11}p_{22}}\right)^i,$$

where $(a)_i = a(a+1)\cdots(a+i-1)$. Then, it can be expressed in terms of the Gauss hypergeometric function

$${}_2F_1(a, b, c; x) = \sum_{i=0}^{\infty} \frac{(a)_i (b)_i}{(c)_i (1)_i} x^i.$$

Note that when $a, b \in \mathbb{Z}_{\leq 0}$, it is a polynomial. The normalizing constant can also be expressed in terms of ${}_2F_1$ for other types of marginal sum vectors. A consequence of this observation is that we can utilize several formulae of the hypergeometric function to evaluate the normalizing constant.

4. Contiguity relation

In the previous section, we expressed the normalizing constant for 2×2 contingency tables with a fixed marginal sum vector in terms of the Gauss hypergeometric function. For $r_1 \times r_2$ contingency tables, the normalizing constant with a fixed marginal sum vector can be expressed in terms of the Aomoto–Gel'fand hypergeometric function of type (r_1, r_1+r_2) [29] (the function ${}_2F_1$ is of type $(2, 4)$). This hypergeometric function is also called the A-hypergeometric function for the product of the (r_1-1)-simplex and (r_2-1)-simplex. The difference holonomic gradient method for these hypergeometric functions utilizes contiguity relations. We illustrate this for the case of the Gauss hypergeometric function; for the general case, see [8].

Example 4 (the case of ${}_2F_1$). Put $f(a) = {}_2F_1(a, b, c; x)$ and

$$F(a) = \begin{pmatrix} f(a) \\ \theta_x f(a) \end{pmatrix}, \quad M(a) = \frac{1}{a-c+1} \begin{pmatrix} bx+a-c+1 & x-1 \\ -abx & a(1-x) \end{pmatrix},$$

where θ_x is the Euler operator $x\partial_x$. Then, we have

$$F(a) = M(a)F(a+1). \tag{3}$$

Now, note the following relations:

$$\frac{1}{a}(a+\theta_x) \bullet f(a) = f(a+1), \tag{4}$$

$$\big(\theta_x(c-1+\theta_x) - x(a+\theta_x)(b+\theta_x)\big) \bullet f(a) = 0. \tag{5}$$

The first relation can be shown from the series expansion and the second relation is the Gauss hypergeometric differential equation. It follows from (4), (5) that

$$\frac{1}{a}(a+\theta_x)\bullet F(a) = F(a+1), \quad \theta_x \bullet F(a) = \begin{pmatrix} 0 & 1 \\ \dfrac{abx}{1-x} & \dfrac{ax+bx-c+1}{1-x} \end{pmatrix} F(a) = A(a)F(a).$$

Next, we have (3) as

$$\frac{1}{a}(a+\theta_x)\bullet F(a) = \frac{1}{a}(aE + A(a))F(a), \quad F(a) = \Big(\frac{1}{a}(aE+A(a))\Big)^{-1} F(a+1) = M(a)F(a+1),$$

where E is the identity matrix.

A relation like $F(a) = M(a)F(a+1)$ is called a *contiguity relation*. In [8], the vector valued function $F(a)$ is called the *Gauss–Manin vector*.

There are several algorithms to obtain contiguity relations [28], [22], [21], [8]. Among them, we choose to use the method of twisted cohomology groups given in [8], because it is the most efficient method for the case of two-way contingency tables.

We briefly summarize the method given in [8]. Consider the hypergeometric series $f(\alpha; x)$ of type (r_1, r_1+r_2). Here, the parameter $\alpha = (\alpha_1, \ldots, \alpha_{r_1+r_2-1})$ stands for the marginal sum vector β and the variable $x = (x_{ij})_{1\le i\le r_1-1, 1\le j\le r_2-1}$ stands for p. It follows from the twisted cohomology group (a vector space spanned by equivalence classes of differential forms) associated to the integral representation of f that the contiguity relation for $\alpha_i \to \alpha_i + 1$ can be obtained as follows.

We consider the twisted cohomology group H (resp. H') standing for the function $f(\alpha; x)$ (resp. $f(\alpha; x)|_{\alpha_i\to\alpha_i+1}$). Both twisted cohomology groups are of dimension $r = \binom{r_1+r_2-2}{r_1-1}$. We take a basis $\varphi_1, \ldots, \varphi_r$ of H such that the "integral" of $(\varphi_1, \ldots, \varphi_r)^T$ gives a constant multiple of the Gauss–Manin vector

$$F(\alpha; x) = (f(\alpha; x), \delta^{(2)} \bullet f(\alpha; x), \ldots, \delta^{(r)} \bullet f(\alpha; x))^T,$$

where $\delta^{(i)}$ is some differential operator with respect to $x = (x_{ij})$. There exist a basis $\varphi'_1, \ldots, \varphi'_r$ of H' and a linear map $\mathcal{U}_i : H' \to H$ such that the integral of $\big(\mathcal{U}_i(\varphi'_1), \ldots, \mathcal{U}_i(\varphi'_r)\big)^T$ gives a constant multiple of the shifted Gauss–Manin vector $F(\alpha; x)|_{\alpha_i\to\alpha_i+1}$. Let $U_i(\alpha; x)$ be a representation matrix of $\mathcal{U}_i$ with respect to the bases $\{\varphi'_i\}$ and $\{\varphi_j\}$:

$$\big(\mathcal{U}_i(\varphi'_1), \ldots, \mathcal{U}_i(\varphi'_r)\big)^T = U_i(\alpha; x)\cdot(\varphi_1, \ldots, \varphi_r)^T.$$

Integrating both sides, we thus obtain the contiguity relation

$$F(\alpha; x)|_{\alpha_i\to\alpha_i+1} = \tilde{U}_i(\alpha; x)F(\alpha; x),$$

where $\tilde{U}_i$ is a constant multiple of U_i. It turns out that the representation matrix U_i can be expressed in terms of a simple diagonal matrix and base transformation matrices which can be obtained by evaluating intersection numbers among differential forms. The contiguity relation for $\alpha_i \to \alpha_i - 1$ can be derived analogously. For more details, see [8]. Here, we illustrate this method in the case of ${}_2F_1$.

Example 5 (the case of ${}_2F_1$ ($r_1 = r_2 = 2$, $r = 2$)). For the parameter (a, b, c) of ${}_2F_1$, we put

$$(\alpha_1, \alpha_2, \alpha_3) = (b, -a, c-b-1).$$

Here, we set $\alpha_0 = -\alpha_1 - \alpha_2 - \alpha_3 = a - c + 1$ for convenience. Since the move $a+1 \to a$ corresponds to $\alpha_2 - 1 \to \alpha_2$ (and $\alpha_0 + 1 \to \alpha_0$) in the new parametrization, the matrix $M(a)$ in Example 4 stands for $U_2(\alpha; x)$. The representation matrix U_2 has the following decomposition (see the Appendix) for more details):

$$U_2 = \frac{\alpha_1(\alpha_2-1)}{\alpha_3} \begin{pmatrix} \frac{1}{\alpha_0}+\frac{1}{\alpha_1} & \frac{1}{\alpha_0} \\ \frac{1}{\alpha_0} & \frac{1}{\alpha_0}+\frac{1}{\alpha_2} \end{pmatrix} \begin{pmatrix} \alpha_1 & -\alpha_1 \\ 0 & -\alpha_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1-x \end{pmatrix} \begin{pmatrix} \frac{1}{\alpha_0+1}+\frac{1}{\alpha_1} & \frac{1}{\alpha_0+1} \\ \frac{1}{\alpha_0+1} & \frac{1}{\alpha_0+1} \end{pmatrix} \begin{pmatrix} \frac{\alpha_1+\alpha_3}{\alpha_2-1} & 1 \\ 1 & \frac{\alpha_2-1+\alpha_3}{\alpha_1} \end{pmatrix}.$$

Apart from the diagonal matrix $\mathrm{diag}(1, 1-x)$, the matrices are expressed by intersection numbers. Since we have $\delta^{(2)} = \frac{1}{\alpha_2}\theta_x$, the matrix U_2 has a small difference with $M(a)$ in Example 4 and we obtain $M(a)$ by adjusting the scale factor $1/\alpha_2$ of θ_x.

By the contiguity relation, we can evaluate the normalizing constant Z and its derivatives. We explain the procedure for the case of ${}_2F_1$. Suppose $a \in \mathbb{Z}_{<-1}$. By the contiguity relation (3), we have

$$\begin{aligned} F(a) &= M(a)F(a+1) \\ &= M(a)M(a+1)F(a+2) \\ &\vdots \\ &= M(a)M(a+1)\cdots M(-2)F(-1). \end{aligned} \tag{6}$$

Then, we can obtain the value of $F(a)$ from the initial value $F(-1) = \left(1 - \frac{b}{c}x, -\frac{b}{c}x\right)^T$ by applying linear transformations. Values of the normalizing constant and its derivatives can be obtained from $F(a)$ with the differential equation for the Gauss hypergeometric function. This method is called the *difference holonomic gradient method* (difference HGM) and can be generalized to the case of $r_1 \times r_2$ contingency tables with the Gauss–Manin vector and contiguity relations given in [8].

We note that a naive evaluation of the polynomial Z is very slow. For example, the polynomial Z of the 2×5 contingency table with the row sum $(4n, 5n)$, the column sum $(5n, n, n, n, n)$ and $p = \left(\begin{smallmatrix} 1 & 1/2 & 1/3 & 1/5 & 1/7 \\ 1 & 1 & 1 & 1 & 1 \end{smallmatrix}\right)$ can be expressed in terms of the Lauricella function $F_D(-4n; -n, -n, -n, -n; n+1; 1/2, 1/3, 1/5, 1/7)$ of 4 variables (see, e.g., [6]). The number of terms is $O(n^4)$. Here is a comparison of the naive summation of F_D and our HGM implementation discussed in the next section.

n	20	30	40
Naive summation (in seconds)	16.0	111.7	456.6
HGM (in seconds)	0.28	0.276	0.284

Thus, the HGM is worth researching.

We briefly introduce an algorithm of difference HGM for $r_1 \times r_2$ contingency tables. The following algorithm computes the Gauss–Manin vector $F(\beta; p)$ which is essentially the same as $F(\alpha; x)$ in the above (for the correspondence between $(\beta; p)$ and $(\alpha; x)$, see [8, Proposition 7.1]). In fact, we give an improvement of Step 2–4 of [8, Algorithm 7.8].

Algorithm 1 (A modified version of [8, Step 1–4 of Algorithm 7.8]).

Input: $\beta = (\beta_1^{(1)}, \dots, \beta_{r_1}^{(1)}; \beta_1^{(2)}, \dots, \beta_{r_2}^{(2)})$: a marginal sum vector, $p = (p_{ij}) \in \mathbb{Q}_{>0}^{r_1 \times r_2}$: probabilities of the cells.

Output: the Gauss–Manin vector $F(\beta; p)$ (which is a vector of size $r = \binom{r_1+r_2-2}{r_1-1}$).

(1) Set $B_0 = (1, \dots, 1, \beta_1^{(1)} + \cdots + \beta_{r_1}^{(1)} - r_1 + 1; \beta_1^{(2)}, \dots, \beta_{r_2}^{(2)})$. Compute $F(B_0; p)$ by the definition. (In this case, the normalizing constant $Z(B_0; p)$ is a polynomial of small degree, and hence the Gauss–Manin vector $F(B_0; p)$ is easily computed.)

