Holonomic Gradient Method for Two Way Contingency Tables *

Yoshihito Tachibana, Yoshiaki Goto, Tamio Koyama, Nobuki Takayama

July 6, 2018

Abstract

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 for an efficient and exact evaluation, and we discuss on complexities of these algorithms and their implementation. We also discuss on a theoretical aspect of the distribution from the viewpoint of the conditional maximum likelihood estimation.

1 Introduction

The holonomic gradient method (HGM) proposed in [12] 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 [5] 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., [14], [22]) has been used for efficient and exact evaluation over the field of rational numbers. We apply the modular method to our evaluation procedure. We explore and discuss the complexities and implementation of these algorithms in Sections 4 and 5.

We here turn from computation to a theoretical quandary 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 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 \(\sigma\) 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 [16] 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

*Elements of sections 2–5 of this paper have previously been published as a research letter in Japanese [24].
references on contingency tables. Among them, we will refer to [11] and [7, Chap 4] herein.

2.1 \( r_1 \times r_2 \) Contingency Table

**Definition 1** \((r_1 \times r_2 \text{ (2 way) contingency table})\) An \( r_1 \times r_2 \) matrix with non-negative integer entries is called the \( r_1 \times r_2 \) contingency table. For a contingency table \( u = (u_{ij}) \), we define the row sum vector by \( \beta^r = \left( \sum_j u_{1j}, \cdots, \sum_j u_{r_1j} \right)^T \), 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 \( u^T \). The column vector obtained by joining \( \beta^r \) and \( \beta^c \) is denoted by \( \beta \), which is called the row column sum vector or the marginal sum vector.

**Example 1** \((2 \times 3 \text{ contingency table and the row sum and the column sum})\) For the \( 2 \times 3 \) contingency table \( u = \begin{pmatrix} 5 & 3 & 6 \\ 7 & 2 & 4 \end{pmatrix} \) 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 \( u^T \) and \( \beta \) are

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

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

\[
P_u^p = \frac{N! p^u}{u! p^u} = \prod_{i,j} p_{ij}^{u_{ij}}, \quad u! = \prod_{i,j} u_{ij}!,
\]

on contingency tables satisfying \( |u| = \sum_{i,j} u_{ij} = N \). The conditional distribution obtained by fixing the row sum vector \( \beta^r \) and the column sum vector \( \beta^c \) is

\[
Z(\beta; p) = \sum_{A u^T = \beta, \ u \in \mathbb{N}_{>0}^{r_1 \times r_2}} \frac{p^u}{u!} \cdot \frac{\partial \log Z}{\partial p_{ij}}.
\]

(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) \( Au^T \) is the marginal sum vector (see Example 2). The expectation of the \( u \)-value at \((i,j)\) of this conditional distribution is equal to

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

(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** \((\text{example of } A)\) When \( u^T = \begin{pmatrix} 5 & 3 & 6 & 7 & 2 & 4 \end{pmatrix}^T \), 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

\[
Au^T = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ 3 \\ 6 \\ 7 \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 14 \\ 13 \\ 12 \\ 5 \\ 10 \end{pmatrix} = \beta.
\]
Example 3 We consider $2 \times 2$ contingency tables with the marginal sum vector $\beta = (5, 7, 8, 4)^T$. All contingency tables $u$ satisfying $Au^T = \beta$ 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

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

3 The Normalizing Constant of $2 \times 2$ 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., [7, p.399, 6.13]). We will illustrate this correspondence for $2 \times 2$ tables.

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

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

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

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

$$
= \frac{p_{11}^{u_{11}}p_{12}^{u_{21}}p_{22}^{u_{22}}}{u_{11}!u_{21}!u_{22}!} \sum_{i=0}^{n} \frac{(-u_{11})_{-i}(-u_{22})_{i}}{(u_{21}+1)_{i}!(1)_{i}} \left( \frac{p_{12}p_{21}}{p_{11}p_{22}} \right)^i.
$$

Note that when $a, b \in \mathbb{Z}_{\leq 0}$, it is a polynomial. 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 $2 \times 2$ contingency tables with a fixed marginal sum vector in terms of the Gauss hypergeometric function. For the $r_1 \times r_2$ contingency table, the normalizing constant with a fixed marginal sum vector can be expressed in terms of the Aomoto-Gelfand hypergeometric function of type $(r_1, r_1 + r_2)$ [28] (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 [5].

Example 4 (the case of $2F_1$) Put $f(a) = 2F_1(a, b; c; x)$ and

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

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

$$
F(a) = M(a)F(a + 1).
$$
Now, note the following relations:

\[ \frac{1}{a}(a + \theta_x) \cdot f(a) = F(a + 1), \]
\[ \{\theta_x(c - 1 + \theta_x) - x(a + \theta_x)(b + \theta_x)\} \cdot f(a) = 0. \]  

The first relation can be shown from the series expansion and the second relation is the Gauss hypergeometric differential equation. It follows from \[3\], \[5\] that we have

\[ \frac{1}{a}(a + \theta_x) \cdot F(a) = F(a + 1), \]
\[ \theta_x F(a) = \begin{pmatrix} aE & 1 \\ ax + bx - c + 1 & 1 \end{pmatrix} F(a) = A(a)F(a). \]

Next, we have \[8\] as

\[ \frac{1}{a}(a + \theta_x) \cdot F(a) = \frac{1}{a}(aE + A(a))F(a), \]
\[ F(a) = \left( \frac{1}{a}(aE + A(a)) \right)^{-1} F(a + 1) = M(a)F(a + 1), \]

where \( E \) is the identity matrix.

A relation like \( F(a) = M(a)F(a + 1) \) is called a contiguity relation. In \[5\], the vector valued function \( F(a) \) is called the Gauss-Manin vector.

There are several algorithms to obtain contiguity relations \[25\], \[18\], \[17\], \[5\]. Among them, we choose to use the method of twisted cohomology groups given in \[5\], because it is the most efficient method for the case of two-way contingency tables.

We briefly summarize the method given in \[5\]. 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 \( \beta \) and the variable \( x = (x_{ij})_{1 \leq i \leq r_1-1, 1 \leq j \leq r_2-1} \) 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. \( H^\prime \)) 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 \( H \) so that the “integral” of \((\varphi_{1}', \ldots, \varphi_r)^T \) gives a constant multiple of the Gauss-Manin vector

\[ F(\alpha;x) = (f(\alpha;x), \delta^{(2)} \cdot f(\alpha;x), \ldots, \delta^{(r)} \cdot f(\alpha;x))^T, \]

where \( \delta^{(i)} \) is some differential operator with respect to \( x = (x_{ij}) \). There exist a basis \( \varphi_{1}', \ldots, \varphi_r \) of \( H^\prime \) and a linear map \( U_i : H^\prime \to H \) such that the integral of \((U_i(\varphi_{1}'), \ldots, U_i(\varphi_r'))^T \) 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 \( U_i \) with respect to the bases \( \{\varphi_{i}'\} \) and \( \{\varphi_{j}\} \):

\[ (U_i(\varphi_{1}'), \ldots, U_i(\varphi_r'))^T = U_i(\alpha;x) \cdot (\varphi_{1}, \ldots, \varphi_r)^T. \]

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. As to details, see \[5\]. 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 $U_2(a; x)$. The representation matrix $U_2$ has the following decomposition

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

The matrices except the diagonal matrix $\text{diag}(1, 1 - x)$ 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 $\theta_x$.

The method put forward in [5] is a polynomial time algorithm with respect to $r_2$ when $r_1$ is fixed.

**Theorem 1** Assume that the complexity of arithmetic of polynomials is $O(1)$, the complexity of multiplying two $n \times n$ matrices is $O(n^3)$, and the complexity of evaluating the determinant of an $n \times n$ matrix is $O(n^3)$. The complexity of obtaining a contiguity relation for the normalizing constant $Z$ for $r_1 \times r_2$ contingency tables is

1. $O(r_1 r_3^2)$ when $r \geq r_1$ where $r = \binom{r_1 + r_2 - 2}{r_1 - 1}$,
2. $O(r_1^3 r_2^3)$ when $r_1$ is fixed,
3. $O(r_1^3 r_2)$ when $r_2$ is fixed,
4. $O(2^6 r_1)$ when $r_1 = r_2$.

**Proof** Appendix will help to follow this proof. By [5, Theorem 5.3], to obtain a contiguity relation, we evaluate five matrices of size $r = \binom{r_1 + r_2 - 2}{r_1 - 1} = \binom{r_1 + r_2 - 2}{(r_1 - 1)(r_2 - 1)}^r$.

(a) one diagonal matrix (the complexity $O(r_1^3 r)$),
(b) two intersection matrices (the complexity $O(r_1^3 r_2^2)$),
(c) two inverse matrices of intersection matrices (the complexity $O(r_1^3 r_2^2)$)

and their product (we explain each complexity later). Since the size of each matrix is $r$, the complexity of multiplication is $O(r^3)$. Thus, the complexity of obtaining a contiguity relation is $O(r^3) + O(r_1^3 r) + O(r_1^3 r_2^2)$.

