# Contingency tables and hypergeometric polynomials associated to hyperplane arrangements

Nobuki Takayama, joint work with Y.Tachibana, Y.Goto, T.Koyama

2018/06/11

1. Yoshihito Tachibana, Yoshiaki Goto, Tamio Koyama, Nobuki Takayama, Holonomic Gradient Method for Two Way Contingency Tables, arxiv:1803.04170
2. Y.Goto, K.Matsumoto, Pfaffian equations and contiguity relations of the hypergeometric function of type (k+1,k+n+2) and their applications, arxiv:1602.01637

$I = (I_1, \ldots, I_{k+1}) \in \mathbf{Z}_{\geq 0}^{k+1}$, $J = (J_1, \ldots, J_{n+1}) \in \mathbf{Z}_{\geq 0}^{n+1}$,
$\sum I_i = \sum J_j$.
$p = (p_{ij})$

$$Z(I, J; p) = \mathrm{C.T.} \prod_{j=1}^{n+1} \left( \sum_{i=1}^{k+1} p_{ij} t_i \right)^{J_j} t^{-I} \tag{1}$$

$t_1 = 1$, $t^{-I} = \prod_{i=1}^{k+1} t_i^{-I_i}$. Note* that

$$Z(I, J; p) = J! \sum \frac{p^u}{u!}$$

where $\sum_i u_{ij} = J_j$ (column sum is $J$), $\sum_j u_{ij} = I_i$ (raw sum is $I$) †.
$Z$ is the normalizing constant (partition function) of a distribution.

*C.T. is the constant term w.r.t. $t$. $J! = \prod_j J_j!$

†We denote these conditions or the left hand sides of them by ☺or ☺$_u$.

Goal 1: Evaluate numerically $Z$ and its derivatives efficiently and accurately [‡].

Motivation from statistics: 2 way contingency table:
$(k+1) \times (n+1)$ matrix with $\mathbf{Z}_{\geq 0}$ entries.

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

$$P(U_{ij} = u_{ij}) = \frac{\exp(-p_{ij}) p_{ij}^{u_{ij}}}{u_{ij}!}$$

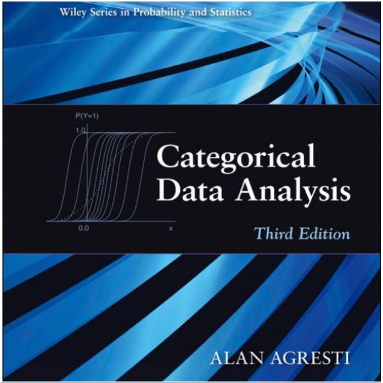
The conditional probability [§] when the row and column sums are fixed to $I$, $J$ is

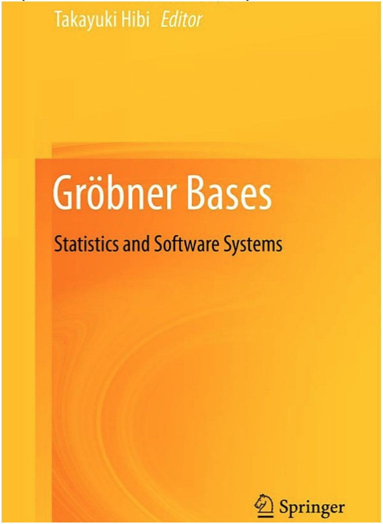
$$P\left( U = u \,\middle|\, \sum_j U_{ij} = I_i, \sum_i U_{ij} = J_j \right) = \frac{p^u / u!}{Z(I, J; p)}$$

[‡] When $I = (4, 14, 5, 2, 1)$, $J = (10, 6, 5, 2, 3)$, there are 229, 174 terms.

[§] $U_{ij}$ is a random variable of the Poisson distribution.