(2) For $k = 1, \dots, r_1 - 1$, define B_k inductively as $B_k = B_{k-1} + (\beta_k^{(1)} - 1) \cdot \delta_k$, where

$$\delta_k = (0, \dots, 0, \underset{k\text{-th}}{1}, 0, \dots, 0, -1; 0, \dots, 0)$$

(note that B_{r_1-1} is β). Evaluate the contiguity matrices $C_k(t)$ that satisfy

$$F(B_{k-1} + (T+1)\delta_k; p) = C_k(T) \cdot F(B_{k-1} + T\delta_k; p), \quad T = 0, 1, \dots, \beta_k^{(1)} - 2.$$

Here, t is an indeterminate and each entry of $C_k(t)$ is an element of $\mathbb{Q}(t)$.

(3) For $k = 1, \dots, r_1 - 1$, compute $F(B_k; p)$ inductively as

$$F(B_k; p) = C_k(\beta_k^{(1)} - 2) \cdots C_k(1) C_k(0) F(B_{k-1}; p). \tag{7}$$

(4) Return $F(B_{r_1-1}; p)$.

By using $F(\beta; p)$, we can compute the normalizing constant $Z(\beta; p)$ and the expectations $E[U_{ij}]$ (see [8, Step 5–7 of Algorithm 7.8]).

Example 6 (cf. [8, Example 7.10]). We consider 3×3 contingency tables whose marginal sum vector is $\beta = (2, 3, 3; 1, 3, 4)$. In this case, the Gauss–Manin vector is of size $\binom{3+3-2}{3-1} = 6$.

(1) We set $B_0 = (1, 1, 6; 1, 3, 4)$, and compute $F(B_0; p)$ by the definition. In this case, the normalizing constant $Z(B_0; p)$ has only eight terms.

(2) We set $B_1 = (2, 1, 5; 1, 3, 4)$, $B_2 = (2, 3, 3; 1, 3, 4) (= \beta)$. By using notations in [8], we put

$$C_1(t) = U_1^{-1}(-5+t, -2-t, -1, 3, 4, 1; x), \quad C_2(t) = U_2^{-1}(-4+t, -2, -2-t, 3, 4, 1; x).$$

Here, $x \in \mathbb{Q}^{(r_1-1)\times(r_2-1)}$ is defined from p. We have

$$C_1(0)F(1, 1, 6; 1, 3, 4; p) = F(2, 1, 5; 1, 3, 4; p),$$
$$C_2(0)F(2, 1, 5; 1, 3, 4; p) = F(2, 2, 4; 1, 3, 4; p), \quad C_2(1)F(2, 2, 4; 1, 3, 4; p) = F(2, 3, 3; 1, 3, 4; p).$$

(3) We compute the product

$$\begin{aligned} C_2(1)C_2(0)C_1(0)F(B_0; p) &= C_2(1)C_2(0)C_1(0)F(1, 1, 6; 1, 3, 4; p) \\ &= C_2(1)C_2(0)F(2, 1, 5; 1, 3, 4; p) \ (= C_2(1)C_2(0)F(B_1; p)) \\ &= C_2(1)F(2, 2, 4; 1, 3, 4; p) \\ &= C_2(1)F(2, 3, 3; 1, 3, 4; p) \ (= F(B_2; p)). \end{aligned}$$

(4) We obtain the Gauss–Manin vector $F(B_2;\, p) = F(\beta;\, p)$.

For example, when $p = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1 & 1/5 & 1/7 \\ 1 & 1 & 1 \end{pmatrix}$, the 6×6 matrix $C_2(t)$ is given as follows.[1]

$$C_2(t) = \begin{pmatrix} \frac{-(35t+29)}{35(t+2)} & \frac{12}{5(t+2)} & \frac{24}{7(t+2)} & \frac{-12}{5(t+2)} & \frac{-24}{7(t+2)} & 0 \\ \frac{1}{5} & -\frac{1}{5} & 0 & \frac{1}{5} & 0 & 0 \\ \frac{1}{7} & 0 & -\frac{1}{7} & 0 & \frac{1}{7} & 0 \\ \frac{-8}{5(t+2)} & \frac{8}{5(t+2)} & 0 & \frac{21t-73}{35(t+2)} & \frac{-88}{35(t+2)} & \frac{88}{35(t+2)} \\ \frac{-6}{7(t+2)} & 0 & \frac{6}{7(t+2)} & \frac{-33}{35(t+2)} & \frac{10t-47}{35(t+2)} & \frac{-33}{35(t+2)} \\ 0 & 0 & 0 & -\frac{1}{35} & \frac{1}{35} & -\frac{1}{35} \end{pmatrix}.$$

Remark 1. The algorithm given in [8] requires more matrix multiplications than Algorithm 1. As [8, Example 7.10], the former algorithm computes the above $F(2, 3, 3; 1, 3, 4;\, p)$ by nine matrix multiplications (each " $\mapsto$" means one multiplication):

$$F(1, 1, 2; 2, 1, 1;\, p) \mapsto F(1, 1, 3; 2, 2, 1;\, p) \mapsto F(1, 1, 4; 2, 3, 1;\, p)$$
$$\mapsto F(1, 1, 5; 2, 3, 2;\, p) \mapsto F(1, 1, 6; 2, 3, 3;\, p) \mapsto F(1, 1, 7; 2, 3, 4;\, p)$$
$$\mapsto F(1, 1, 6; 1, 3, 4;\, p) \mapsto F(2, 1, 5; 1, 3, 4;\, p) \mapsto F(2, 2, 4; 1, 3, 4;\, p) \mapsto F(2, 3, 3; 1, 3, 4;\, p).$$

On the other hand, Algorithm 1 needs only the last three steps.

We give the complexity to construct the matrix $C_k(t)$. The Appendix will help the reader follow the argument. By [8, Theorem 5.3], the matrix $U_k^{\pm 1}$ for the contiguity relation is the product of five matrices of size $r = \binom{r_1+r_2-2}{r_1-1} = \frac{(r_1+r_2-2)!}{(r_1-1)!\,(r_2-1)!}$:

(a) one diagonal matrix whose entries are rational functions in p,

(b) two intersection matrices whose entries are rational functions in β,

(c) two inverse matrices of intersection matrices

(cf. Example 5). For U_k^{-1}, by substituting

- $\beta_k^{(1)}$ and $\beta_{r_1}^{(1)}$ with certain polynomials in t of degree 1,
- the other $\beta_j^{(i)}$'s and p with certain rational numbers,

we obtain the matrix $C_k(t)$. By this construction and the formula for (a), (b), (c) in [8], it turns out that when we construct $C_k(t)$, we treat rational functions in t whose denominator and numerator are of degree at most 12. As long as we have tried on a computer for cases $5 \times r_i$, $r_i \leq 12$, the degrees of numerators

[1] This is obtained by our program `gtt_ekn3` as

```
gtt_ekn3.downAlpha3(2,2,2 | arule=gtt_ekn3.alphaRule_num([-5+t,-2,-1-t,3,4,1],2,2),
                            xrule=gtt_ekn3.xRule_num([[1,1/2,1/3],[1,1/5,1/7],[1,1,1]],2,2)).
```

and denominators are much smaller than 12 and no big number (large number so that FFT multiplication algorithms are used) appears in the matrix $C_k(t)$; when we use the modular method, all numbers in the matrix are elements in a finite field. Thus, we assume in the following theorem that the complexity of arithmetics of polynomials in one variable is $O(1)$.

Theorem 1. *Let $r_1, r_2 \geq 2$. Assume that the complexity of arithmetics is $O(1)$, the complexities of multiplying two $n \times n$ matrices and evaluating the determinant of an $n \times n$ matrix are $O(n^\omega)$ for some $2 \leq \omega < 3$. The complexity of obtaining the matrix $C_k(t)$ in Algorithm 1 for $r_1 \times r_2$ contingency tables is $O(r^\omega)$, where $r = \binom{r_1+r_2-2}{r_1-1}$. Especially, it is*

(1) *$O(r_2^{\omega r_1})$ when r_1 is fixed,*

(2) *$O(r_1^{\omega r_2})$ when r_2 is fixed,*

(3) *$O(2^{2\omega r_1})$ when $r_1 = r_2$.*

Proof. As explained later, the complexity to construct the above matrices (a), (b) and (c) are $O(r_1^\omega r)$, $O(r_1^2 r^2)$ and $O(r_1^2 r^2)$, respectively. Since the size of each matrix is r, the complexity of multiplication is $O(r^\omega)$. Thus, the complexity to obtain a contiguity relation is $O(r^\omega) + O(r_1^\omega r) + O(r_1^2 r^2)$. Since r is larger than r_1^2 in general, the complexity is equal to $O(r^\omega)$.

(1) We fix r_1 and assume $r_2 \gg r_1$. By the Stirling formula $\log n! \sim n \log n - n$, we have

$$\begin{aligned}\log r &\sim (r_1 + r_2)\log(r_1 + r_2) - r_2 \log r_2 \\ &= r_1 \log r_2 + r_1 \log\left(1 + \frac{r_1}{r_2}\right) + r_2 \log\left(1 + \frac{r_1}{r_2}\right) \sim r_1 \log r_2.\end{aligned}$$

Then we obtain $r \sim r_2^{r_1}$ and the complexity is $O(r_2^{\omega r_1})$.

(2) This can be obtained by a similar argument to Claim (1).

(3) If $r_1 = r_2$, then by the Stirling formula, we have

$$\log r \sim 2r_1 \log 2r_1 - 2r_1 \log r_1 = 2r_1 \log 2,$$

which implies $r \sim 2^{2r_1}$. Thus, the complexity is $O(2^{2\omega r_1})$.

Now, we explain the complexity of obtaining the matrices (a), (b), (c).

(a) As [8, Theorem 5.3], each nonzero entry of the diagonal matrix is the ratio of determinants of two $r_1 \times r_1$ matrices. Thus the complexity of evaluation is $O(r_1^\omega r)$.

(b) The entries of intersection matrices are intersection numbers of $(r_1 - 1)$-th twisted cohomology groups, which can be evaluated by the formula in [8, Fact 3.2]. The complexity of evaluating an intersection number by this formula is $O(r_1^2)$, and hence the complexity of obtaining the intersection matrix is $O(r_1^2 r^2)$.

(c) By the proof of [8, Proposition A.1], the inverse matrix of an intersection matrix is expressed as a product of two diagonal matrices and one intersection matrix. The complexity of obtaining the diagonal matrices is $O(r_1 r)$, since that of their nonzero entry is $O(r_1)$. Therefore, the complexity of

obtaining the inverse matrix of the intersection matrix is dominated by the complexity $O(r_1^2 r^2)$ of obtaining the intersection matrix. □

In this section we conducted a complexity analysis of the method for obtaining the contiguity relation. The theoretical complexity is of a polynomial order when r_i is fixed and our implementation shows that this step is efficient for small sized contingency tables. However, a naive evaluation of the composition of linear transformations (6) is slow, even for small contingency tables, because of large numbers when $|a|$ is large.