1. If $r \geq r_1$, the complexity is $O(r_1 r_3^3)$.
2. We fix $r_1$. By the Stirling formula $\log n! \sim n \log n - n$, we can show that $\log r \sim r_1 \log r_2$. Then we have $r \sim r_1^2$ and the complexity is $O(r_1^3 r_1^2)$.
3. The claim 3 can be obtained by a similar argument to the claim 2.
4. If $r_1 = r_2$, then the Stirling formula implies $r \sim 2^6 r_1$. Thus, the complexity is $O(2^6 r_1)$.

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

(a) As [5, 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 $O(r_1^3 r)$.

(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 [5, 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 $O(r_1^3 r_2^2)$.

1 see Appendix as to more details.
(c) By the proof of [5, 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 entries 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.

By the contiguity relation, we can evaluate the normalizing constant $Z$ and its derivatives. Let us explain the procedure for the case of $2F_1$.

Suppose $a \in \mathbb{Z}_{<-1}$. By the contiguity relation (3), we have

$$F(a) = M(a)F(a+1) = M(a)M(a+1)F(a+2) = \cdots = M(a)M(a+1)\cdots M(-2)F(-1).$$

Then, we can obtain the value of $F(a)$ from the initial value $F(-1) = (1 - \frac{1}{2}x, -\frac{1}{2}x)^T$ 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 in the case of $r_1 \times r_2$ contingency tables with the Gauss-Manin vector and contiguity relations given in [5]. In this section we conduct a complexity analysis of their 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 because of large numbers even for small contingency tables when $|a|$ is large.

5 Modular method

To perform exact and efficient evaluations for rational numbers by difference HGM, we need a fast and exact evaluation of a composite of linear transformations for vectors with rational number entries. This problem has hitherto been explored and there are several implementations, e.g., LINBOX [11]. For the purposes of empirical application, we study an efficient implementation with a distributed computation method on Risa/Asir [21]. Our implementation is published as the package gtt-ekn for Risa/Asir. Function names in this section are those in this package. We also give a complexity analysis of the difference HGM with the modular method.

5.1 Complexity

Before discussing on the complexity of our implementation by the modular method, we give the complexity of a naive method (called the algorithm plain), which performs the difference HGM by linear transformations (6) without the modular method.

**Proposition 1** Let the size of the square matrix for linear transformations be $r = \left( \begin{array}{c} r_1 + r_2 - 1 \\ r_1 - 1 \end{array} \right)$, the number of times of linear transformations $n$. We suppose the complexity of arithmetic operations of $m$ digits numbers is $O(m^2)$, the input numbers have digits of a fixed size, and digits increase by $d$ by each linear transformation. Under these assumptions, the complexity of the algorithm plain is $O(r^2 d^2 n^3)$.

**Proof** One linear transformation consists of multiplications of $r^2$ times and sums of $r(r-1)$ times. When the initial value vector consists of $x$ digits integers, the complexity of the algorithm plain is

$$(2r^2 - r) \sum_{i=0}^{n} (x + di)^2 = \frac{2r^2 d^2 n^3}{6} + O(rn^2).$$
There are two bottle necks in a naive difference HGM. The first one is that the algorithm plain becomes slower when marginal sums become bigger because \( n \) in the proposition is proportional to marginal sums. The second one is that use of rational numbers need to get common denominators and to perform divisions of common factors of numerators and denominators. Although these operations does not increase the order of the complexity, they increase the constant factor of the complexity because of these extra operations. We use the following methods to avoid these bottle necks.

1. We evaluate the denominator and the numerator of the image of the iteration of linear transformations separately and divide the common factor of the denominator and the numerator in the last step (\texttt{gmat_fac_int} in the \texttt{gtt}

2. For a set of prime numbers \( \{P_i\} \) we perform linear transformations over the finite field \( \mathbb{F}_{P_i} \). We reconstruct the result by linear transformations over \( \mathbb{Q} \) from the results by linear transformations over finite fields by the Chinese remainder theorem and the algorithm \texttt{IntegerToRational} ([15 Algorithm 6.25]). This method is called the modular method. The linear linear transformations over finite fields may be done on distributed processors (\texttt{gmat_fac_int} in the \texttt{gtt}

The modular method is a powerful computation scheme and is applied to several problems in computer algebra, see, e.g., [13], [22]. We propose to apply the modular method to the difference HGM. We summarize our methods as follows.

**Algorithm 1 (\texttt{gmat_fac_int} (generalized matrix factorial by integers))**

Input: \( M(k) \): a matrix of a variable \( k \), \( F \): a vector, \( S \): a starting \( k \), \( E \): an end point of \( k \), \( \text{Inc} \): increment of \( k \).

Output: \( M(E) \cdot \cdots \cdot M(S + \text{Inc})M(S)F \): image of \( F \) by applying linear transformations.

1. Normalize \( \text{Inc} \) as \( E := \frac{E - S}{\text{Inc}} \), \( S := 0 \), \( \text{Inc} := 1 \). Change the variable \( k \) of \( M \) to fit to this normalization. Let \( DNF, NMF, DNM, NMM \) be denominators and numerators of \( F \), \( M \) respectively (\( DNF, DNM \) are integers, \( NMF \) is a vector, \( NMM \) is a matrix). Put \( I = 0 \).
2. \( NMF := NMM(I) \times NMF \), \( DNF := DNM(I) \times DNF \).
3. \( I := I + 1 \). If \( I \) is less than \( E \), then goto 2 else return \( NMF \).

**Algorithm 2 (\texttt{gmat_fac_iter} (generalized matrix factorial by itor), modular method)**

Input: \( M(k) \), \( F \), \( S \), \( E \), \( \text{Inc} \), \( \text{PList} \): a list of prime numbers, \( \text{IDLList} \): a list of processes for distributed computation.

Output: A candidate value of \( M(E) \cdot \cdots \cdot M(S + \text{Inc})M(S)F \).

1. Normalize \( \text{Inc} \) as \( E := \frac{E - S}{\text{Inc}} \), \( S := 0 \), \( \text{Inc} := 1 \). Change the variable \( k \) of \( M \) to fit to this normalization. Let \( DNF, NMF, DNM, NMM \) be denominators and numerators of \( F \), \( M \) respectively. Put \( I = 0 \).
2. For each prime number \( P_i \) in \( \text{PList} \), perform linear transformations of \( F \) over \( \mathbb{F}_{P_i} \). If \( DNF \) or \( DNM(I) \) is not invertible modulo \( P_i \) (unlucky case), then skip this prime number \( P_i \). Let the output be \( G^t \). This step may be distributed to processes in the \( \text{IDLList} \).
3. Apply the Chinese remainder theorem to construct a vector \( G \) satisfying \( G \equiv G^t \pmod{P_i} \) where \( P = \prod_{P_i \in \text{PList}} P_i \).
4. Return a candidate value by the procedure \texttt{IntegerToRational}(\( G, P \)).

The complexity of the Algorithm \texttt{gmat_fac_int} is same with the algorithm plain. The complexity of the modular method \texttt{gmat_fac_iter} is given as follows. It is linear with respect to \( n \) (the size of the marginal sum vector), which will also be demonstrated by timing data in next subsections.

**Theorem 2** Let \( n \) be the times of the linear transformations and the size of the square matrix \( r = \begin{pmatrix} r_1 + r_2 - 2 \\ r_1 - 1 \end{pmatrix} \). 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. We assume that the complexity of four arithmetic operations of \( m \) digits numbers is \( O(m^2) \). The complexity of \texttt{gmat_fac_iter} is

\[
\max \left\{ O \left( \frac{N_p^2 d_p^2}{C} \right), O \left( (N_p d_p)^6 \right) \right\}.
\]
Proof  We estimate the complexity of each step of \texttt{g_mat_fac_iterator}.

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

2. In order to find an integer $x$ such that $x \equiv x_i \mod p_i$, $(i = 1, 2)$, we may solve $mp_1+np_2 = x_1-x_2$ and put $x = x_1-mp_1$. Suppose $p_1 \geq p_2$. Substitute $p_1$ in $mp_1+np_2 = x_1-x_2 =: y$ by $p_1 = q_2p_2+p_3$, $(p_3$ is the remainder of the division of $p_1$ by $p_2)$, then we have $(mq_2+n)p_2+mp_3 = y$. Let $(m',n')$ be a solution of $mp_1+np_2 = y$. Then, $m = n', n = m'-n'q_2$ is a solution of $mp_1+np_2 = y$. Thus, solving a linear indefinite equation is reduced to solving a linear indefinite equation of smaller coefficients.