References on contingency tables (MSC2010: 62H17).



$$E[U_{ij}|☺] = \sum_{☺} \frac{u_{ij} p^u / u!}{Z(I, J; p)} = p_{ij} \frac{\partial}{\partial p_{ij}} \log Z \qquad (2)$$

**Proposition**

*$E[U_{ij}|☺_U]$ is invariant by the torus action $p_{ij} \mapsto p_{ij} p_i p'_j$, $p_i, p'_j \in \mathbf{R}_{>0}$.*

**Theorem**

$$\mathbf{R}_{>0}^{(k+1)(n+1)} / \sim \ni (p_{ij}) \mapsto E[U_{ij}|☺] \in \mathrm{relint\,New}(Z) \qquad (3)$$

*is an isomorphism* [¶].

Goal 2: Find the inverse image numerically [‖].

---

[¶] $\sim$ is the equivalence relation w.r.t. the torus action. This theorem is a special case of Th. 1 of N.Takayama, S.Kuriki, A.Takemura, A-Hypergeometric Distributions and Newton Polytopes, Advances in Applied Mathematics 99 (2018) 109–133.

[‖] conditional maximal likelihood estimation (CMLE).

Let us explain the idea of our method ** for $2 \times 2$ case.

$\bar{u} = \begin{pmatrix} J_1 & 0 \\ J_2 - I_1 & I_2 \end{pmatrix}$.

$$Z = \frac{p^{\bar{u}}}{\bar{u}!} {}_2F_1\left(-J_1, -I_2, J_2 - I_2 + 1; \frac{p_{12}p_{21}}{p_{11}p_{22}}\right) \tag{4}$$

$f(a) = {}_2F_1(a, b, c; x) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k (1)_k} x^k$,
$(a)_k = a(a+1)\cdots(a+k-1)$. $F(a) = (f(a), xdf/dx(a))^T$.

$$F(a) = (E + A(a)/a)^{-1} F(a+1), \tag{5}$$

where $A(a) = \begin{pmatrix} 0 & 1 \\ abx/(1-x) & (ax + bx - c + 1)/(1-x) \end{pmatrix}$.

$$F(a) = M(a)M(a+1)\cdots M(-2)F(-1), \quad M(a) = (E + A(a)/a)^{-1}.$$

"factorial" of contiguity relation (5).

**holonomic gradient method, HGM

$$\tilde{p} = \begin{array}{c} \\ 1 \\ 2 \\ \\ k+1 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & & & k+1 & k+2 & k+3 & & k+n+2 \end{array} \\ \begin{pmatrix} 1 & 0 & \cdots & 0 & p_{11} & p_{12} & \cdots & p_{1,n+1} \\ 0 & 1 & \cdots & 0 & p_{21} & p_{22} & \cdots & p_{2,n+1} \\ & & \cdots & & & & \cdots & \\ 0 & 0 & \cdots & 1 & p_{k+1,1} & & \cdots & p_{k+1,n+1} \end{pmatrix} \end{array}$$

$L_j = \tilde{p}_j \cdot t$ where $\tilde{p}_j$ is the $j$-th column vector [††] of $\tilde{p}$. $\alpha_j \in \mathbf{C} \setminus \mathbf{Z}$, $\sum_{j=1}^{k+n+2} \alpha_j = 0$.

$$\nabla = d_t + \sum_j \alpha_j d_t \log L_j \tag{6}$$

$$\tilde{P} = \{\tilde{p} \mid \text{ any } (k+1) \times (n+1) \text{ minor of } \tilde{p} \neq 0\}$$

$$T_p = \{t' \in \mathbf{C}^k \mid L_j(p;t) \neq 0 \text{ for all } j.\}, \quad p \in \tilde{P}$$

[††] $L_1 = t_1 = 1, L_2 = t_2, \ldots, L_{k+n+2} = \sum_i p_{i,n+1} t_i$

$$\mathcal{J}=\{(j_1,\ldots,j_{k+1})\,|\,1\le j_1<j_2<\cdots<j_{k+1}\le k+n+2\}$$

$$_q\mathcal{J}_p=\{J\in\mathcal{J}\,|\,q\notin J, p\in J\}$$

$$\varphi\langle J\rangle = d_t\log(L_{j_2}/L_{j_1})\wedge d_t\log(L_{j_3}/L_{j_1})\wedge\cdots\wedge d_t\log(L_{j_{k+1}}/L_{j_1}) \quad (7)$$