5. Efficient evaluation of a composition of linear transformations

To perform exact and efficient evaluations by the difference HGM, we need a fast and exact evaluation of a composition of linear transformations for vectors with rational number entries. This problem has hitherto been explored and there are several implementations, e.g., LINBOX [15]. For the purposes of empirical application, we study several methods to evaluate the composition of linear transformations such as (6) or (7). Our implementation is published as the package `gtt_ekn3` for Risa/Asir [24]. The function names in this section are those in this package.

5.1. ***Our benchmark problems.*** We use four benchmark problems to compare the various methods. The timing data are taken on a machine with

CPU	Intel(R) Xeon(R) CPU E5-4650 2.70 GHz
the number of CPU's	32
the number of cores	8
OS	Debian 9.8
memory	256 GB
software system	Risa/Asir (2018) version 20190328 with GMP [9]

Benchmark Problem 1. Evaluate

$$f = {}_2F_1\left(-36N, -11N, 2N; \frac{1-\frac{1}{N}}{56}\right), \ N \in \mathbb{N}.$$

It stands for the 2×2 contingency tables with the row sums $(36N, 13N-1)$ and the column sums $(38N-1, 11N)$. The parameter (p_{ij}) is set to $\begin{pmatrix} 1 & \frac{1-1/N}{56} \\ 1 & 1 \end{pmatrix}$.

Benchmark Problem 2. Evaluate the expectation for the 3×5 contingency tables with the row sums $(N, 2N, 12N)$, the column sums $(N, 2N, 3N, 4N, 5N)$, and the parameter p given by

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{5} & \frac{1}{7} \\ 1 & \frac{1}{11} & \frac{1}{13} & \frac{1}{17} & \frac{1}{19} \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Benchmark Problem 3. Evaluate the expectation for the 5×5 contingency tables with the row sums $(4N, 4N, 4N, 4N, 4N)$, the column sums $(2N, 3N, 5N, 5N, 5N)$, and the parameter p given by

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{5} & \frac{1}{7} \\ 1 & \frac{1}{11} & \frac{1}{13} & \frac{1}{17} & \frac{1}{19} \\ 1 & \frac{1}{23} & \frac{1}{29} & \frac{1}{31} & \frac{1}{37} \\ 1 & \frac{1}{37} & \frac{1}{41} & \frac{1}{43} & \frac{1}{47} \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Benchmark Problem 4. Evaluate the expectation for the 7×7 contingency tables with the row sums $(N, 2N, 3N, 4N, 5N, 6N, 7N)$, the column sums $(N, 2N, 3N, 4N, 5N, 6N, 7N)$, and the parameter p given by

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{5} & \frac{1}{7} & \frac{1}{11} & \frac{1}{13} \\ 1 & \frac{1}{17} & \frac{1}{19} & \frac{1}{23} & \frac{1}{29} & \frac{1}{31} & \frac{1}{37} \\ 1 & \frac{1}{41} & \frac{1}{43} & \frac{1}{47} & \frac{1}{53} & \frac{1}{59} & \frac{1}{61} \\ 1 & \frac{1}{67} & \frac{1}{71} & \frac{1}{73} & \frac{1}{79} & \frac{1}{83} & \frac{1}{89} \\ 1 & \frac{1}{97} & \frac{1}{101} & \frac{1}{103} & \frac{1}{107} & \frac{1}{109} & \frac{1}{113} \\ 1 & \frac{1}{127} & \frac{1}{131} & \frac{1}{137} & \frac{1}{139} & \frac{1}{149} & \frac{1}{151} \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

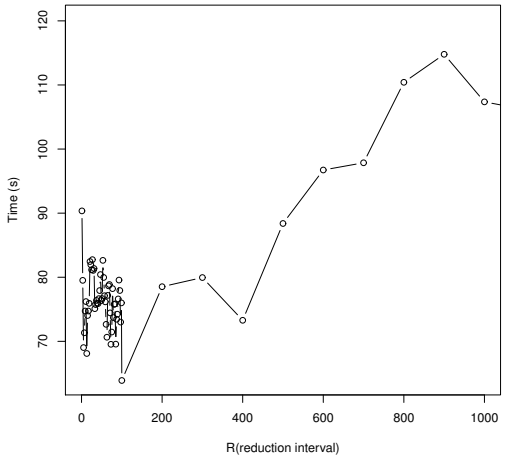
5.2. ***Floating point arithmetic.*** If we can evaluate the composition of linear transformations (7) accurately over floating point numbers, we can utilize GPU's or other hardware for efficient evaluation. Unfortunately, we lose the precision during the iteration of linear transformations in general. For example, let us evaluate the case of $N = 100$ for our 2×2 Benchmark Problem 1 with double arithmetic. The output by the double precision floating point arithmetic is `4.08315e+94`, but the answer is `4.48194745579962e+94` where we use the double value expression in the standard form, e.g., `4.08e+94` means 4.08×10^{94}. The output by double has only one digit of accuracy.

5.3. ***Intermediate swell of integers.*** We denote by $M(n)$ the complexity of the multiplication of two n-digits integers. The book [4] is a survey on algorithms and complexities on integer arithmetic.

Arithmetic over $\mathbb{Q}$ is more expensive than arithmetic over $\mathbb{Z}$, because the reduction of a rational number needs the computation of GCD of the numerator and the denominator. The best known complexity of the operation of GCD is $O(M(n) \log n)$ for two n-digits numbers (see, e.g., [16], [4]). The complexity of the Euclidean algorithm for GCD is $O(n^2)$.[2]

One way to avoid reductions in $\mathbb{Q}$ in our iterations of linear transformations (7) is to evaluate numerators and denominators separately and compute the GCD of the numerator and the denominator every R step of the linear transformations. We will call this sequential method `g_mat_fac_int` (generalized matrix factorial over integers). A reduction performing in every R step is necessary. In fact, our evaluation

[2]Timing data over $\mathbb{Q}$ in the version 1 of this paper at arxiv is very slow, because asir 2000 uses the Euclidean algorithm for the reductions in $\mathbb{Q}$ as default. The system asir 2018 based on GMP uses faster GCD algorithms as default.



R (reduction interval)	1	3	5	7	9	11	13	15
Time (s)	90.352	79.5147	69.024	71.335	74.7312	76.2025	68.1058	74.0283

Figure 1. Intermediate reduction.

problems make intermediate swell of integers by the method `g_mat_fac_int`. For example, the table below shows sizes of the numerators and the denominators by the separate evaluation without the intermediate reduction in our Benchmark Problem 1:

N	digits of num./den.	digits of num./den. after reduction	time
300	$1.97 \times 10^5 / 1.96 \times 10^5$	$3.35 \times 10^4 / 3.28 \times 10^4$	0.92s
500	$3.47 \times 10^5 / 3.47 \times 10^5$	$5.87 \times 10^4 / 5.76 \times 10^4$	1.56s

After the reduction, the numerators and the denominators become smaller as shown in the second column of the table.

We have no theoretical estimate for the best choice of R for intermediate reductions. Figure 1 shows timing data of our Benchmark Problem 2 with $N = 100$. The horizontal axis is the interval R of the intermediate reduction and the vertical axis is the timing. The graph indicates that we should choose R such that $5 \leq R \leq 100$.

5.4. ***Multimodular method.*** It may be standard to use the modular method when we have an intermediate swell of integers. We refer to, e.g., [11] and its references for the complexity analysis on modular methods.

Algorithm 2 (`g_mat_fac_itor` (generalized matrix factorial by itor), modular method). [3]

Input: $M(k)$ (matrix), F (vector), $S < E$ (indices), P_{list} (a list of prime numbers), C_{list} (a list of processes for a distributed computation).

Output: A candidate value of $M(E) \cdots M(S+2)M(S+1)M(S)F$ or "failure".

[3] We use "itor" as an abbreviation of the procedure `IntegerToRational`.

(1) Let F_n, F_d (scalar), M_n, M_d (scalar) be numerators and denominators of F and M respectively.

(2) For each prime number P_i in P_{list}, perform the linear transformations

$$\prod_{i=0}^{E-S} (M_n(S+i)M_d(S+i)^{-1})F_n F_d^{-1}$$

of F over $\mathbb{F}_{P_i}$. If the integer F_d or M_d is not invertible modulo P_i (unlucky case), then skip this prime number P_i and set P_{list} to $P_{\text{list}} \setminus \{P_i\}$. Let the output be G_i. This step may be distributed to processes in the C_{list}.

(3) Apply the Chinese remainder theorem to construct a vector G over $\mathbb{Z}/P\mathbb{Z}$ satisfying $G \equiv G_i \mod P_i$ where $P = \prod_{P_i \in P_{\text{list}}} P_i$.

(4) Return a candidate value by the procedure `IntegerToRational`(G, P) (rational reconstruction).

The complexity of the modular method `g_mat_fac_itor` is estimated as follows.

Theorem 2. *Let n be the number of the linear transformations and the size of the square matrix $r = \binom{r_1+r_2-2}{r_1-1}$. Suppose that each prime number P_i is d_p digits number and we use N_p prime numbers. C is the number of processes. The complexity of* `g_mat_fac_itor` *is approximated as*

$$\max\left\{O\left(\frac{nr^2N_pM(d_p)}{C}\right), O\left(r(d_pN_p)^2\right)\right\}$$

when n is in a bounded region where the rational reconstruction succeeds and the asymptotic complexity of the Chinese remainder theorem approximates well the corresponding exact complexity in the region.

Proof. We estimate the complexity of each step of `g_mat_fac_itor`.

(1) The complexity of one linear transformation is $O(r^2M(d_p))$. The linear transformation is performed n times for N_p prime numbers. Then the complexity is $O(nr^2N_pM(d_p))$ on a single process. This step can be distributed into C processes, then the complexity is $O(\frac{nr^2N_pM(d_p)}{C})$.

(2) The complexity to find an integer x such that $x \equiv x_i \mod p_i$ $(i = 1, \ldots, N_p)$ is discussed in [11, Theorem 6] under the assumption that an inborn FFT scheme is used. It follows from the estimate that the reconstruction complexity $C_n(N_p)$ of N_p primes of d_p digits is bounded by

$$\left(\tfrac{2}{3} + o(1)\right)M(d_pN_p)\max\left(\frac{\log N_p}{\log\log(d_pN_p)}, 1 + O(N_p^{-1})\right).$$

(3) The rational reconstruction algorithm `IntegerToRational`, see, e.g., [5], [19], is a variation of the Euclidean algorithm and its complexity is bounded by $O((N_pd_p)^2)$. We have r numbers to reconstruct.