We denote by $p_l$ the number (the remainder) which is obtained by the $i$-th step of the procedure above. We estimate the complexity of this procedure of solving linear indefinite equations. Let $f_i$ be the number of digits of $p_l$ as a $b$-adic number. By choosing a suitable base $b$, we may assume that the number of the linear indefinite equations obtained in this procedure is $f_1$ at most. Put $f_0 = f_1 + 1$. $f_k$ is at most $f_0 - k$. Hence, the number of digits of $p_k,q_k$ is at most $O(f_0 - k), k = 1, \ldots, f_0$. When the number of digits of $y$ is $O(f_0)$, the number of digits of the solution of the last linear indefinite equation is $O(f_0)$ at most. Then, the complexity of the step $i$ of the construction $m'-n'q_k$ is bounded by $O(f_0+f_1+f_1+i+\cdots+f_{i-1}f_1)$. The complexity of the division is bounded by $O(f_0^2)$ and it is smaller than $O(f_1f_0)$. Therefore, the complexity of the step of constructing $x$ by the Chinese remainder theorem is bounded by $\sum_{k=1}^{f_0}O(f_0(f_0-k)^2)$. The sum is bounded by $O(f_0^4)$. We apply the Chinese remainder theorem by a tournament method. In other words, we construct $x_{i+1}$ which satisfies $x_{i+1} \equiv x_i \mod P_i$ and $x_{i+1} \equiv x_{i+1} \mod P_i+1$ at first. Secondly, we construct $x_{i+3}$ which satisfies $x_{i+3} \equiv x_{i+3} \mod P_iP_{i+1}$ and $x_{i+3} \equiv x_{i+2,i+3} \mod P_{i+2}P_{i+3}$. And we repeat this tournament. Since the number of digits of the prime numbers is $d_p$, $f_0$ in the last step of the tournament is $N_p^sd_p$. We solve linear indefinite equations $\log(N_p^sd_p)$ times during the tournament. Hence the complexity is $\sum_{i=1}^{\log(N_p^sd_p)}\left(\frac{N_p^sd_p}{2}\right)^4$, which is equal to $O(N_p^6d_p^616\log(N_p^sd_p))$. It is bounded by $O(N_p^6d_p^6)$. We perform this procedure for each element of a vector of length $r$, then the complexity of this step is $O(r(N_p^6d_p)^6)$.

3. The algorithm \texttt{IntegerToRational} is a variation of the Euclidean algorithm and its complexity is bounded by $O(N_p^6d_p)$.

Summarizing these estimates, we obtain the conclusion. \hfill \Box

5.2 Computational Experiments for $2 \times 2$ Tables

We give timing data of the algorithms plain, \texttt{g_mat_fac_int} (integer arithmetic operations), \texttt{g_mat_fac_iterator} (modular method) for evaluating

\[ f = 2F_1(-36N,-11N,2N;\frac{1}{56}), \quad N \in \mathbb{N}. \]

The timing data are taken on a machine with

\begin{itemize}
  \item \textbf{cpu}: Intel(R) Xeon(R) CPU E5-4650 2.70GHz
  \item \textbf{the number of cpu’s}: 32
  \item \textbf{the number of cores}: 8
  \item \textbf{os}: debian 7.8
  \item \textbf{memory}: 256GB
  \item \textbf{software system}: Risa/Asir version20150126
\end{itemize}
The figure 1 illustrates that the time of computation is proportional to $N^2$. It takes about 6 minutes when $N = 100$.

Figure 1: Algorithm plain

It is also proportional to $N^2$, but the constant factor is smaller than the algorithm plain (note that the scale of $N$ is different with the algorithm plain).

Figure 2: Algorithm $g_{\text{mat-fac-int}}$ (by integer arithmetic operations)
The timing data in the Figure 3 is that of the algorithm `g_mat_facitor` (modular method) with the process number \( C = 16 \), and the number of primes \( N_p = 1000 \) of which the number of digits is about \( d_p = 100 \). Note that the computation time is linear with respect to \( N \). The Figure 4 gives the timing data with respect to the number of distributed processes. The distributed computation is executed by the `ex_rpc` function of Risa/Asir [13, Chapter 20].

The timing data is proportional to \( \frac{1}{C} \) where \( C \) is the number of processes. Our complexity analysis is supported by these experimental data. The next figure illustrates the modular method is stronger than the algorithm `g_mat_facint` (by integer arithmetic operations) when \( N \) becomes larger by superimposing the Figure 2 to the Figure 3.
In this example, the modular method becomes faster than the non-modular method when $N$ is larger than about 600. We have demonstrated the modular method mainly studied in computer algebra (see, e.g., [14], [22]) is useful and effective for statistical computation by the difference HGM both by complexity analysis (Theorem 2) and by computational experiments.

5.3 $5 \times 5$, $7 \times 7$ Contingency Tables

We will demonstrate that our method works for larger contingency tables. The Figure 6 is a timing data for the case of $5 \times 5$ contingency tables. The size of the matrix of the linear transformation is $r = \binom{8}{4} = 70$. The step 1 is the derivation time of the linear transformation (contiguity relation) and the step 2 is the timing data of the algorithm $\text{g_mat_fac_itor}$ (modular method).

We use the marginal sum vector $(4N, 4N, 4N, 4N, 4N, 2N, 3N, 5N, 5N, 5N)^T$. The number of primes is 200, of which number of digits is 100. We use 8 processes. The computation is performed in the linear
time with respect to $N$ with the modular method. This timing data demonstrates that our method works well in the case of $5 \times 5$ contingency tables.

It also works well for $7 \times 7$ contingency tables. The size of the matrix of the linear transformation is $r = \binom{12}{6} = 924$. We use the marginal sum vector $(4N, 4N, 5N, 5N, 5N, 5N, 5N, 6N, 3N, 4N, 4N, 4N, 6N, 6N)^T$. The number of primes is 200, of which number of digits is 100. We use 16 processes.

![Figure 7: 7 \times 7 contingency table](image)

6 Zero Cells

The contiguity relations derived by [5] 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 we cannot therefore 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 $E[U_{ij}] = \frac{\partial \log Z}{\partial p_{ij}}$ 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 2** Let $\beta$ be the marginal sum vector and $L$ a generic line in $p$-space. If we evaluate $E[U_{ij}]$ at $2\beta_1$ points $p \in \mathbb{R}^{r_1 \times r_2}$ 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 $\beta_1$ at most. Apply an interpolation algorithm by rational function, e.g., Stoer-Bulirsch algorithm [23], [19]. Then, we can obtain the exact value by interpolation. □

**Example 6** Let the marginal sums and the cell probabilities be

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

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

```
[5150] import("gtt_ekn.rst");
0
[5151] E=gtt_ekn.cBasistoE_0([3,4,3],[3,4,3],[1,1/2,0],[1,1/3,1/4],[1,1,1]) | randinit=20;
```
Although the interpolation method is applicable to any pattern of 0-cells, a more efficient method involves utilizing hypergeometric functions restricted on some $p_{ij} = 0$'s. In general, contingency relations and Pfaffian systems for such hypergeometric functions become complicated. In [5], a method is put forward to evaluate intersection numbers and contiguity relations when only one $p_{ij}$ is zero.

### 7 Sufficient Statistics as $\sigma$-algebra

It is often that 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 new formulation of parameters of interest, nuisance parameters, and sufficient statistics.

Classical formulations of sufficient statistics as $\sigma$-algebras appear in [3], [10], and so on. Our formulation is different because we treat parameters as random variables instead of considering a family of probability measures. This Bayesian statistical approach enable us to consider $\sigma$-algebras on parameter spaces. We express nuisance parameters and parameters of interest as sub $\sigma$-algebras of the $\sigma$-algebra generated by all parameters and data.

The treatment of nuisance parameters and parameters of interest is an important issue in statistics.

The distinction between those parameters which are of interest versus those which are nuisance, may seem easy. In fact, it seems to be only a matter of declaring that $\mu$ is a parameter of interest or $\nu$ 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 $\sigma$-algebras on parameter spaces. In probability theory and stochastic processes, $\sigma$-algebra is important as a natural way to express information (see, e.g. [3]). Discussions in this section are based on conditional expectations with respect to $\sigma$-algebra. For basic properties of conditional expectation, see [24].

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

In standard parameter estimation, we assume a probability space $(\Omega', \mathcal{F}', \mathbf{P}')$ 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 $\sigma$-algebra on $\Omega$ generated by

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

The measurable space $(\Omega, \mathcal{F})$ is deemed to be the product measurable space of $(\Omega', \mathcal{F}')$ and $(\Theta, \mathcal{B}(\Theta))$ [27] p75. For $A \in \mathcal{F}'$, let $f_A : \Theta \to \mathbf{R}$ 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 $\theta$ be a measurable map from $\Omega$ to $\Theta$ defined by

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

This implies that parameters can be regarded as a $\Theta$-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 7 Let \((\Omega', \mathcal{F}', \mathbf{P}')\) be the probability space \((\mathbf{R}, \mathcal{B}(\mathbf{R}), N(\mu, \sigma^2))\), where \(N(\mu, \sigma^2)\) is the Gaussian distribution on \(\mathbf{R}\) with mean \(\mu\) and variance \(\sigma^2\). In this case, the parameter space is \(\Theta = \{ (\mu, \sigma^2) \in \mathbf{R}^2 | \sigma^2 > 0 \}\) and the parameter \(\theta\) 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 \(\sigma\)-algebra \(\mathcal{G}\) of \(\mathcal{F}\), we use \(L^1(\mathcal{G})\) to denote the linear space of random variables which are integrable and \(\mathcal{G}\)-measurable. When two elements \(X\) and \(Y\) of \(L^1(\mathcal{G})\) satisfy \(X(\omega) = Y(\omega)\) for all \(\omega \in \Omega\), we say that \(X = Y\). Note that \(X = Y\) almost surely does not imply that \(X = Y\).