**Theorem (Goto, Matsumoto(2016) *, contiguity relation)**

*Put* $F=(\varphi\langle J\rangle\,|\quad J\in{}_{k+n+2}\mathcal{J}_1)$ *and assume* $\tilde{p}\in\tilde{P}$ *Then,*

$$L_iF\equiv\left(CP_i^{-1}D_iQ_iC^{-1}\right)F \quad \text{in} \quad H^k(\Omega^\bullet(T_p),\nabla) \qquad (8)$$

*where* $C, P_i, Q_i$ *are intersection matrices* † *among* $\varphi\langle J\rangle$ *and* $D_i$ *is a diagonal matrix with rational function entries of* $p$*.*

---

*Y.Goto, K.Matsumoto, Pfaffian equations and contiguity relations of the hypergeometric function of type (k+1,k+n+2) and their applications, arxiv:1602.01637, to appear in Funkcialaj Ekvacioj.

† $H^k(\mathcal{E}_0^\bullet(T_p),\nabla^\vee)\times H^k(\Omega^\bullet(T_p),\nabla)\to\mathbf{C}$ is called the intersection form. Note that $H^k(\mathcal{E}_0^\bullet(T_p),\nabla^\vee)\simeq H^k(\Omega^\bullet(T_p),\nabla^\vee)$

Integrating the both sides of (8) with $\int_\Delta \prod L_j^{\alpha_j} \cdot$, we have a contiguity relation for $Z$.

**Theorem**

*Fix $k$. Then the complexity of constructing the contiguity relation is $O(n^{3(k+1)})$.*

How do we evaluate efficiently $M(a)M(a-1)\cdots M(-1)$?
$\Rightarrow$ the modular method in computer algebra; evaluate in $\mathbf{Z}/s\mathbf{Z}$ for several prime numbers $s$ and reconstruct the answer in $\mathbf{Q}$ by the Chinese remainder theorem.

**Theorem**

*Let $n$ be the number of linear transformations and put $r = \binom{k+n}{k}$ [‡]. The complexity of the modular method is $\max(O(|J|), O(r))$, $|J| = \sum J_j$.*

(Numerical evidences.)

[‡] the rank of the twisted cohomology group.

Have we solved two goals? $\Rightarrow$ Not completely. We have assumed that $\tilde{p} \in \tilde{P}$ [§].

## Proposition

*Let $\beta_1$ be the total degree of $Z$ and $L$ a generic line in p-space. If we evaluate $E[U_{ij}]$[¶] at $2\beta_1$ points $p \in \mathbf{R}_{>0}^{(k+1)\times(n+1)}$ on a line $L$, then the exact value of $E[U_{ij}]$ can be obtained at any point on $L$.*

However, this method is not efficient $\Rightarrow$ open questions [‖] for hyperplane arrangements of the case that some of $p_{ij} = 0$.

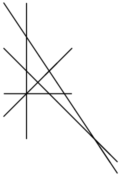


Figure: $V(t_2 t_3 (p_{21} t_2 + p_{31} t_3) \prod_{j=2}^{3} (p_{1j} + p_{2j} t_2 + p_{3j} t_3))$

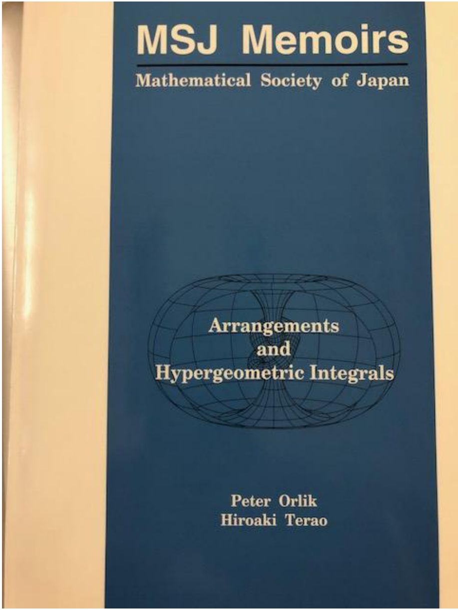
---

[§] This is the condition that hyperplane arranement is in a generic position.