Since the complexity of step (2) is smaller than other parts, we obtain the conclusion. □

The complexity is linear with respect to n (which is proportional to the size of the marginal sum vector in our benchmark problems) when the first argument of the "max" in the theorem is dominant. However, when n becomes larger, the rational reconstruction fails or gives a wrong answer. This is why we make

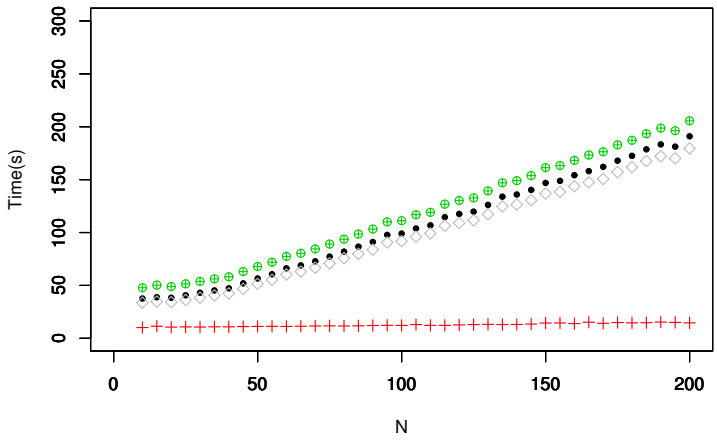


Figure 2. 5×5 contingency table, the Benchmark Problem 3 with 32 processes.

the assumption that n is in a bounded region. Note that the complexity estimate in the theorem is not an asymptotic complexity and is an approximate evaluation of it.

Let us present an example that this approximate evaluation works. Figure 2 is a graph of the timing data for the Benchmark Problem 3 with $N_p = 400$ and $d_p = 100$ by the decimal digits. The top point graph is the total time, the second top point graph is the time of the generalized matrix factorial (the execution time of Algorithm 2), the third point graph is the time of the distributed generalized matrix factorial by modulo P_i's (the step (2) of Algorithm 2). The last point graph is the time to obtain contiguity relations. Contiguity relations for several directions are obtained by distributing the procedures into 32 processes. Note that the point graph is linear with respect to N, which is proportional to the number of the linear transformations n. The timing data imply that the first argument of "max" of Theorem 2 is dominant in this case. In fact, when $N = 200$, the step for reconstructing rational numbers only takes about 8 seconds and linear transformations over finite fields take from 35 seconds to 52 seconds.

We should ask if our multimodular method is efficient on real computer environments. The following table is a comparison of timing data of the sequential method `g_mat_fac_int` (with a distributed computation of contiguity relations by 32 processors) and the multimodular method `g_mat_fac_itor` by 32 processors for the Benchmark Problem 3.

N	90	200
`g_mat_fac_int` with the reduction interval $R = 100$	21.57	45.40
`g_mat_fac_int` without the intermediate reduction	68.17	227.23
`g_mat_fac_itor` by 32 processors	103.23	205.57

Unfortunately, the multimodular method is slower than the sequential method `g_mat_fac_int` with a relevant choice of R on our best computer, however it is faster than the case of a bad choice of $R = \infty$.

When the size of the contingency table becomes larger, the rank r becomes larger rapidly. For example, $r = 20$ for the 5×5 contingency tables and $r = 924$ for the 7×7 contingency tables. Figure 3 shows timing data of our Benchmark Problem 4 of 7×7 contingency tables with the multimodular method by

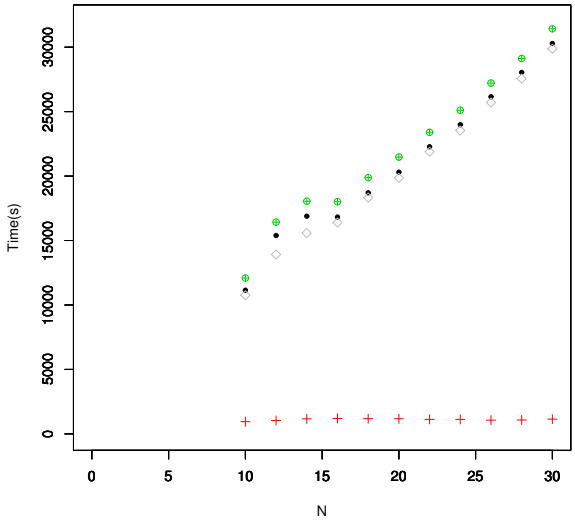


Figure 3. 7×7 contingency table, the Benchmark Problem 4 with 32 processes.

32 processors. We can also see linear timing with respect to N, but the slope is much larger than the 5×5 case as shown in our complexity analysis.

5.5. ***Binary splitting method.*** It is well-known that the binary splitting method for the evaluation of the factorial $m!$ of a natural number m is faster method than a naive evaluation of the factorial by $m! = m \times (m-1)!$. The binary splitting method evaluates $m(m-1)\cdots(\lfloor m/2 \rfloor + 1)$ and $\lfloor m/2 \rfloor (\lfloor m/2 \rfloor - 1)\cdots 1$ and obtains $m!$. This procedure can be recursively executed. This binary splitting can be easily generalized to our generalized matrix factorial; we may evaluate, for example, $M(a)M(a+1)\cdots M(\lfloor a/2 \rfloor - 1)$ and $M(\lfloor a/2 \rfloor)\cdots M(-2)$ to obtain $M(a)M(a+1)\cdots M(-2)$, $a < -2$ in (6). This procedure can be recursively applied. However, what we want to evaluate is the application of the matrix to the vector $F(-1)$. The matrix multiplication is slower than the linear transformation. Then, we cannot expect that this method is efficient for our problem when the size of the matrix is not small and the length of multiplication is not very long. However, there are cases that the binary splitting method is faster. Here is an output by our package `gtt_ekn3.rr`.

```
[1828] import("gtt_ekn3.rr")$
[4014] cputime(1)$
0sec(1.001e-05sec)
[4015] gtt_ekn3.expectation(Marginal=[[1950,2550,5295],[1350,1785,6660]],
                  P=[[17/100,1,10],[7/50,1,33/10],[1,1,1]]|bs=1)$ //binary splitting
3.192sec(3.19sec)
[4016] gtt_ekn3.expectation(Marginal,P)$
    4.156sec(4.157sec)
```

5.6. ***Benchmark of constructing contiguity relations.*** We gave a complexity analysis of finding contiguity relations. When r_1 is fixed, it is $O(r_2^{3r_1})$. The Figure 4 shows timing data to obtain contiguity

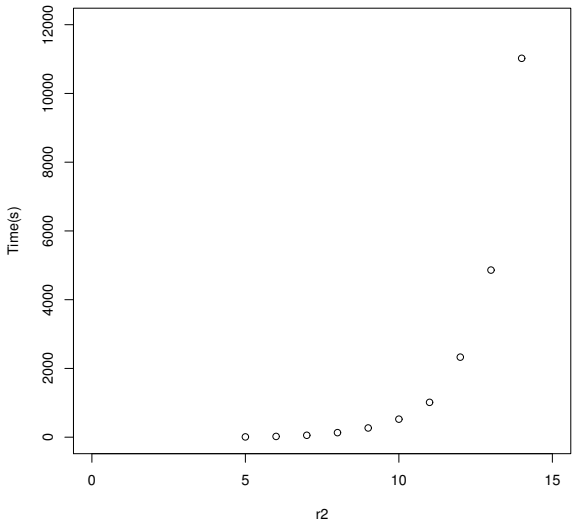


Figure 4. Time to obtain contiguity relations.

relations for $5 \times r_2$ contingency tables where the parameter p is

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1/p_1 & 1/p_2 & \cdots & 1/p_{r_2-1} \\ 1 & 1/p_{r_2} & 1/p_{r_2+1} & \cdots & 1/p_{2(r_2-1)} \\ 1 & \cdots & & & \\ 1 & 1/p_{(r_1-1)(r_2-1)+1} & \cdots & & \end{pmatrix}$$

(p_i is the i-th prime number), the row sum vector is $(a_1, 400, 400, 400, 400)$, and the column sum vector is $(200, 300, 500, 500, \ldots, 500)$. As is shown by our complexity analysis, when r_2 becomes larger, it rapidly becomes harder to obtain contiguity relations.

6. Zero cells

The contiguity relations derived by [8] are valid only when there are no zero cells in the contingency table. If there is a zero ($p_{ij} = 0$ and $u_{ij} = 0$) in the contingency table, a denominator of the contiguity relation is zero in general and therefore we cannot use their identity. One method to avoid this difficulty is interpolation. Note that the normalizing constant Z is a rational function in p_{ij} and the expectation $E[U_{ij}] = p_{ij} \frac{\partial \log Z}{\partial p_{ij}}$ is also a rational function. Because it is a rational function, we can obtain the exact value by evaluating it on a sufficient number of rational p_{ij}'s.

Proposition 1. *Let β be the marginal sum vector and L a generic line in p-space. If we evaluate $E[U_{ij}]$ at $2\beta_1$ points $p \in \mathbb{R}_{>0}^{r_1 \times r_2}$ on a line L, then the exact value of $E[U_{ij}]$ can be obtained at any point on L.*

Proof. When we restrict $E[U_{ij}]$ to the line L, it is a rational function in one variable. The degree of the denominator and the numerator is β_1 at most. Apply an interpolation algorithm by rational function, e.g., Stoer–Bulirsch algorithm [27], [23]. Then, we can obtain the exact value by interpolation. □

Example 7. Let the marginal sums and the parameter p (cell probability) be

$$\begin{array}{ccc|c} * & * & * & 3 \\ * & * & * & 4 \\ * & * & * & 3 \\ \hline 3 & 4 & 3 & \end{array}, \quad p = \begin{pmatrix} 1 & 1/2 & 0 \\ 1 & 1/3 & 1/4 \\ 1 & 1 & 1 \end{pmatrix}$$

Then, we can evaluate the expectation matrix ($E[U_{ij}]$) by the difference HGM and interpolation. Below is an output of our package `gtt_ekn3`. Here the `randinit` parameter specifies an interval of random nonzero p_{ij}'s where (i, j)'s are positions of zero cells.

```
[5150] import("gtt_ekn3.rr");
0
[5151] E=gtt_ekn3.cBasistoE_0(0,[[3,4,3],[3,4,3]],[[1,1/2,0],[1,1/3,1/4],[1,1,1]] | randinit=20);
[ 71076/56575 98649/56575 0 ]
[ 157581/113150 28069/22630 77337/56575 ]
[ 39717/113150 114957/113150 92388/56575 ]
// Expectation (exact value)
[5153] number_eval(E);  // Expectation (approximate value)
[ 1.25631462660186 1.74368537339814 0 ]
[ 1.39267344233319 1.2403446752099 1.36698188245692 ]
[ 0.351011931064958 1.01596995139196 1.63301811754308 ]
```

Although the interpolation method is applicable to any pattern of zero cells, a more efficient method involves utilizing hypergeometric functions restricted on some $p_{ij} = 0$'s. In general, contiguity relations and Pfaffian systems for such hypergeometric functions become complicated. In [7], a method is put forward to evaluate intersection numbers and contiguity relations when only one p_{ij} is zero.

7. Sufficient statistics as σ-algebra

Often we decompose parameters for contingency tables into row and column probabilities and odds ratios. When only odds ratios are the parameters of interest, CMLE is an appropriate method to estimate those odds ratios. However, this decomposition is no longer elementary when contingency tables contain zero cells. To facilitate a mathematically clear discussion of CMLE in the next section, we offer a formulation of parameters of interest, nuisance parameters, and sufficient statistics. Theorems 3, 4, and 5 explain what sufficient statistics are for the two-way contingency tables admitting zero cells. In order to prove these theorems, we utilize the notion of sufficient σ-algebra.