Let \(\vartheta\) be the sub \(\sigma\)-algebra of \(\mathcal{F}\) generated by a random variable \(\theta\). It represents the information of \(\theta\). We formulate notions of nuisance parameters, sufficient parameters, and parameters of interest as sub \(\sigma\)-algebras of \(\vartheta\).

Let \(X\) and \(Y\) be \(\mathbf{R}\)-valued random variables and \(\theta\) be a \(\Theta\)-valued random variable, which we will call a parameter. We assume that \(X\) is integrable. Because the value of the conditional expectation \(E(X|Y, \theta)\) of \(X\) with respect to \(\sigma(Y, \theta)\), which is the \(\sigma\)-algebra generated by \(Y\) and \(\theta\), is uniquely determined by the outcome of \(Y\) and \(\theta\), \(E(X|Y, \theta)\) can be regarded as a function \(f(Y, \theta)\) of \(Y\) and \(\theta\), i.e., we can define \(f\) as a function from \(\mathbf{R} \times \Theta \) to \(\mathbf{R}\) by

\[ f(y, c) := E(X|Y, \theta)(\omega) \quad (Y(\omega) = y, \theta(\omega) = c). \]

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

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

Definition 2 Sub \(\sigma\)-algebra \(\mathcal{I}\) is said to be of interest with respect to a pair of sub \(\sigma\)-algebras \((\mathcal{D}, \mathcal{G})\) if, for all \(X \in L^1(\mathcal{D})\), \(E(X|\mathcal{G})\) 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 \(\sigma\)-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 1 Note that the condition of Definition 3 is equivalent to stating that the equation

\[ E(X|\sigma(\mathcal{S}, \mathcal{N})) = E(X|\mathcal{S}) \quad \text{a.s.} \]  

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

\[ E(X|\sigma(\mathcal{S}, \mathcal{N})) = E \left( E(X|\sigma(\mathcal{S}, \mathcal{N})) | \mathcal{S} \right) \]

\[ = E(X|\mathcal{S}) \quad \text{for all } \mathcal{S} \ni \mathcal{D} = \sigma(X), \quad S = \sigma(T), \quad N = \sigma(\theta). \]
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 [9].

**Example 8** For random variables $X_1, \ldots, X_n, \theta$, suppose that

1. $0 \leq \theta \leq 1$

2. The conditional probability of $X_1, \ldots, X_n$ for given $\theta$ 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 $(\mathcal{D}, \mathcal{N})$.

To describe a sub $\sigma$-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 $(\mathcal{S}, \Sigma)$. For $B \subset \mathcal{S}$ and $g \in G$, we put

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

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

Let $\mathcal{O}^*$ be the family of the element in $\Sigma$ invariant under the action of $G$, i.e., we put

$$
\mathcal{O}^* := \{B \in \Sigma : g \cdot B = B\}.
$$

**Lemma 1** $\mathcal{O}^*$ is a sub $\sigma$-algebra of $\Sigma$.

**Proof** The proof is straightforward. To be sure how to prove it, we only show that $B_n \in \mathcal{O}^* (n \in \mathbb{N})$ implies $\bigcup_{n=1}^{\infty} B_n \in \mathcal{O}^*$. Suppose that $B_n (n \in \mathbb{N})$ is an element of $\mathcal{O}^*$. Since we have $G \cdot \bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} G \cdot B_n = \bigcup_{n=1}^{\infty} B_n$, $\bigcup_{n=1}^{\infty} B_n$ is an element of $\mathcal{O}^*$. \qed

A measurable map $X : (\Omega, \mathcal{F}) \to (\mathcal{S}, \Sigma)$ induces a sub $\sigma$-algebra of $\mathcal{F}$ by

$$
\mathcal{O} := \{\{X \in B\} \mid B \in \mathcal{O}^*\}.
$$

Note that $\{X \in B\}$ is the inverse image $X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}$. This notation is often used in the probability theory and we use it in the sequel. We call $\mathcal{O}$ as the $\sigma$-algebra generated by the orbits of group $G$.

**Lemma 2** Let $f : \mathcal{S} \to \mathbb{R}$ be a function. Suppose that a measurable map $X : (\Omega, \mathcal{F}) \to (\mathcal{S}, \Sigma)$ is surjective. Then, all of the following four conditions are equivalent:

(a). $f$ is $\mathcal{O}^*$-measurable.

(b). $f(g \cdot x) = f(x)$ holds for any $g \in G$ and any $x \in \mathcal{S}$.

(c). $f(X)$ is $\mathcal{O}$-measurable.

(d). $f(g \cdot X) = f(X)$ holds for any $g \in G$.

**Proof** [(a)$\Rightarrow$(c)]. Suppose that $f$ is $\mathcal{O}^*$-measurable. For any $B \in \mathcal{B}(\mathbb{R})$, we have $f^{-1}(B) \in \mathcal{O}^*$. By the definition of $\mathcal{O}$, $X^{-1}(f^{-1}(B)) = \{f(X) \in B\} \in \mathcal{O}$ holds. Hence, $f(X)$ is $\mathcal{O}$-measurable.

[(c)$\Rightarrow$(d)]. Suppose that $f(X)$ is $\mathcal{O}$-measurable. Take an arbitrary $a \in \mathbb{R}$. Then, $\{f(X) = a\} \in \mathcal{O}$ implies that there exists $B \in \mathcal{O}^*$ such that $\{X \in B\} = \{f(X) = a\} = \{X \in f^{-1}(a)\}$. Since $X : \Omega \to \mathcal{S}$...
is surjective, we have $B = f^{-1}(a)$. Thus, $f^{-1}(a)$ is an element of $O^*$, and we have $G \cdot f^{-1}(a) = f^{-1}(a)$. This implies $g \cdot f^{-1}(a) = f^{-1}(a)$ holds for any $g \in G$, and we have
\[
\{f(g \cdot X) = a\} = \{g \cdot X \in f^{-1}(a)\} = \{X \in g^{-1} \cdot f^{-1}(a)\} = \{X \in f^{-1}(a)\} = \{f(X) = a\}.
\]
Hence, $f(g \cdot X) = f(X)$ holds for all $g \in G$.

[(d)$\Leftrightarrow$(b)]. Since $X : \Omega \rightarrow S$ is surjective, $f(g \cdot x) = f(x)$ holds for any $x \in S$ and $g \in G$ if and only if $f(g \cdot X) = f(X)$ holds for any $g \in G$.

[(b)$\Leftrightarrow$(a)]. Suppose the function $f$ is invariant under the action of $G$. Take an arbitrary $B \in \Sigma$. By the invariance of the function $f$, we have
\[
G \cdot f^{-1}(B) = \{g \cdot x | g \in G, x \in S, f(x) \in B \} = \{x \in S | f(x) \in B \} = f^{-1}(B).
\]
Hence, $f^{-1}(B)$ is included in $O^*$. This implies that $f$ is $O^*$-measurable. \hfill $\Box$

Since $O$ is a sub $\sigma$-algebra of $\sigma(X)$, $Y \in L^1(O)$ can be regarded as a function of $X$. By Lemma 2, we say that a random variable $Y$ is invariant under the action of group $G$ if $Y$ is $O$-measurable.

We apply the above discussion on group actions to sufficient $\sigma$-algebras. Let $D$ be a sub $\sigma$-algebra of $F$. Let $\Theta$ be a set, and $\theta : (\Omega, F) \rightarrow (\Theta, B(\Theta))$ be a measurable map. We regard $\theta$ and $\Theta$ as the parameter and the space of parameters respectively. Let $S$ be a set, and $T : \Omega \rightarrow S$ be an $D$-measurable map. For $X \in L^1(D)$, $E(X[T, \theta])$ can be regarded as a function on $S \times \Theta$. In other words, there exists a function $f_X : S \times \Theta \rightarrow \mathbb{R}$ such that $f_X(T(\omega), \theta(\omega)) = E(X[T, \theta](\omega))$ for all $\omega \in \Omega$.

**Lemma 3** We assume the same notation as above. Suppose that an action of group $G$ on $\Theta$ satisfies $f_X(t, g \cdot c) = f_X(t, c)$ for all $t \in S$, $c \in \Theta$, $g \in G$, and $X \in L^1(D)$, and put
\[
O := \{\{(T, \theta) \in B\} | B \in \Sigma \times B(\Theta), G \cdot B = B\}.
\]
Then, $O$ is of interest with respect to $(D, \sigma(T, \theta))$.