[¶] We denote $E[U_{ij} | \circledcirc_U]$ by $E[U_{ij}]$.

[‖] Y.Goto, 1805.01714

This book will help to solve the open question of constructing contiguity relations efficiently for any hyperplane arrangement.



What is the space

$$\mathbf{R}_{\geq 0}^{(k+1)(n+1)} / \sim$$

It is not a manifold!

1. Algraic geometry: Related to the Chow quotient by M.Kapranov (1992).

2. Measure theoretic (statistic).

$U_{ij} : \Omega_{ij} \to \mathbf{Z}$, $P(U_{ij} = u_{ij}) = \exp(-\theta_{ij})\theta_{ij}^{u_{ij}} / u_{ij}!$.
$\Omega = \prod \Omega_{ij} \times \Theta$, $\Theta = \{(\theta_{ij}) \,|\, \theta_{ij} \in \mathbf{R}_{\geq 0}\}$.

$$\mathcal{O} = \sigma\left( ☺_U, \frac{\theta_{ij}\theta_{k\ell}}{\theta_{i\ell}\theta_{kj}}, Z_{ij}(\theta) \right).^{**} \tag{9}$$

where $Z_{ij}(c) = 1$ when $c_{ij} > 0$ and $= 0$ when $c_{ij} = 0$.

## Theorem

$$E[X|\sigma(☺_U, \theta)] = E[X|\sigma(☺_\theta, \mathcal{O})] = E[X|\mathcal{O}]$$

*for any* $X \in \mathcal{L}^1(\sigma(U))$ $^{\dagger\dagger}$.

$^{**}\sigma(Y)$ is the $\sigma$ algebra generated by $Y^{-1}(B)$. $\mathcal{O}$ is "of interest".

$^{\dagger\dagger}$cf. $E[U_{ij}|☺_U]$ is invariant by the torus action. $☺_\theta$ is nuissance.

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 |

Categorical data for males $\begin{pmatrix} 1 & 2 & 28 \\ 0 & 4 & 47 \\ 35 & 32 & 71 \end{pmatrix}$.

Categorical data for females $\begin{pmatrix} 0 & 4 & 95 \\ 3 & 18 & 98 \\ 51 & 59 & 105 \end{pmatrix}$.

CMLE for males: $\begin{pmatrix} 0.458167657900967 & 1 & \underline{6.25676090279981} \\ 0 & 1 & \underline{5.25200491199345} \\ 1 & 1 & 1 \end{pmatrix}$.

CMLE for females:
$\begin{pmatrix} 0 & 1 & \underline{13.2714773737657} \\ 0.193351042187373 & 1 & \underline{3.04872586155291} \\ 1 & 1 & 1 \end{pmatrix}$.

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

‡‡

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

Generalized odds ratios: $\begin{pmatrix} 1 & \underline{11.2} & 2.66666666666667 \\ 1 & 1 & 1 \end{pmatrix}$.

$11.2 = \frac{32\times 7}{4\times 5}$.

‡‡ Data of the previous page https://cran.r-project.org/web/packages/LearnBayes/index.html. Data of this papge https://www.pmda.go.jp/files/000148557.pdf

Summary:

1. Exact numerical evaluation of the hypergeometric polynomial $Z$ can be done efficiently with contiguity relations and the modular method.
2. It has applications to conditional maximal likelihood estimation (CMLE) for two way contingency tables.

Future challenge: Each column is death month, each row is birth month *.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 |
| 1 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 1 | 3 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 2 |
| 0 | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 2 |
| 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 1 | 1 | 2 | 0 | 0 | 2 | 0 | 1 | 1 | 0 |
| 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

* Diaconis, Sturmfels (1998), Algebraic algorithms for sampling from conditional distributions. Andrews, Herzberg (1985), Data, Springer, page 429.