Classical formulations of sufficient statistics as σ-algebras appear in, e.g., [3], [14]. Our formulation is different because we treat parameters as random variables instead of considering a family of probability measures. This Bayesian statistical approach enables us to consider σ-algebras on parameter spaces. We express nuisance parameters and parameters of interest as sub-σ-algebras of the σ-algebra generated by all parameters. A Bayesian approach to sufficient statistics is presented in, e.g., Chapter 2 of the textbook by M. Schervish [26]. This book studies sufficient statistics by conditional probabilities given parameter valued random variables. We study them by a more general approach of conditional expectations given σ-algebras. The technical details are lengthy and, in this section and the next, we state only fundamental

notions and theorems which we need to study two-way contingency tables. Proofs for them are given in the preprint of this paper at arxiv (1803.04170). A general framework of the theory will be given in [13].

The treatment of nuisance parameters and parameters of interest is an important issue in statistics. The distinction between those parameters which are salient and of interest versus those which are not, may seem easy. However, it seems to be only a matter of declaring that μ is a parameter of interest or ν is a nuisance parameter. As we will see in the next section, when a group acts on parameter spaces and the group is regarded as the space of nuisance parameters, the distinction between them is not trivial. From a geometric perspective, the cause of this difficulty is that determining whether a parameter is "of interest" or a "nuisance" depends on a coordinate system. To formulate the "of interest" notion independently of a specific coordinate system, we will consider σ-algebras on parameter spaces. In probability theory and stochastic processes, σ-algebra is important as a natural way to express information (see, e.g. [12]). Discussions in this section are based on conditional expectations with respect to σ-algebra. For basic properties of conditional expectation, see [30].

Let Θ be a set. The set Θ stands for the parameter spaces. Let $\mathcal{B}(\Theta)$ be a σ-algebra on Θ, then $(\Theta, \mathcal{B}(\Theta))$ is a measure space. In the case where Θ is a topological space, we assume that $\mathcal{B}(\Theta)$ is the Borel algebra on Θ.

In standard parameter estimation, we assume a probability space $(\Omega', \mathcal{F}', \mathbf{P}'_c)$ with a parameter $c \in \Theta$. Let us define our probability space from the standard setting. Suppose $(\Theta, \mathcal{B}(\Theta), \mu)$ is a probability space. Put $\Omega := \Omega' \times \Theta$. Let $\mathcal{F}$ be the σ-algebra on Ω generated by

$$A \times B := \{(\omega, c) \in \Omega \mid \omega \in A,\ c \in B\} \quad (A \in \mathcal{F}',\ B \in \mathcal{B}(\Theta)).$$

The measurable space $(\Omega, \mathcal{F})$ is deemed to be the product measurable space of $(\Omega', \mathcal{F}')$ and $(\Theta, \mathcal{B}(\Theta))$ [30, p75]. For $A \in \mathcal{F}'$, let $f_A : \Theta \to \mathbb{R}$ be the function defined by $f_A(c) := \int_A \mathbf{P}'_c(d\omega)$ $(c \in \Theta)$. If f_A is $\mathcal{B}(\Theta)$-measurable for any $A \in \mathcal{F}'$, we can define a measure $\mathbf{P}$ on $\mathcal{F}$ by $\mathbf{P}(A \times B) := \int_B f_A(c)\mu(dc)$ $(A \in \mathcal{F}',\ B \in \mathcal{B}(\Theta))$. Thus, our probability space is defined as the product space under the measurable condition of f_A.

Let θ be a measurable map from Ω to Θ defined by

$$\theta : \Omega \ni (\omega', c) \mapsto c \in \Theta.$$

This implies that parameters can be regarded as a Θ-valued random variable. Although random variables are usually denoted by capital letters, we use lower case letters to denote random variables that are regarded as parameters.

Example 8. Let $(\Omega', \mathcal{F}', \mathbf{P}'_c)$ be the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), N(\mu, \sigma^2))$, where $N(\mu, \sigma^2)$ is the Gaussian distribution on $\mathbb{R}$ with mean μ and variance σ^2. In this case, the parameter space is

$$\Theta = \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid \sigma^2 > 0\}$$

and the parameter θ as a measurable map is defined by

$$\theta : \Omega \ni (x, (\mu, \sigma^2)) \mapsto (\mu, \sigma^2) \in \Theta.$$

We restart from a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, which is not necessarily a product space. For a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$, we use $\mathcal{L}^1(\mathcal{G})$ to denote the linear space of random variables which are integrable and $\mathcal{G}$-measurable. When two elements X and Y of $\mathcal{L}^1(\mathcal{G})$ satisfy $X(\omega) = Y(\omega)$ for all $\omega \in \Omega$, we state that X and Y are equal and denote $X = Y$. Note that $X = Y$ almost surely does not imply that $X = Y$. Let ϑ be the sub-σ-algebra of $\mathcal{F}$ generated by a random variable θ. It represents the information of θ. We formulate notions of nuisance parameters, sufficient parameters, and parameters of interest as sub-σ-algebras of ϑ.

For a pair of random variables X and Y, Y is $\sigma(X)$-measurable if and only if Y equals to $f(X)$ for a Borel measurable function f. See, e.g., [30, p206].

Let X and Y be $\mathbb{R}$-valued random variables and θ be a Θ-valued random variable, which we will call a parameter. We assume that X is integrable. The conditional expectation $\mathbf{E}(X|Y, \theta)$ can be regarded as a function of (Y, θ), i.e., we can take a Borel measurable function f from $\mathbf{R} \times \Theta$ to $\mathbf{R}$ such that

$$f(Y, \theta) = \mathbf{E}(X|Y, \theta) \quad \text{a.s.}$$

Because the equation $f(y, c_1) = f(y, c_2)$ may hold even if $c_1 \neq c_2$, the conditional expectation $\mathbf{E}(X|Y, \theta)$ is measurable with respect to a sub-σ-algebra strictly smaller than $\sigma(Y, \theta)$. This suggests that taking conditional expectation can reduce the information of θ.

Let us express this loss of information of θ in terms of σ-algebra. Let $\mathcal{D}$ and $\mathcal{G}$ be sub-σ-algebras of $\mathcal{F}$. In some applications, such as Theorem 3 discussed later, it is assumed that $\mathcal{D}$ is the sub-σ-algebra generated by all observable statistics and $\mathcal{G}$ is a sub-σ-algebra generated by a fraction of the observable statistics and a fraction of the parameters. Note that $\mathcal{G}$ may include some information of parameters. For $X \in \mathcal{L}^1(\mathcal{D})$, the conditional expectation $\mathbf{E}(X|\mathcal{G})$ can be measurable for a sub-σ-algebra which is strictly smaller than $\mathcal{G}$.

Definition 2. A sub-σ-algebra $\mathcal{I}$ is said to be *of interest* with respect to a pair of sub-σ-algebras $(\mathcal{D}, \mathcal{G})$ if, for all $X \in \mathcal{L}^1(\mathcal{D})$, there exists a version of $\mathbf{E}(X|\mathcal{G})$ which is $\mathcal{I}$-measurable.

Notions of nuisance and sufficiency describe a special case of such information loss.

Definition 3. Let $\mathcal{D}$, $\mathcal{S}$ and $\mathcal{N}$ be sub-σ-algebras of $\mathcal{F}$. When $\mathcal{S}$ is of interest with respect to $(\mathcal{D}, \sigma(\mathcal{S}, \mathcal{N}))$, we deem that $\mathcal{S}$ is sufficient for $(\mathcal{D}, \mathcal{N})$ or that $\mathcal{N}$ is nuisance for $(\mathcal{D}, \mathcal{S})$.

Remark 2. Note that the condition of Definition 3 is equivalent to stating that the equation

$$\mathbf{E}(X|\sigma(\mathcal{S}, \mathcal{N})) = \mathbf{E}(X|\mathcal{S}) \quad \text{a.s.} \tag{8}$$

holds for any $X \in \mathcal{L}^1(\mathcal{D})$. In fact, we have

$$\begin{aligned} \mathbf{E}(X|\sigma(\mathcal{S}, \mathcal{N})) &= \mathbf{E}(\mathbf{E}(X|\sigma(\mathcal{S}, \mathcal{N}))|\mathcal{S}) && (\mathbf{E}(X|\sigma(\mathcal{S}, \mathcal{N})) \in \mathcal{L}^1(\mathcal{S})) \\ &= \mathbf{E}(X|\mathcal{S}) && (\text{tower property}). \end{aligned}$$

Remark 3. In statistics, a statistic T is sufficient with respect to a parameter θ if the conditional distribution of observed data X given the statistic $T = t$ does not depend on the parameter θ. This condition is

formally expressed as

$$p(x|t,\theta) = p(x|t).$$

In similar tests and the Neyman–Scott problem, θ is denoted as a nuisance parameter or an uninteresting parameter [2]. We express this condition in terms of measure theory in Definition 3. In our definition, we use σ-algebra instead of statistics and parameters. Traditional definitions can be reduced to ours by

$$\mathcal{D} = \sigma(X), \quad \mathcal{S} = \sigma(T), \quad \mathcal{N} = \sigma(\theta).$$

Intuitively, $\mathcal{D}$, $\mathcal{S}$, and $\mathcal{N}$ denote the information of the observed data, the sufficient statistics, and the nuisance parameters, respectively.

In addition, we utilize conditional expectations instead of conditional probabilities because the latter can only be defined for a limited class of probability space and conditions.

Fundamental theorems on sufficient statistics can be generalized in our formulation on the sufficient sigma field [13].

Example 9. For random variables $X_1, \ldots, X_n, \theta$, suppose that

(1) $0 \le \theta \le 1$, and

(2) the conditional probability of $X_1, \ldots, X_n$ for given θ is

$$\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \quad (x_i \in \{0,1\})$$

Then, putting $\mathcal{D} := \sigma(X_1, \ldots, X_n)$, $\mathcal{N} := \sigma(\theta)$, $\mathcal{S} := \sigma(X_1 + \cdots + X_n)$, $\mathcal{S}$ is sufficient for $(\mathcal{D}, \mathcal{N})$.