**Proof** The group action on $\Theta$ induces an group action on the Cartesian product $S \times \Theta$ by
\[
g \cdot (t, c) = (t, g \cdot c) \quad (g \in G, (t, c) \in S \times \Theta).
\]
Applying Lemma 2 in the case of the group action on $S \times \Theta$, $f_X(T, \theta)$ is $O$-measurable for any $X \in L^1(D)$. Hence, $O$ is of interest with respect to $(D, \sigma(T, \theta))$. \hfill $\Box$

We will use the following lemmas in the next section.

**Lemma 4** Let a measurable function $\theta : \Omega \rightarrow \Theta$ be surjective and $G$ be a sub $\sigma$-algebra of $\sigma(\theta) := \{\theta^{-1}[B] | B \in B(\Theta)\}$. Then, $\theta G := \{\theta(B) | B \in B(\Theta)\}$ is a sub $\sigma$-algebra of $B(\Theta)$.

**Proof** Since $\theta$ is surjective, $\Theta = \theta(\Omega)$ is an element of $\theta G$.

Let $A \in \theta G$. There exists $B \in G$ such that $A = \theta(B)$. By $G \subset \sigma(\theta)$, there exists $C \in B(\Theta)$ such that $B = \theta^{-1}C$. Since surjectivity of $\theta$ implies that $\theta(\theta^{-1}S) = S$ holds for any $S \subset \Theta$, we have $A = \theta(B) = \theta(\theta^{-1}C) = C$. By $\theta^{-1}A = \theta^{-1}C = B \in G$, we have $\theta^{-1}A^c = (\theta^{-1}A)^c \in G$. By surjectivity of $\theta$, $A^c = \theta(\theta^{-1}A^c)$ is an element of $\theta G$.

Suppose $A_n \in \theta G$ for $n \in \mathbb{N}$. Analogously, we have $\theta^{-1}A_n \in G$. Consequently, $\theta^{-1} \bigcup_{n \in \mathbb{N}} A_n = \bigcup_{n \in \mathbb{N}} \theta^{-1}A_n \in G$ implies $\bigcup_{n \in \mathbb{N}} A_n = \theta(\theta^{-1} \bigcup_{n \in \mathbb{N}} A_n) \in \theta G$. \hfill $\Box$

**Lemma 5** Suppose that a measurable function $\theta : \Omega \rightarrow \Theta$ is surjective. Let $f_\lambda : \Theta \rightarrow \mathbb{R}$ $(\lambda \in \Lambda)$ be measurable functions. Then, we have
\[
\sigma(f_\lambda \circ \theta : \lambda \in \Lambda) = \theta^{-1} \sigma(f_\lambda : \lambda \in \Lambda),
\]
where $\sigma(f_\lambda : \lambda \in \Lambda)$ is the $\sigma$-algebra generated by $\{f_\lambda^{-1}B | \lambda \in \Lambda, B \in B(\mathbb{R})\}$. \hfill (8)
Proof Obviously, the right hand side of (5) includes the left hand side. We show the opposite inclusion. By the surjectivity of $\theta$, we have $f_X^{-1}B = \theta^{-1}f_X^{-1}B = \theta(f_X \circ \theta)^{-1}B \in \theta \sigma(f_X \circ \theta : \lambda \in \Lambda)$ for any $B \in B(\mathbb{R})$. By Lemma 3 $\theta \sigma(f_X \circ \theta : \lambda \in \Lambda)$ is a sub-$\sigma$-algebra of $B(\Theta)$. Hence, we have

$$\sigma(f_X : \lambda \in \Lambda) \subset \theta \sigma(f_X \circ \theta : \lambda \in \Lambda).$$

(9)

Note that we have

$$C = \theta^{-1} \theta C \quad (C \in \sigma(\theta)).$$

(10)

In fact, since there exists $C' \in B(\Theta)$ such that $C = \theta^{-1} C'$, we have $\theta^{-1} \theta C = \theta^{-1} \theta \theta^{-1} C' = \theta^{-1} C' = C$.

Let $A \in \theta^{-1} \sigma(f_X : \lambda \in \Lambda)$. There exists $B \in \sigma(f_X : \lambda \in \Lambda)$ such that $A = \theta^{-1} B$. By (10), there exists $C \in \sigma(f_X \circ \theta : \lambda \in \Lambda)$ such that $B = \theta^{-1} C$. Equation (10) and $\sigma(f_X \circ \theta : \lambda \in \Lambda) \subset \sigma(\theta)$ implies $A = \theta^{-1} \theta C = C \in \sigma(f_X \circ \theta : \lambda \in \Lambda)$. Consequently, the opposite inclusion holds.

Lemma 6 Let $V$ and $W$ be finite-dimensional vector spaces over $\mathbb{R}$, $V \oplus W$ be the direct sum of $V$ and $W$, $\pi : V \oplus W \to V$ be the projection, and $V^*$ be the dual space of $V$. Then, we have

$$\{B \in B(V \oplus W) | B + W = B\} = \sigma(f \circ \pi : f \in V^*).$$

(11)

Here, we put $B + W := \{v + w | v \in B, w \in W\}$.

Proof Since $\pi^{-1} B + W = \pi^{-1} B$ holds for any $B \in B(V)$, $\{B \in B(V \oplus W) | B + W = B\}$ includes $\pi^{-1} B(V)$. Let $\iota : V \to V \oplus W$ be the canonical injection. Suppose that $B \in B(V \oplus W)$ satisfies $B + W = B$. Since we have

$$x \in \pi^{-1} \iota^{-1} B \iff \iota \pi(x) \in B$$

$$\iff \pi(x) \in B$$

$$\iff \pi(x) + (x - \pi(x)) \in B + W$$

$$\iff x \in B + W$$

$$\iff x \in B$$

$\pi^{-1} \iota^{-1} B \subset B$ holds. Since we can show the opposite inclusion analogously, we have $\pi^{-1} \iota^{-1} B = B$. By $\iota^{-1} B \in B(V)$, $B$ is an element of $\pi^{-1} B(V)$. Then, we have

$$\{B \in B(V \oplus W) | B + W = B\} = \pi^{-1} B(V).$$

Since $f \in V^*$ is a continuous map from $V$ to $\mathbb{R}$, $B(V)$ includes $\sigma(f : f \in V^*)$. Let $\{f_1, \ldots, f_n\}$ be a basis of $V^*$. Since any open subsets of $V \cong \mathbb{R}^n$ is a countable union of open sets of the form

$$\bigcap_{i=1}^n f_i^{-1}(\{x \in \mathbb{R} | a_i < x < b_i\}) \quad (a_i, b_i \in \mathbb{Q}),$$

we have $B(V) \subset \sigma(f_1, \ldots, f_n) \subset \sigma(f : f \in V^*)$. Consequently, $B(V) = \sigma(f : f \in V^*)$ holds and we have

$$\{B \in B(V \oplus W) | B + W = B\} = \pi^{-1} \sigma(f : f \in V^*).$$

By Lemma 3 the right hand side of the above equation equals to $\sigma(f \circ \pi : f \in V^*)$. \hfill \Box

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 $n$. Suppose that Poisson random variables $X_k \sim \text{Pois}(c_k)$, $(k = 1, \ldots, n)$ 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 $\mathcal{A}$-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

$$P(X = x | AX = b, \theta = c) = \frac{\prod_{j=1}^n \frac{c_j^{x_j}}{x_j!} \exp(-c_j)}{\sum_{Ay=b} \prod_{j=1}^n \frac{c_j^{y_j}}{y_j!} \exp(-c_j)} = \frac{\prod_{j=1}^n \frac{c_j^{x_j}}{x_j!}}{\sum_{Ay=b} \prod_{j=1}^n \frac{c_j^{y_j}}{y_j!}}.$$
An application of conditional distributions in statistics is the elimination of nuisance parameters. By Definition\(\text{3}\) and Remark\(\text{2}\) 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 $[7]$). 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 $A$-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 $\sigma$-$\sigma$-algebra under less restrictive conditions on $\theta$ and $c$.

The parameter $c$ of $A$-distribution moves on the set $\Theta := \mathbb{R}_{\geq 0}^n$. Consider the action of the multiplicative group $G := \mathbb{R}_{> 0}$ on the space $\Theta$ 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 $\Theta$ induces group action on $\mathbb{Z}_{\geq 0}^d \times \Theta$ by

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

Applying Lemma\(\text{3}\) in the case where $D = \sigma(X)$, $S = \mathbb{Z}_{\geq 0}^d$, and $T = AX$, we have the following theorem:

**Theorem 3** The sub $\sigma$-$\sigma$-algebra

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

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

**Proof** For any $g \in G$, we have

$$g \cdot \frac{\prod_{j=1}^d \theta_j^{x_j}}{\sum_{A \in B} \prod_{j=1}^d \theta_j^{y_j}/y_j!} = \frac{\prod_{j=1}^d \left( \theta_j^{x_j} \prod_{i=1}^n g_i^{a_{ij}x_i} \right) / x_j!}{\sum_{A \in B} \prod_{j=1}^n \left( \theta_j^{y_j} \prod_{i=1}^d g_i^{a_{ij}y_i} \right) / y_j!} = \frac{\prod_{j=1}^d \theta_j^{x_j} / x_j!}{\sum_{A \in B} \prod_{j=1}^d \theta_j^{y_j} / y_j!} = \frac{\prod_{j=1}^d \theta_j^{x_j}}{\sum_{A \in B} \prod_{j=1}^d \theta_j^{y_j}}.$$

Since the conditional distribution of $X$ with respect to $(AX, \theta)$ is invariant under the action of $G$ on $\mathbb{Z}_{\geq 0}^d \times \Theta$, for any $X \in L^1(\sigma(X))$, the conditional expectation $\mathbb{E}(X \mid AX, \theta)$ is also invariant under the action. By Lemma\(\text{3}\), $O$ is of interest with respect to $(\sigma(X), \sigma(AX, \theta))$. $\square$

Note that the quotient space $\Theta/G$ by the group action $G$ is not a manifold. Therein lies the difficulty with describing the space of parameters of interest and hence why we utilized the notion of $\sigma$-$\sigma$-algebra of interest.