In order to clarify our formulation by the σ-algebra, we will prove that $\mathcal{S}$ is sufficient. For $x = (x_1, \ldots, x_n)^\top \in \mathbf{R}^n$, we denote by $|x|$ the sum of elements of x. Put $X := (X_1, \ldots, X_n)^\top$ and $T := |X| = X_1 + \cdots + X_n$. By [30, p206], for any $Y \in \mathcal{L}^1(\mathcal{D})$, we can take a Borel measurable function $f : \mathbf{R}^d \to \mathbf{R}$ such that $Y = f(X)$. Let $g : \{0, 1, \ldots, n\} \to \mathbf{R}$ be a function defined by

$$g(t) := \binom{n}{t}^{-1} \sum_{x \in \{0,1\}^n} \delta_{t,|x|} f(x).$$

Then, $g(T)$ is $\mathcal{S}$-measurable. For any $B, C \in \mathcal{B}(\mathbf{R})$, we have (with I_B and I_C the indicator functions of B and C)

$$\begin{aligned}
\mathbf{E}(Y; T \in B, \theta \in C) &= \mathbf{E}(Y I_B(T) I_C(\theta)) = \mathbf{E}(f(X) I_B(|X|) I_C(\theta)) \\
&= \int \sum_{x \in \{0,1\}^n} f(x) I_B(|x|) I_C(\theta) \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} p(\theta)\, d\theta \\
&= \int \sum_{x \in \{0,1\}^n} \sum_{t=0}^{n} \delta_{t,|x|} f(x) I_B(|x|) I_C(\theta) \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} p(\theta)\, d\theta \\
&= \int \sum_{x \in \{0,1\}^n} \sum_{t=0}^{n} \delta_{t,|x|} f(x) I_B(t) I_C(\theta) \theta^t (1-\theta)^{n-t} p(\theta)\, d\theta
\end{aligned}$$

$$
\begin{aligned}
&= \int \sum_{t=0}^{n} \binom{n}{t}^{-1} \sum_{x\in\{0,1\}^n} \delta_{t,|x|} f(x) I_B(t) I_C(\theta) \binom{n}{t} \theta^t (1-\theta)^{n-t} p(\theta)\, d\theta \\
&= \mathbf{E}\left(g(T) I_B(t) I_C(\theta)\right) \\
&= \mathbf{E}\left(g(T);\, T \in B, \theta \in C\right).
\end{aligned}
$$

Since $\sigma(\mathcal{S}, \mathcal{N})$ is generated by $\{T \in B\} \cap \{\theta \in C\}$ $(B, C \in \mathcal{B}(\mathbf{R}))$, by [30, 1.6. Lemma (a)], we have $\mathbf{E}(Y; A) = \mathbf{E}(g(T); A)$ for any $A \in \sigma(\mathcal{S}, \mathcal{N})$. Consequently, $g(T)$ is a version of $\mathbf{E}(Y|\sigma(\mathcal{S}, \mathcal{N}))$ and $\mathcal{S}$-measurable. Hence, $\mathcal{S}$ is sufficient for $(\mathcal{D}, \mathcal{N})$.

To describe a sub-σ-algebra of interest in our application to the $\mathcal{A}$-distribution, we consider orbits of some group action. Suppose that a group G acts on a measurable space (S, Σ). For $B \subset S$ and $g \in G$, we put

$$
g \cdot B := \{g \cdot b \mid b \in B\}, \qquad\qquad G \cdot B := \{g \cdot b \mid g \in G,\ b \in B\}.
$$

Note that $G \cdot B = B$ holds if and only if $g \cdot B = B$ for any $g \in G$.

8. Application to the conditional MLE problem

In this section, we discuss a conditional MLE problem for $\mathcal{A}$-distributions.

Let A be an integer matrix of size $d \times n$, and b be an integer vector of length d. Suppose that Poisson random variables $X_k \sim \mathrm{Pois}(c_k)$, $(k = 1, \ldots, n)$ are mutually independent. We denote the conditional distribution of the random vector $X := (X_1, \ldots, X_n)^\top$ given $AX = b$ as an $\mathcal{A}$-distribution. The parameters of the $\mathcal{A}$-distribution are $c = (c_1, \ldots, c_n)^\top$ and $b = (b_1, \ldots, b_n)^\top$. The probability mass function of the $\mathcal{A}$-distribution is given as

$$
\mathbf{P}(X = x \mid AX = b, \theta = c) = \frac{\prod_{j=1}^{n} \frac{c_j^{x_j}}{x_j!} \exp(-c_j)}{\sum_{Ay=b} \prod_{j=1}^{n} \frac{c_j^{y_j}}{y_j!} \exp(-c_j)} = \frac{\prod_{j=1}^{n} \frac{c_j^{x_j}}{x_j!}}{\sum_{Ay=b} \prod_{j=1}^{n} \frac{c_j^{y_j}}{y_j!}}.
$$

An application of conditional distributions in statistics is the elimination of nuisance parameters. By Definition 3 and Remark 3, the conditional distribution of a statistic given the occurrence of a sufficient statistic of a nuisance parameter does not depend on the value of the nuisance parameter. This is an important property in similar tests and the Neyman–Scott problems (see, e.g., [2] and [10]). Hence, by the conditional distribution, we can estimate the parameter of interest without being affected by the nuisance parameter. From this perspective, we can regard the $\mathcal{A}$-distribution as the conditional distribution given the sufficient statistic AX, and the nuisance parameter corresponding to AX is $A\theta$. The traditional definition does not offer a mathematically clear description of the parameter of interest for this case. This is the motivation for the discussions in the previous section. The space of parameters of interest is naturally described as a sub-σ-algebra under less restrictive conditions on θ and c.

The parameter c of the $\mathcal{A}$-distribution moves on the set $\Theta := \mathbb{R}^n_{\geq 0}$. Consider the action of the multiplicative group $G := \mathbb{R}^d_{>0}$ on the space Θ defined as

$$g \cdot c = \left(c_j \prod_{i=1}^{d} g_i^{a_{ij}} \right)_{j=1,\ldots,n} \quad (g \in G,\ c \in \Theta).$$

This group action on Θ induces a group action on $\mathbb{Z}^d_{\geq 0} \times \Theta$ by

$$g \cdot (b, c) = (b, g \cdot c) \quad (g \in G,\ (b, c) \in \mathbb{Z}^d_{\geq 0} \times \Theta).$$

Theorem 3. *The sub-σ-algebra*

$$\mathcal{O} := \{\{(AX, \theta) \in B\} \mid B \in \mathcal{B}(\mathbb{Z}^d_{\geq 0}) \times \mathcal{B}(\Theta),\ G \cdot B = B\}$$

is of interest with respect to $(\sigma(X), \sigma(AX, \theta))$.

Note that the quotient space Θ/G by the group action G is not a manifold. Therein lies the difficulty in describing the space of parameters of interest and hence why we utilized the notion of σ-algebra of interest.

For a vector $v = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n$, we use $J(v)$ to denote the set of subscript j that satisfies $v_j \neq 0$. We also use $|J(v)|$ to denote the number of elements in $J(v)$, and we put $J(v)^c := \{j \in \mathbb{N} \mid j \notin J(v)\}$.

For $\alpha = (\alpha_1, \ldots, \alpha_n)^\top \in \mathbb{R}^n$, let R_α be the function from $\Theta = \mathbb{R}^n_{\geq 0}$ to $\mathbb{R}$ defined by

$$R_\alpha(c) := \begin{cases} \prod_{j \in J(\alpha)} c_j^{\alpha_j} & (c_j \neq 0 \text{ for all } j \in J(\alpha)) \\ 0 & (c_j = 0 \text{ for some } j \in J(\alpha)) \end{cases} \qquad (c = (c_1, \ldots, c_n)^\top \in \Theta).$$

Let $Z : \Theta \to \mathbb{R}^n$ be the function defined by $Z(c) := (Z_1(c), \ldots, Z_n(c))^\top$ $(c \in \Theta)$ where

$$Z_j(c) := \begin{cases} 1 & (c_j > 0) \\ 0 & (c_j = 0). \end{cases}$$

Theorem 4. *Let $\hat{\theta} : \Omega \to \mathbb{Z}^d_{\geq 0} \times \Theta$ be the measurable function defined by $\hat{\theta}(\omega) = (AX(\omega), \theta(\omega))$. If $\hat{\theta}$ is surjective, then*

$$\mathcal{O} = \sigma\big(AX,\ R_\alpha(\theta),\ Z(\theta); \alpha \in \ker A\big). \tag{9}$$

This theorem implies that the sub-σ-algebra of interest $\mathcal{O}$ stands for generalized odds ratios, which are, intuitively, parameters of interest. Note that the parameter may lie on the border θ_i.

As an interesting and important case of $\mathcal{A}$-distributions, we consider the $r_1 \times r_2$ contingency table. Let u_{ij} be independent Poisson random variables with parameter $\theta_{ij} \geq 0$ $(1 \leq i \leq r_1,\ 1 \leq j \leq r_2)$. The parameter $\theta := (\theta_{ij})$ lies on the set $\Theta := \mathbb{R}^{r_1 \times r_2}_{\geq 0}$. As in the previous section, we regard θ as a measurable function from $(\Omega, \mathcal{F})$ to $(\Theta, \mathcal{B}(\Theta))$. Note that we can assume that θ is surjective without loss of generality. Let $\mathcal{D}$ be the sub-σ-algebra generated by all u_{ij}, and $\mathcal{G}$ be the sub-σ-algebra generated by

$$\theta_{ij}\ (1 \leq i \leq r_1,\ 1 \leq j \leq r_2), \quad \sum_{i=1}^{r_1} u_{ij}\ (1 \leq j \leq r_2), \quad \sum_{j=1}^{r_2} u_{ij}\ (1 \leq i \leq r_1).$$

For all $X \in \mathcal{L}^1(\mathcal{D})$, the conditional expectation $\mathbf{E}(X|\mathcal{G})$ is invariant under the action of the multiplicative group $G := \mathbb{R}_{>0}^{r_1+r_2}$ on Θ defined by

$$g \cdot c := (g_i g_{r_1+j} c_{ij}) \quad (g = (g_i) \in G,\ c = (c_{ij}) \in \Theta).$$

For $1 \le i, k \le r_1$ and $1 \le j, \ell \le r_2$, let $R_{ijk\ell} : \Theta \to \mathbb{R}$ be a function defined by

$$R_{ijk\ell}(c) := \begin{cases} \frac{c_{ij}c_{k\ell}}{c_{i\ell}c_{kj}} & (c_{ij}c_{k\ell}c_{i\ell}c_{kj} \neq 0) \\ 0 & (c_{ij}c_{k\ell}c_{i\ell}c_{kj} = 0) \end{cases} \quad (c = (c_{ij}) \in \Theta).$$

Note that $R_{ijk\ell}$ is a function obtained from the odds ratio. For $1 \le i \le r_1$ and $1 \le j \le r_2$, we define a function $Z_{ij} : \Theta \to \mathbb{R}$ by

$$Z_{ij}(c) := \begin{cases} 1 & (c_{ij} > 0) \\ 0 & (c_{ij} = 0) \end{cases} \quad (c = (c_{ij}) \in \Theta).$$

The functions Z_{ij} $(1 \le i \le r_1,\ 1 \le j \le r_2)$ hold information on the position of zero cells. The functions $R_{ijk\ell}$ and Z_{ij} are invariant with respect to the action of group G.

The following theorem states that $A\theta$ is a nuisance parameter.

Theorem 5. $\sigma(AX, \theta) = \sigma(A\theta, \mathcal{O})$.

Corollary 1. *$\sigma(A\theta)$ is nuisance for $(\sigma(X), \mathcal{O})$.*

Proof. By Theorem 3, for any $Y \in \mathcal{L}^1(\sigma(X))$, $\mathbf{E}(Y|\sigma(AX, \theta))$ is $\mathcal{O}$-measurable. The equation in Theorem 5 implies that $\mathbf{E}(Y|\sigma(AX, \theta)) = \mathbf{E}(Y|\sigma(A\theta, \mathcal{O}))$. Hence, $\mathcal{O}$ is of interest with respect to $(\sigma(X), \sigma(A\theta, \mathcal{O}))$. Therefore $\sigma(A\theta)$ is nuisance for $(\sigma(X), \mathcal{O})$. □

9. Examples of CMLE problems

In the first part of this paper, we propose some efficient methods to evaluate the normalizing constant of the conditional distribution of fixed row and column sums for solving CMLE problems. In the second part, we clarify a statistical meaning of considering the conditional distribution. When the independence of rows and columns (the null model) is rejected under a test, it will be natural to estimate parameters of interest under the alternative hypothesis based on CMLE we have discussed. More precisely, Theorem 4 and 5 claim that when AX is given, $\sigma(R_\alpha(\theta), Z(\theta))$ are of interest and $\sigma(A\theta)$ is a nuisance. In the case of contingency tables, generalized odds ratios $R_\alpha(p)$ and positions of zero cells $Z(p)$ are of interest and row and column probabilities Ap are a nuisance when the marginal sums of the table are given. We present examples of estimating generalized odds ratios by CMLE.

Example 10. We generate categorical data concerning the number of hours slept and time of going to bed from a student sample in the LearnBayes package[4] of the system R for statistical computing.

Rows are categorized by time spent sleeping. The categories are sleeping less than 6 hours, 6–7 hours, and more than 7 hours. Columns are categorized by the time of going to bed. The categories are going to

[4]https://cran.r-project.org/web/packages/LearnBayes/index.html

bed before midnight, between midnight and 1am, and after 1am. We wish to analyze these categorical data by the Poisson random model $U_{ij} \sim \mathrm{Pois}(p_{ij})$. The independence of rows and columns is rejected by the χ^2 test with the threshold p-value 0.05. Then, we regard the column sum $\sum_i p_{ij}$ and the row sum $\sum_j p_{ij}$ as nuisance parameters. These represent probabilities of the event standing for j-th row and one standing for i-th column when the rows and the columns are independent. We perform CMLE under the condition that column sums $\sum_i u_{ij}$ and row sums $\sum_j u_{ij}$ are given.

Categorical data for all:

Bed time \ Hours slept	less than 6 hour	6–7	more than 7 hours
Before 24	1	6	123
24–25	3	22	145
After 25	86	91	176

We omit titles and express this table as $\begin{pmatrix} 1 & 6 & 123 \\ 3 & 22 & 145 \\ 86 & 91 & 176 \end{pmatrix}$. Categorical data for males:

$$\begin{pmatrix} 1 & 2 & 28 \\ 0 & 4 & 47 \\ 35 & 32 & 71 \end{pmatrix}$$

Categorical data for females:

$$\begin{pmatrix} 0 & 4 & 95 \\ 3 & 18 & 98 \\ 51 & 59 & 105 \end{pmatrix}$$

Because this CMLE can be solved by the $\mathcal{A}$-distribution discussed previously, we apply our algorithm for evaluating normalizing constants and their derivatives to the method for estimating the conditional maximum likelihood in [29, §4]. We obtain the following estimates. CMLE (p_{ij}) for all:

$$\begin{pmatrix} 0.176556059977815 & 1 & 10.5634953362788 \\ 0.144532927997885 & 1 & 3.39969669537228 \\ 1 & 1 & 1 \end{pmatrix}$$

CMLE for males:

$$\begin{pmatrix} 0.458167657900967 & 1 & \underline{6.25676090279981} \\ 0 & 1 & \underline{5.25200491199345} \\ 1 & 1 & 1 \end{pmatrix}$$

CMLE for females:

$$\begin{pmatrix} 0 & 1 & \underline{13.2714773737657} \\ 0.193351042187373 & 1 & \underline{3.04872586155291} \\ 1 & 1 & 1 \end{pmatrix}$$

As explained in the previous section, the space of parameters of interest should be regarded as the collection of different orbits by the torus action. When the parameter value obtained via CMLE is (p_{ij}),

values on the orbit $(g_i h_j p_{ij})$, $g_i, h_j \in \mathbb{R}_{>0}$ are equivalent parameters. Since the normalized elements of the second column and the third row are 1, we have $g_3 h_1 = g_3 h_2 = g_3 h_3 = 1$ and $g_1 h_2 = g_2 h_2 = g_3 h_2 = 1$. Then, we have $g_i h_j = 1$ for all i, j. The condition whereby this normalization is possible ($p_{i2} \neq 0$, $p_{3j} \neq 0$) defines a subspace of the parameters of interest. The subspace is isomorphic to $\mathbb{R}^4_{\geq 0}$ by the quotient topology. The correspondence is given by

$$(p_{ij}) \mapsto \begin{pmatrix} \frac{p_{11}p_{32}}{p_{12}p_{31}} & 1 & \frac{p_{13}p_{32}}{p_{12}p_{33}} \\ \frac{p_{21}p_{32}}{p_{22}p_{31}} & 1 & \frac{p_{23}p_{32}}{p_{22}p_{33}} \\ 1 & 1 & 1 \end{pmatrix} \tag{10}$$

In this chart, males and females exhibit different tendencies. For example, the underlined values at $(1, 3)$ and $(2, 3)$ positions are close in the case of males but not for females.

The number obtained by replacing p_{ij} by the frequency u_{ij} in (10) is called a generalized odds ratio. Generalized odds ratios for our data are as follows. Odds ratios for all:

$$\begin{pmatrix} 0.176356589147287 & 1 & 10.5994318181818 \\ 0.144291754756871 & 1 & 3.40779958677686 \\ 1 & 1 & 1 \end{pmatrix}$$

Odds ratios for males:

$$\begin{pmatrix} 0.457142857142857 & 1 & 6.30985915492958 \\ 0 & 1 & 5.29577464788732 \\ 1 & 1 & 1 \end{pmatrix}$$

Odds ratios for females:

$$\begin{pmatrix} 0 & 1 & 13.3452380952381 \\ 0.19281045751634 & 1 & 3.05925925925926 \\ 1 & 1 & 1 \end{pmatrix}$$

Note that, as proved in [29, Theorem 5], these generalized odds ratios approximate CMLE because we have a sufficient sample size.

When the sample size is relatively small, a generalized odds ratio may not approximate the corresponding CMLE well. We present one example.

Example 11. The categorical data below are taken from emergency safety information on diclofenac sodium for influenza encephalitis and encephalopathy.[5]
Categorical data:

	acetaminophen	diclofenac sodium	mefenamic acid
death	4	7	2
survival	32	5	6

[5]Pharmaceuticals and Medical Devices Agency, Japan, 2000, https://www.pmda.go.jp/files/000148557.pdf

We omit titles and express this table as $\left(\begin{smallmatrix} 4 & 7 & 2 \\ 32 & 5 & 6 \end{smallmatrix}\right)$. By applying our algorithm and the method in [29], we obtain the following CMLE.

$$\begin{pmatrix} 1 & \underline{10.5557279737263} & 2.62096714359908 \\ 1 & 1 & 1 \end{pmatrix}$$

Generalized odds ratios are

$$\begin{pmatrix} 1 & \underline{11.2} & 2.66666666666667 \\ 1 & 1 & 1 \end{pmatrix}$$

See the numbers underlined above. We observe that the odds ratio is larger than the CMLE. In other words, the effect of nuisance parameters increases the risk in this case. Finally, we briefly note how subsequent data released from the same institute in 2001 appeared to show that diclofenac sodium was in fact more associated with survival, rather than death. This reminds us of some of the difficulties inherent in statistical analyses. Here are those new data:[6]

	acetaminophen	diclofenac sodium	mefenamic acid
death	23	13	6
survival	78	25	9

Our algorithm outputs CMLE

$$\begin{pmatrix} 1 & 1.7567483756645 & 2.24788463785377 \\ 1 & 1 & 1 \end{pmatrix}$$

and odds ratios:

$$\begin{pmatrix} 1 & 1.76347826086957 & 2.26086956521739 \\ 1 & 1 & 1 \end{pmatrix}.$$

Appendix

We will explain the derivation of the matrix U_2 of Example 5 with twisted cohomology groups by following [8] and the program `gtt_ekn3/ekn_pfaffian_8.rr` of the package `gtt_ekn3`.

We start with the integral representation of ${}_2F_1$:

$$\frac{\Gamma(b)\Gamma(c-b)}{\Gamma(c)} \cdot {}_2F_1(a,b,c;x) = \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-xt)^{-a}dt = (-1)^b \int_0^{-1} t^b(1+xt)^{-a}(1+t)^{c-b-1}\frac{dt}{t}.$$

We replace the parameters a, b, c by

$$(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (a-c+1, b, -a, c-b-1),$$

where $\alpha_0 = -\alpha_1 - \alpha_2 - \alpha_3$ stands for the exponent at infinity. The decrement of a stands for an increment of α_2 (and decrement of α_0). The identity we want to derive is $F(a) = M(a)F(a+1)$, which is a special case of

$$\mathbf{S}(\alpha; x) = \frac{1}{\alpha_2} U_2(\alpha_{(2)}; x)\mathbf{S}(\alpha_{(2)}; x), \quad \alpha_{(2)} := (\alpha_0+1, \alpha_1, \alpha_2-1, \alpha_3)$$

[6] http://idsc.nih.go.jp/disease/influenza/iencepha.html

in [8, Corollary 6.3] ($\alpha_{(2)}$ stands for $a+1$). The function `upAlpha(2,1,1)` in the program derives $\frac{1}{\alpha_2}U_2$. $\mathbf{S}(\alpha; x)$ is the vector consisting of the hypergeometric series $S(\alpha; x)$ defined in [8, Section 6] and its derivatives (Gauss–Manin vector). When $c \in \mathbb{N}_0$, it can be expressed in terms of ${}_2F_1$ as

$$\mathbf{S}(\alpha; x) = \begin{pmatrix} S \\ \frac{1}{\alpha_2}\theta_x S \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1/\alpha_2 \end{pmatrix} \begin{pmatrix} S \\ \theta_x S \end{pmatrix} = \frac{1}{(-a)!\,(-b)!\,(c-1)!} \begin{pmatrix} 1 & 0 \\ 0 & 1/\alpha_2 \end{pmatrix} \begin{pmatrix} {}_2F_1 \\ \theta_x {}_2F_1 \end{pmatrix}.$$

Hence, the matrix $M(a)$ can be expressed as

$$M(a) = -a \begin{pmatrix} 1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \left(\frac{1}{\alpha_2} U_2(\alpha_{(2)}) \right) \begin{pmatrix} 1 & 0 \\ 0 & 1/(\alpha_2 - 1) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \alpha_2 \end{pmatrix} U_2(\alpha_{(2)}) \begin{pmatrix} 1 & 0 \\ 0 & 1/(\alpha_2 - 1) \end{pmatrix}.$$

It follows from [8, Theorem 5.3] that the representation matrix U_2 can be expressed as

$$U_2(\alpha_{(2)}; x) = C(\alpha) P_2(\alpha)^{-1} D_2(x) Q_2(\alpha_{(2)}) C(\alpha_{(2)})^{-1}.$$

We use the notation $|\tilde{x}\langle ij\rangle|$, which is the determinant of the minor matrix consisting of the i-th column and the j-th column of the matrix $\tilde{x} = \left(\begin{smallmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & x & 1 \end{smallmatrix}\right)$, where the numbering starts with 0 (see [8] as to details). We put $\varphi\langle ij\rangle = \frac{|\tilde{x}\langle ij\rangle| dt}{L_i L_j}$, where $L_0 = 1$, $L_1 = t$, $L_2 = 1 + xt$, and $L_3 = 1 + t$. We have the following expressions with these notations.