For a vector $v = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$, 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 $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$, let $R_\alpha$ be the function from $\Theta = \mathbb{R}_{\geq 0}^n$ to $\mathbb{R}$ defined by

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

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

$$Z_j(c) := \begin{cases} 1 & (c_j > 0) \\ 0 & (c_j = 0) \end{cases}.$$
Lemma 7  The random variables $AX, R_\alpha(\theta) (\alpha \in \ker A)$, and $Z(\theta)$ are $\mathcal{O}$-measurable.

Proof  Obviously, $AX$ is $\mathcal{O}$-measurable. Let $\pi : \mathbb{Z}_\geq 0 \times \Theta \to \Theta$ be the projection. By some calculations, we have

$$R_\alpha \circ \pi(g \cdot (t, c)) = R_\alpha \circ \pi((t, c)), \quad Z \circ \pi(g \cdot (t, c)) = Z \circ \pi((t, c))$$

for any $\alpha \in \ker A, g \in G$, and $(t, c) \in \mathbb{Z}_\geq 0 \times \Theta$. Consequently, the functions $R_\alpha \circ \pi$ and $Z \circ \pi$ are invariant under $G$. Applying Lemma 2 in the case where $X = (AX, \theta), R_\alpha \circ \pi(X) = R_\alpha(\theta)$ and $Z \circ \pi(X) = Z(\theta)$ are $\mathcal{O}$-measurable. \hfill \Box

Let $\{e_1, \ldots, e_n\}$ be the standard basis of $\mathbb{R}^n$, i.e., the $i$-th component of $e_i$ is 1 and the other components are 0. For the $d \times n$ matrix $A$, $\ker A$ and $\text{Im} A^\top$ can be written as

$$\ker A = \left\{ \sum_{j=1}^n x_j e_j \mid \sum_{j=1}^n a_{ij} x_j = 0 \right\}, \quad \text{Im} A^\top = \sum_{i=1}^d \mathbb{R} \sum_{j=1}^n a_{ij} e_j,$$

where $a_{ij}$ is the $(i, j)$-component of $A$. For $z \in \{0, 1\}^n$, let $R^{J(z)} := \sum_{j \in J(z)} \mathbb{R} e_j$ be the sub vector space of $\mathbb{R}^n$ spanned by $e_j (j \in J(z))$, $p_z : \mathbb{R}^n \to R^{J(z)} \left( \sum_{j=1}^n x_j e_j \mapsto \sum_{j \in J(z)} x_j e_j \right)$ be the projection, and $\hat{\iota}_z : R^{J(z)} \to \mathbb{R}^n$ be the canonical injection. For $\alpha \in \mathbb{R}^n$, we denote by $L_\alpha$ the linear map from $\mathbb{R}^n$ to $\mathbb{R}$ defined by

$$L_\alpha(x) = \sum_{j=1}^n a_j x_j \quad \left( x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n \right).$$

Lemma 8  Under the same notation as above, the following equation holds for any $z \in \{0, 1\}^n$:

$$\left\{ B \in \mathcal{B}(R^{J(z)}) \mid B + p_z \text{Im} A^\top = B \right\} = \sigma(L_\alpha \hat{\iota}_z : \alpha \in R^{J(z)} \cap \ker A) \quad (12)$$

Proof  With a map

$$\langle \cdot, \cdot \rangle : R^{J(z)} \times R^{J(z)} \to \mathbb{R} \left( \left( \sum_{j \in J(z)} x_j e_j, \sum_{j \in J(z)} y_j e_j \right) \mapsto \sum_{j \in J(z)} x_j y_j \right),$$

$R^{J(z)}$ is an inner product space. By the equation

$$R^{J(z)} \cap \ker A = \{ v \in R^{J(z)} \mid \langle v, w \rangle = 0 \text{ for all } w \in p_z \text{Im} A^\top \},$$

we have

$$R^{J(z)} = \left( R^{J(z)} \cap \ker A \right) \oplus p_z \text{Im} A^\top.$$

By Lemma 8, we have

$$\left\{ B \in \mathcal{B}(R^{J(z)}) \mid B + p_z \text{Im} A^\top = B \right\} = \sigma \left( f \circ \pi : f \in \left( R^{J(z)} \cap \ker A \right)^* \right),$$

where $\pi : R^{J(z)} \to R^{J(z)} \cap \ker A$ is the projection and $(R^{J(z)} \cap \ker A)^*$ denotes the dual space of $R^{J(z)} \cap \ker A$. Since we have

$$\left\{ f \circ \pi : f \in \left( R^{J(z)} \cap \ker A \right)^* \right\} = \{ L_\alpha \hat{\iota}_z : \alpha \in R^{J(z)} \cap \ker A \},$$

Equation (12) holds. \hfill \Box

Lemma 9  For $z \in \{0, 1\}^n$, let $\iota_z$ be the canonical injection from $\Theta_z := \{ c \in \Theta \mid Z(c) = z \}$ to $\Theta$. Then, the inclusion

$$\iota_z^{-1} \mathcal{O}^{**} \subset \iota_z^{-1} \sigma(R_\alpha : \alpha \in \ker A)$$

holds. Here, we put $\mathcal{O}^{**} := \{ B \in \mathcal{B}(\Theta) \mid G \cdot B = B \}$. (13)
Lemma 11 Let $\pi : \mathbb{Z}^d_{\geq 0} \times \Theta \to \Theta$ and $\pi' : \mathbb{Z}^d_{\geq 0} \times \Theta \to \mathbb{Z}^d_{\geq 0}$ be the projections. Put $O^* := \{B \in B(\mathbb{Z}^d_{\geq 0} \times \Theta)[G \cdot B = B]\}$. Then, the following equation holds:

$$O^* = \sigma(\pi', R_\alpha \circ \pi, Z \circ \pi; \alpha \in \ker A).$$  

(21)
Let $B \in O^*$. For $t \in \mathbb{Z}^d_{\geq 0}$, let $\iota_t : \Theta \to \mathbb{Z}^d_{\geq 0} \times \Theta$ be an inclusion map defined by $\iota_t(c) = (t, c)$, and put $B_t := B \cap \iota_t(\Theta)$. Then, $\iota_t^{-1} B_t$ is a Borel set of $\Theta$. Since the equation $G \cdot B = B$ implies $G \iota_t^{-1} B = \iota_t^{-1} B$, $\iota_t^{-1} B$ is an element of $O^*$. By Lemma 10 and Lemma 5 we have $\pi^{-1} \iota_t^{-1} B_t \in \pi^{-1} \sigma (R_\alpha, Z; \alpha \in ker A) = \sigma (R_\alpha \circ \pi, Z \circ \pi; \alpha \in ker A)$. Hence, $B_t = \pi^{-1} \iota_t^{-1} B_t \cap \iota_t(\Theta) \in \sigma (\pi', R_\alpha \circ \pi, Z \circ \pi; \alpha \in ker A)$ implies $B = \bigcup_{t \in \mathbb{Z}^d_{\geq 0}} B_t \in \sigma (\pi', R_\alpha \circ \pi, Z \circ \pi; \alpha \in ker A)$. We have (21).

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 the equation

$$O = \sigma (AX, R_\alpha(\theta), Z(\theta); \alpha \in ker A)$$

(22)

holds.