$$\begin{aligned}
D_2(x) &= \operatorname{diag}\left(\frac{|\tilde{x}\langle 21\rangle|}{|\tilde{x}\langle 01\rangle|}, \frac{|\tilde{x}\langle 23\rangle|}{|\tilde{x}\langle 03\rangle|} \right) = \operatorname{diag}(1, 1-x) = \begin{pmatrix} 1 & 0 \\ 0 & 1-x \end{pmatrix}, \\
C(\alpha) &= \begin{pmatrix} \mathcal{I}(\varphi\langle 01\rangle, \varphi\langle 01\rangle) & \mathcal{I}(\varphi\langle 01\rangle, \varphi\langle 02\rangle) \\ \mathcal{I}(\varphi\langle 02\rangle, \varphi\langle 01\rangle) & \mathcal{I}(\varphi\langle 02\rangle, \varphi\langle 02\rangle) \end{pmatrix} = 2\pi\sqrt{-1} \begin{pmatrix} \frac{1}{\alpha_0} + \frac{1}{\alpha_1} & \frac{1}{\alpha_0} \\ \frac{1}{\alpha_0} & \frac{1}{\alpha_0} + \frac{1}{\alpha_2} \end{pmatrix}, \\
Q_2(\alpha) &= \begin{pmatrix} \mathcal{I}(\varphi\langle 01\rangle, \varphi\langle 01\rangle) & \mathcal{I}(\varphi\langle 01\rangle, \varphi\langle 02\rangle) \\ \mathcal{I}(\varphi\langle 03\rangle, \varphi\langle 01\rangle) & \mathcal{I}(\varphi\langle 03\rangle, \varphi\langle 02\rangle) \end{pmatrix} = 2\pi\sqrt{-1} \begin{pmatrix} \frac{1}{\alpha_0} + \frac{1}{\alpha_1} & \frac{1}{\alpha_0} \\ \frac{1}{\alpha_0} & \frac{1}{\alpha_0} \end{pmatrix}, \\
P_2(\alpha) &= \begin{pmatrix} \mathcal{I}(\varphi\langle 21\rangle, \varphi\langle 01\rangle) & \mathcal{I}(\varphi\langle 21\rangle, \varphi\langle 02\rangle) \\ \mathcal{I}(\varphi\langle 23\rangle, \varphi\langle 01\rangle) & \mathcal{I}(\varphi\langle 23\rangle, \varphi\langle 02\rangle) \end{pmatrix} = 2\pi\sqrt{-1} \begin{pmatrix} \frac{1}{\alpha_1} & -\frac{1}{\alpha_2} \\ 0 & -\frac{1}{\alpha_2} \end{pmatrix},
\end{aligned}$$

where $\mathcal{I}$ is the intersection form on the twisted cohomology group. The inverse matrices of them can also be expressed in terms of intersection numbers as in [8, Appendix]. This method is implemented as the function `invintMatrix_k` in our package and it outputs

$$\begin{aligned}
P_2(\alpha)^{-1} &= \frac{1}{(2\pi\sqrt{-1})^2} \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \begin{pmatrix} \mathcal{I}(\varphi\langle 31\rangle, \varphi\langle 01\rangle) & \mathcal{I}(\varphi\langle 31\rangle, \varphi\langle 03\rangle) \\ \mathcal{I}(\varphi\langle 32\rangle, \varphi\langle 01\rangle) & \mathcal{I}(\varphi\langle 32\rangle, \varphi\langle 03\rangle) \end{pmatrix} \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_3 \end{pmatrix} \\
&= \frac{1}{2\pi\sqrt{-1}} \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \begin{pmatrix} \frac{1}{\alpha_1} & -\frac{1}{\alpha_3} \\ 0 & -\frac{1}{\alpha_3} \end{pmatrix} \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_3 \end{pmatrix} = \frac{1}{2\pi\sqrt{-1}} \begin{pmatrix} \alpha_1 & -\alpha_1 \\ 0 & -\alpha_2 \end{pmatrix}, \\
C(\alpha)^{-1} &= \frac{1}{(2\pi\sqrt{-1})^2} \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \begin{pmatrix} \mathcal{I}(\varphi\langle 31\rangle, \varphi\langle 31\rangle) & \mathcal{I}(\varphi\langle 31\rangle, \varphi\langle 32\rangle) \\ \mathcal{I}(\varphi\langle 32\rangle, \varphi\langle 31\rangle) & \mathcal{I}(\varphi\langle 32\rangle, \varphi\langle 32\rangle) \end{pmatrix} \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \\
&= \frac{1}{2\pi\sqrt{-1}} \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \begin{pmatrix} \frac{1}{\alpha_3} + \frac{1}{\alpha_1} & \frac{1}{\alpha_3} \\ \frac{1}{\alpha_3} & \frac{1}{\alpha_3} + \frac{1}{\alpha_2} \end{pmatrix} \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} = \frac{\alpha_1\alpha_2}{2\pi\sqrt{-1}\cdot\alpha_3} \begin{pmatrix} \frac{\alpha_1+\alpha_3}{\alpha_2} & 1 \\ 1 & \frac{\alpha_2+\alpha_3}{\alpha_1} \end{pmatrix}.
\end{aligned}$$

These matrices can be obtained in our program as

$D_2(x) =$ `repMatrix(2,1,1)`, $Q_2(\alpha)/(2\pi\sqrt{-1}) =$ `intMatrix([0,2],[0,3],1,1)`,
$P_2(\alpha)/(2\pi\sqrt{-1}) =$ `intMatrix([2,0],[0,3],1,1)`, $(2\pi\sqrt{-1})P_2(\alpha)^{-1} =$ `invintMatrix_k([2,0],[0,3],1,1)`,
$C(\alpha)/(2\pi\sqrt{-1}) =$ `intMatrix([0,3],[0,3],1,1)`, $(2\pi\sqrt{-1})C(\alpha)^{-1} =$ `invintMatrix_k([0,3],[0,3],1,1)`.

The argument (1, 1) stands for $(r_1 - 1, r_2 - 1)$.

Acknowledgement

This work was supported by MEXT/JSPS KAKENHI Grant Numbers JP 25220001, 17K05279, 18J01507, JST CREST Grant Number JP19209317 and by Research Institute for Mathematical Sciences, a Joint Usage/Research Center located in Kyoto University. We deeply appreciate several constructive criticisms by the reviewers, which made big improvements of our algorithms and implementation.

References

[1] A. Agresti, *Categorical data analysis*, 3rd ed., Wiley-Interscience, Hoboken, NJ, 2013.

[2] S. Amari, *Information geometry and its applications*, Applied Mathematical Sciences **194**, Springer, 2016.

[3] P. Billingsley, *Probability and measure*, 3rd ed., Wiley, 1995.

[4] R. P. Brent and P. Zimmermann, *Modern computer arithmetic*, Cambridge Monographs on Applied and Computational Mathematics **18**, Cambridge University Press, 2011.

[5] J. von zur Gathen and J. Gerhard, *Modern computer algebra*, 2nd ed., Cambridge University Press, Cambridge, 2003.

[6] Y. Goto, "Contiguity relations of Lauricella's F_D revisited", *Tohoku Math. J.* (2) **69**:2 (2017), 287–304.

[7] Y. Goto, "Intersection numbers of twisted cycles and cocycles for degenerate arrangements", 2018. arXiv 1805.01714

[8] Y. Goto and K. Matsumoto, "Pfaffian equations and contiguity relations of the hypergeometric function of type $(k+1, k+n+2)$ and their applications", *Funkcial. Ekvac.* **61**:3 (2018), 315–347.

[9] T. Granlund and the GMP development team, GNU Multiple Precision Arithmetic Library, 1991–2019, http://gmplib.org.

[10] T. Hibi et al., *Gröbner Bases: statistics and software systems*, Springer, 2013.

[11] J. van der Hoeven, "Faster chinese remaindering", 2016. hal-01403810.

[12] I. Karatzas and S. E. Shreve, *Brownian motion and stochastic calculus*, Graduate Texts in Mathematics **113**, Springer, 1988.

[13] T. Koyama, "A new formulation of sufficient σ-algebra", In preparation.

[14] D. Landers and L. Rogge, "Minimal sufficient σ-fields and minimal sufficient statistics: two counterexamples", *Ann. Math. Statist.* **43** (1972), 2045–2049.

[15] LINBOX: exact computational linear algebra, http://www.linalg.org.

[16] N. Möller, "On Schönhage's algorithm and subquadratic integer GCD computation", *Math. Comp.* **77**:261 (2008), 589–607.

[17] H. Nakayama, K. Nishiyama, M. Noro, K. Ohara, T. Sei, N. Takayama, and A. Takemura, "Holonomic gradient descent and its application to the Fisher–Bingham integral", *Adv. in Appl. Math.* **47**:3 (2011), 639–658.

[18] M. Noro and K. Yokoyama, "A modular method to compute the rational univariate representation of zero-dimensional ideals", *J. Symbolic Comput.* **28**:1-2 (1999), 243–263.

[19] M. Noro and K. Yokoyama, *Computation of Gröbner bases: introduction to computational algebra*, University of Tokyo Press, 2003. In Japanese.

[20] M. Ogawa, *Algebraic statistical methods for conditional inference of discrete statistical models*, PhD thesis, University of Tokyo, 2015.

[21] K. Ohara and N. Takayama, "Pfaffian systems of A-hypergeometric systems, II: Holonomic gradient method", 2015. arXiv 1505.02947

[22] T. Oshima, *Fractional calculus of Weyl algebra and Fuchsian differential equations*, MSJ Memoirs **28**, Mathematical Society of Japan, Tokyo, 2012.

[23] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes: the art of scientific computing*, 3rd ed., Cambridge University Press, 2007.

[24] Risa/Asir, a computer algebra system, http://www.math.kobe-u.ac.jp/Asir.

[25] T. Sasaki and T. Takeshima, "A modular method for Gröbner-basis construction over **Q** and solving system of algebraic equations", *J. Inform. Process.* **12**:4 (1989), 371–379.

[26] M. J. Schervish, *Theory of statistics*, Springer, 1995.

[27] J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, Springer, 1980.

[28] N. Takayama, "Gröbner basis and the problem of contiguous relations", *Japan J. Appl. Math.* **6**:1 (1989), 147–160.

[29] N. Takayama, S. Kuriki, and A. Takemura, "A-hypergeometric distributions and Newton polytopes", *Adv. in Appl. Math.* **99** (2018), 109–133.

[30] D. Williams, *Probability with martingales*, Cambridge University Press, 1991.

Received 2018-06-14. Revised 2020-01-05. Accepted 2020-03-24.

YOSHIHITO TACHIBANA: tatibana@math.kobe-u.ac.jp
Kobe University, Kobe 657-8501, Japan

YOSHIAKI GOTO: goto@res.otaru-uc.ac.jp
Otaru University of Commerce, Otaru 047-8501, Japan

TAMIO KOYAMA: koyama@wakhok.ac.jp
Wakkanai Hokusei Gakuen University, Wakkanai 097-0013, Japan

NOBUKI TAKAYAMA: takayama@math.kobe-u.ac.jp
Kobe University, Kobe 657-8501, Japan