Proof By Lemma 5 and Lemma 11 we have

$$O = \hat{\theta}^{-1} O^* = \hat{\theta}^{-1} \sigma (\pi', R_\alpha \circ \pi, Z \circ \pi; \alpha \in ker A)$$

$$= \sigma (\pi'(\hat{\theta}), R_\alpha \circ \pi(\hat{\theta}), Z \circ \pi(\hat{\theta}); \alpha \in ker A) = \sigma (AX, R_\alpha(\theta), Z(\theta); \alpha \in ker A).$$

Theorem 5 The following equation holds:

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

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

To show Theorem 5 we prepare the following lemma:

Lemma 12 Let \( F : R^n \to R^d \) be a linear map and \( \iota : R^n_{>0} \to R^n \) be the inclusion. For \( \alpha \in \ker F \), let \( R_\alpha : R^n_{>0} \to R \) be a function defined by \( R_\alpha(x) := \prod_{i=1}^n x_i^{\alpha_i} \). Then, we have

\[
B(R^n_{>0}) = \sigma(F_\iota, R_\alpha; \alpha \in \ker F).
\]

Proof It is obvious that the left-hand side includes the right-hand side. We show the opposite inclusion. Let \( \{\alpha_1, \ldots, \alpha_m\} \) be a basis of \( \ker F \). Then the differential map

\[
\varphi : R^n_{>0} \to (\text{Im} F_\iota) \times R_m \quad (x \mapsto (F_\iota(x), R_{\alpha_1}(x), \ldots, R_{\alpha_m}(x)))
\]

is surjective. By the general theory of the exponential family (23 p. 125, 4 Theorem 3.6), \( \varphi \) is also injective. Hence, \( \varphi \) is a diffeomorphism between \( R^n_{>0} \) and \( (\text{Im} F_\iota) \times R_m \), and we have

\[
B(R^n_{>0}) = \varphi^{-1}B((\text{Im} F_\iota) \times R_m) = \varphi^{-1}\sigma(p, q_1, \ldots, q_m) = \sigma(p, q_1\varphi, \ldots, q_m\varphi)
\]

\[
= \sigma(F_\iota, R_{\alpha_1}(x), \ldots, R_{\alpha_m}(x) ; \alpha \in \ker F)
\]

Here, \( p : (\text{Im} F_\iota) \times R_m \) and \( q_i : (\text{Im} F_\iota) \times R_m \) are the projections. \( \square \)

Proof of Theorem 5 Recall that, for \( z \in \{0, 1\}^n \), we put \( \Theta_z = \{c \in \Theta | Z(c) = z\} \) and that \( \iota_z : \Theta_z \to \Theta \) is the inclusion. Applying Lemma 12 in the case where \( F = A\iota_z \), we have

\[
B(\Theta_z) = \sigma(A\iota_z, R_\alpha\iota_z; \alpha \in \ker A\iota_z).
\]

(23)

The equation \( \ker A\iota_z = R^{J(z)} \cap \ker A \) implies

\[
\sigma(A\iota_z, R_\alpha\iota_z; \alpha \in \ker A\iota_z) = \sigma(A\iota_z, R_\alpha\iota_z; \alpha \in R^{J(z)} \cap \ker A)
\]

(24)

By Equations (23) and (24), we have

\[
B(\Theta_z) = \sigma(A\iota_z, R_\alpha\iota_z; \alpha \in R^{J(z)} \cap \ker A) \subset \iota_z^{-1}\sigma(A, R_\alpha; \alpha \in R^{J(z)} \cap \ker A)
\]

\[
\subset \iota_z^{-1}\sigma(A, R_\alpha, Z; \alpha \in \ker A) \subset \sigma(A, R_\alpha, Z; \alpha \in \ker A).
\]

(25)

Any \( B \in B(\Theta) \) can be decomposed as \( B = \bigcup_{z \in \{0, 1\}^n} (B \cap \Theta_z) = \bigcup_{x \in \{0, 1\}^n} \iota_z^{-1}B \). By (26), \( B \) is an element of \( \sigma(A, R_\alpha, Z; \alpha \in \ker A) \). Hence, we have

\[
B(\Theta) = \sigma(A, R_\alpha, Z; \alpha \in \ker A)
\]

(26)

The \( \sigma \)-algebra generated by \( \theta \) is the pull-back of the left-hand side of (26) with respect to \( \theta \). By Lemma 5 the pull-back of the right-hand side of (26) equals to \( \sigma(A\theta, R_\alpha(\theta), Z(\theta); \alpha \in \ker A) \). Hence, we have

\[
\sigma(\theta) = \sigma(A\theta, R_\alpha(\theta), Z(\theta); \alpha \in \ker A).
\]

This equation implies

\[
\sigma(AX, \theta) = \sigma(AX, A\theta, R_\alpha(\theta), Z(\theta); \alpha \in \ker A)
\]

By Theorem 4, we have

\[
\sigma(A\theta, O) = \sigma(A\theta, \sigma(AX, R_\alpha(\theta), Z(\theta); \alpha \in \ker A)) = \sigma(A\theta, AX, R_\alpha(\theta), Z(\theta); \alpha \in \ker A) = \sigma(AX, \theta)
\]

(27)

\( \square \)
9 Examples of CMLE problems

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 9 We generate categorical data concerning the number of hours slept and time of going to bed from a student sample in the LearnBayes package\footnote{https://cran.r-project.org/web/packages/LearnBayes/index.html} 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 bed before midnight, between midnight and 1am, and after 1am. We wish to analyze these categorical data by the Poisson random model $U_{ij} \sim \text{Pois}(p_{ij})$. The independence of rows and columns is rejected by the $\chi^2$ 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.

Categorial 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 $\begin{pmatrix} 1 & 6 & 123 \\ 3 & 22 & 145 \\ 86 & 91 & 176 \end{pmatrix}$. Categorial data for males:

$\begin{pmatrix} 0 & 5 & 28 \\ 3 & 4 & 47 \\ 35 & 32 & 71 \end{pmatrix}$

Categorial data for females:

$\begin{pmatrix} 0 & 3 & 18 \\ 4 & 32 & 98 \\ 51 & 50 & 105 \end{pmatrix}$

Because this CMLE can be solved by the $A$-distribution discussed previously, we apply our algorithm for evaluating normalizing constants and their derivatives to the method for estimating the conditional maximum likelihood in \cite[§4]{26}. We obtain the following estimates. CMLE ($p_{ij}$) for all:

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

CMLE for males:

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

CMLE for females:

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

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 $g_3h_1 = g_3h_2 = g_3h_3 = 1$ and $g_1h_2 = g_2h_2 = g_3h_2 = 1$. Then, we have $g_ih_j = 1$ 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 $R_{\geq 0}^4$ by the quotient topology. The correspondence is given by

$$(p_{i,j}) \mapsto \begin{pmatrix} p_{11}p_{i2} & 1 & p_{13}p_{i2} \\ p_{21}p_{i2} & 1 & p_{23}p_{i2} \\ p_{31}p_{i2} & 1 & p_{33}p_{i2} \\ 1 & 1 & 1 \end{pmatrix} \eqno(27)$$

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 $p_{i,j}$ by the frequency $u_{i,j}$ in (27) is called a generalized odds ratio. Generalized odds ratios for our data are as follows. Odds ratios for all:

$$
\begin{pmatrix}
0.176356589147287 & 10.5994318181818 & 1.050779958677686 \\
0.144291754756871 & 3.40779958677686 & 1.050779958677686 \\
1 & 1 & 1
\end{pmatrix}
$$

Odds ratios for males:

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

Odds ratios for females:

$$
\begin{pmatrix}
0 & 1 & 13.3452380952381 & 3.05925925925926 \\
0.1928104571634 & 1 & 3.05925925925926 & 1 \\
1 & 1 & 1
\end{pmatrix}
$$

Note that, as proved in [26, 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 10 The categorical data below are taken from emergency safety information on diclofenac sodium for influenza encephalitis and encephalopathy.[3]

Categorical data:

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

We omit titles and express this table as $$\begin{pmatrix} 4 & 7 & 2 \\ 32 & 5 & 6 \end{pmatrix}$$. By applying our algorithm and the method in [26], we obtain the following CMLE.

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

Generalized odds ratios are

$$
\begin{pmatrix}
1 & 11 & 2.666666666666667 \\
1 & 1 & 1
\end{pmatrix}
$$

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:

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

[4]http://idsc.nih.go.jp/disease/influenza/iencepha.html
The decrement of a vector consisting of the hypergeometric series $S_{gtt}$ and the program $c$ vector. When $(F_1 0$, Appendix (and odds ratios: these notations.

Our algorithm outputs CMLE $M$. Hence, the matrix $M$ follows from [5, Theorem 5.3] that the representation matrix $M$. We start with the integral representation of $F(a) = M(a)F(a + 1)$, which is a special case of

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

in [5] Corollary 6.3 ($\alpha(2)$ stands for $a + 1$). The function $upAlpha(2,1,1)$ in the program derives $\frac{1}{\alpha} U_2$. $S(\alpha;x)$ is the vector consisting of the hypergeometric series $S(\alpha;x)$ defined in [5] Section 6] and its derivatives (Gauss-Manin vector). When $c \in \mathbb{N}_0$, it can be expressed in terms of $F_1$ as

$$S(\alpha;x) = \frac{S}{\alpha} \theta,S = \left(\begin{array}{cc}1 & 0 \\ 0 & 1/\alpha_2\end{array}\right) \left(\begin{array}{c}S \\ \theta,S\end{array}\right) = \frac{1}{(-\alpha)(-\beta)(c-1)!} \left(\begin{array}{cc}1 & 0 \\ 0 & 1/\alpha_2\end{array}\right) \left(\begin{array}{c}F_1 \\ \theta_2 F_1\end{array}\right).$$

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

$$M(a) = -a \left(\begin{array}{cc}1 & 0 \\ 0 & 1/\alpha_2\end{array}\right) \left(\begin{array}{cc}1 & 0 \\ 0 & 1/(\alpha_2 - 1)\end{array}\right) = \left(\begin{array}{cc}1 & 0 \\ 0 & 1/(\alpha_2 - 1)\end{array}\right) U_2(\alpha(2)) \left(\begin{array}{c}1 \\ 1\end{array}\right).$$

It follows from [5] 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 $|\tilde{x}(ij)|$, which is the determinant of the minor matrix consisting of the $i$-th column and the $j$-th column of the matrix $\tilde{x} = \left(\begin{array}{ccc}0 & 1 & x \\ 0 & 1 & 1\end{array}\right)$, where the numbering starts with 0 (see [5] as to details). We put $\varphi(ij) = \frac{|\tilde{x}(ij)|}{L_0 L_1}$, where $L_0 = 1, L_1 = t, L_2 = 1 + xt,$ and $L_3 = 1 + t$. We have the following expressions with these notations.

$$D_2(x) = \text{diag} \left(\begin{array}{ccc}|\tilde{x}(21)| & |\tilde{x}(23)| \\ |\tilde{x}(01)| & |\tilde{x}(03)|\end{array}\right) = \text{diag}(1, 1 - x) = \left(\begin{array}{cc}1 & 0 \\ 0 & 1 - x\end{array}\right),$$

$$C(\alpha) = \frac{I(\varphi(01), \varphi(01)) I(\varphi(02), \varphi(01)) I(\varphi(01), \varphi(02)) I(\varphi(02), \varphi(02))}{I(\varphi(02), \varphi(01)) I(\varphi(03), \varphi(01)) I(\varphi(01), \varphi(02)) I(\varphi(03), \varphi(02))} = 2\pi \sqrt{-1} \left(\begin{array}{cc}1 & 1 \\ \frac{1}{\alpha_0} & \frac{1}{\alpha_1}\end{array}\right),$$

$$Q_2(\alpha) = \frac{I(\varphi(01), \varphi(01)) I(\varphi(02), \varphi(01)) I(\varphi(01), \varphi(02)) I(\varphi(02), \varphi(02))}{I(\varphi(02), \varphi(01)) I(\varphi(03), \varphi(01)) I(\varphi(01), \varphi(02)) I(\varphi(03), \varphi(02))} = 2\pi \sqrt{-1} \left(\begin{array}{cc}1 & 1 \\ \frac{1}{\alpha_0} & \frac{1}{\alpha_1}\end{array}\right),$$

$$P_2(\alpha) = \frac{I(\varphi(21), \varphi(01)) I(\varphi(23), \varphi(01)) I(\varphi(22), \varphi(02)) I(\varphi(23), \varphi(02))}{I(\varphi(23), \varphi(01)) I(\varphi(23), \varphi(01)) I(\varphi(21), \varphi(02)) I(\varphi(23), \varphi(02))} = 2\pi \sqrt{-1} \left(\begin{array}{cc}1 & 1 \\ \frac{1}{\alpha_0} & \frac{1}{\alpha_1}\end{array}\right).$$

$^a\alpha_0 = -\alpha_1 - \alpha_2 - \alpha_3$ stands for the exponent at infinity.
where $\mathcal{I}$ is the intersection form on the twisted cohomology group. The inverse matrices of them can be also expressed in terms of intersection numbers as in [3, Appendix]. This method is implemented as the function \texttt{invintMatrix.k} in our package and it outputs

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

$$C(\alpha)^{-1} = \frac{1}{2\pi \sqrt{-1}} \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \begin{pmatrix} \frac{1}{\alpha_1} - \frac{1}{\alpha_3} & \alpha_1 \\ \alpha_1 & 0 \end{pmatrix} = \frac{1}{2\pi \sqrt{-1}} \begin{pmatrix} \alpha_1 - \alpha_1 \\ 0 & -\alpha_2 \end{pmatrix},$$

These matrices can be obtained in our program as

$$D_2(\alpha) = \text{repMatrix}(2, 1, 1), \quad C(\alpha)/(2\pi \sqrt{-1}) = \text{intMatrix}([0, 3], [0, 3], 1, 1),$$

$$P_2(\alpha)/(2\pi \sqrt{-1}) = \text{intMatrix}([2, 0], [0, 3], 1, 1), \quad Q_2(\alpha)/(-1) = \text{intMatrix}([0, 2], [0, 3], 1, 1),$$

$$(2\pi \sqrt{-1})P_2(\alpha)^{-1} = \text{invintMatrix.k}([2, 0], [0, 3], 1, 1), \quad (2\pi \sqrt{-1})C(\alpha)^{-1} = \text{invintMatrix.k}([0, 3], [0, 3], 1, 1)$$

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

\textbf{Acknowledgements:} This work was supported by MEXT/JSPS KAKENHI Grand Numbers JP 25220001, 17K05279, 18J01507 and by Research Institute for Mathematical Sciences, a Joint Usage/Research Center located in Kyoto University.

\section*{References}

[1] Agresti, A., Categorical Data Analysis, 3rd edition, Wiley Series in Probability and Statistics, Wiley, 2013.

[2] Amari, S., Information Geometry and Its Application, Applied Mathematical Sciences, Springer Japan, 2016.

[3] Billingsley, P., Probability and Measure, Wiley Series in Probability and Statistics, Wiley, 1995.

[4] Brown, L. D. Fundamentals of statistical exponential families with applications in statistical decision theory. \textit{Lecture Notes-Monograph Series}, 9:1–279, 1986.

[5] Goto, Y., Matsumoto, K., Pfaffian equations and contiguity relations of the hypergeometric function of type$(k + 1, k + n + 2)$ and their applications, \texttt{arXiv:1602.01637}, to appear in Funkcialaj Ekvacioj.

[6] Goto, Y., Intersection numbers of twisted cycles and cocycles for degenerate arrangements, \texttt{arXiv:1805.01714}.

[7] T.Hibi et al, Gröbner Bases : Statistics and Software Systems, Springer, 2013.

[8] Karatzas, I., and Shreve, S., Brownian motion and stochastic calculus, Graduate texts in mathematics, Springer-Verlag, 1988.

[9] Koyama, T., A new formulation of sufficient $\sigma$-algebra, in preparation.

[10] Landers, D., and Rogge, L., Minimal sufficient $\sigma$-fields and minimal sufficient statistics two counterexamples, The Annals of Mathematical Statistics 43, 6 (1972), 2045–2049.

[11] LINBOX: exact computational linear algebra. \texttt{http://www.linalg.org}

[12] Nakayama, H., Nishiyama, K., Noro, M., Ohara, K., Sei, T., Takayama, N., Takemura, A., Holonomic Gradient Descent and its Application to Fisher-Bingham Integral, Advances in Applied Mathematics 47 (2011), 639-658.
[13] M.Noro, N.Takayama, Risa/Asir Drill 2012, [http://www.math.kobe-u.ac.jp/Asir](http://www.math.kobe-u.ac.jp/Asir), 2012 (in Japanese).

[14] Noro, M., Yokoyama, K., A Modular Method to Compute the Rational Univariate Representation of Zero-Dimensional Ideals. Journal of Symbolic Computation 28/1(1999), 243-263.

[15] M.Noro, K.Yokoyama, Computation of Gröbner Basis — introduction to computational algebra, University of Tokyo press, 2003 (in Japanese).

[16] M.Ogawa, Algebraic Statistical Methods for Conditional Inference of Discrete Statistical Models, PhD thesis, University of Tokyo, 2015.

[17] Ohara, K., Takayama, N., Pfaffian Systems of A-Hypergeometric Systems II — Holonomic Gradient Method, arXiv:1505.02947.

[18] Oshima, T., Fractional Calculus of Weyl Algebra and Fuchsian Differential Equations, MSJ Memoirs 28, 2013.

[19] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., Numerical Recipes 3rd Edition: The Art of Scientific Computing, Cambridge University Press, 2007.

[20] References for HGM [http://www.math.kobe-u.ac.jp/OpenXM/Math/hgm/ref-hgm.html](http://www.math.kobe-u.ac.jp/OpenXM/Math/hgm/ref-hgm.html)

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

[22] Sasaki, T., Takeshima, T., A modular method for Gröbner-basis construction over Q and solving system of algebraic equations. Journal of Information Processing, 12(1989), no.4, 371-379.

[23] Stoer, J., Bulirsch, R., Introduction to Numerical Analysis, Springer-Verlag, New York-Heidelberg, 1980.

[24] Tachibana, Y., Goto, Y., Takayama, N., An implementation of the difference holonomic gradient method to two way contingency tables, RIMS Kokyuroku, 2054(2017), 105–117.

[25] Takayama, N., Gröbner basis and the problem of contiguous relations, Japan Journal of Applied Mathematics, 6 (1989), 147-160.

[26] Takayama, T., Kuriki, S., Takemura, A., A-hypergeometric distributions and newton polytopes, Advances in Applied Mathematics 99 (2018), 109 – 133.

[27] Williams, D., Probability with Martingales, Cambridge mathematical textbooks, Cambridge University Press, 1991.