A-超幾何多項式の近似公式と分割表統計

高山信毅 (神戸大学) (栗木, 竹村との共同研究)

arxiv: 1510.02269

hgm OpenXM search.

整数成分の $d \times n$ 行列 $A = (a_{ij})$. A の第 i 列ベクトル a_i を $\mathbb{Z}^d$ の元とみたとき, この列ベクトル達は $\mathbb{Z}^d$ を生成していると仮定する. ある行 i について $a_{ij} > 0$ と仮定.

$\beta \in \mathbf{N}_0 A = \mathbf{N}_0 a_1 + \cdots + \mathbf{N}_0 a_n$ に対して多項式

$$Z_A(\beta; p) = \sum_{Au=\beta, u \in \mathbf{N}_0^d} \frac{p^u}{u!} \tag{1}$$

を A-超幾何多項式とよぶ. $u! = \prod_{i=1}^n u_i!$, $p^u = \prod_{i=1}^n p_i^{u_i}$.

A の行が生成する $\mathbf{Z}$-加群が $(1, 1, \ldots, 1)$ を含む. $p > 0$, $\beta \in \mathbf{N}^d$ を固定. 列ベクトルの集合 $\{\bar{a}_i^T\}$ は $\mathrm{Ker}\,(A : \mathbf{Z}^n \to \mathbf{Z}^d)$ の基底. 横ベクトル $\bar{a}_i$ をならべて行列 $\bar{A}^T$ を作る. $Am = \beta$, $m^{\bar{a}_i} = p^{\bar{a}_i}$ を満す $m \in \mathbf{R}^n_{>0}$ がただひとつ存在する (ips).

Theorem

$k \to +\infty$ で,

$$Z(k\beta; p) \sim \frac{p^{km}}{\Gamma(km+1)} \frac{(2\pi k)^{(n-d)/2}}{\left(\det(\bar{A}M^{-1}\bar{A}^T)\right)^{1/2}} \tag{2}$$

$M = \mathrm{diag}(m)$.

(証明は初等的な確率論, 定理 3 に帰着. MLE(maximal likelihood estimation) 問題の解の副産物).
$P(U = u) = \frac{p^u}{u!}/Z_A(\beta; p)$, $u \in \mathbb{N}_0^d$ は $Au = \beta$ の上の確率分布. p はパラメーター.

$$Z_A(\beta;p)=\sum_{Au=\beta,u\in\mathbf{N}_0^d}\frac{p^u}{u!}$$

例 22(2×2 分割表): $A=\begin{pmatrix}0&0&1&1\\1&0&1&0\\0&1&0&1\end{pmatrix}$, $\beta=(37,36,12)^T$. u を次の形式で書く: $\begin{pmatrix}u_1&u_2\\u_3&u_4\end{pmatrix}$. $Au=\beta$ を満たす u 達は

$$u=\begin{pmatrix}11&0\\25&12\end{pmatrix},\ldots,u=\begin{pmatrix}4&7\\32&5\end{pmatrix},\ldots,u=\begin{pmatrix}0&11\\36&1\end{pmatrix}.$$

$Z_A(\beta;p)=\frac{p_1^{11}p_3^{25}p_4^{12}}{11!25!12!}{}_2F_1\left(-11,-12,26;y\right),y=\frac{p_2p_3}{p_1p_4}.,$

$\mathrm{Ker}\,A=\mathbb{Z}\begin{pmatrix}-1&1\\1&-1\end{pmatrix}$, $\frac{m_2m_3}{m_1m_4}=\frac{p_2p_3}{p_1p_4}, Am=\beta.$

$Am = \beta$, $\frac{m_2 m_3}{m_1 m_4} = \frac{p_2 p_3}{p_1 p_4}$. $p_i > 0$, $m_i > 0$, $\beta \in \mathbb{N}^3$, $\beta_1 - \beta_3 \geq 0$, $\beta_2 + \beta_3 - \beta_1 > 0$.

$$\frac{p_1^{k(\beta_2+\beta_3-\beta_1)} p_3^{k(\beta_1-\beta_3)} p_4^{k\beta_3}}{(k(\beta_2+\beta_3-\beta_1))!(k(\beta_1-\beta_3))!(k\beta_3)!}$$

$$\cdot {}_2F_1\left(-k(\beta_2+\beta_3-\beta_1), -k\beta_3, k(\beta_1-\beta_3)+1; \frac{p_2 p_3}{p_1 p_4}\right)$$

$$\sim \quad \frac{p^{km}}{\prod \Gamma(km_i+1)} \frac{2\pi k}{\sqrt{\sum \frac{1}{m_i}}} \quad \text{(J.Cornfield, 1959)}$$

$\beta_2 + \beta_3 - \beta_1$	$-a$	0
β_1	$c-1$	$-b$
	β_2	β_3

$p(\xi) = (\exp \xi_1, \ldots, \exp \xi_n)$, $\psi(\xi) = \log Z(\beta; p(\xi))$ とおく.

$$E[U_i] = \frac{p_i \partial_i \bullet Z}{Z}|_{p=p(\xi)} = \frac{\partial \psi(\xi)}{\partial \xi_i},$$

ここで $\partial_i = \frac{\partial}{\partial p_i}$.
性質 (不変性). $\xi - \xi' \in \mathrm{Im}\, A^T$ の時. $E[U](\xi) = E[U](\xi')$.

情報幾何 (参考: 甘利, 情報幾何学の新展開, 2014, サイエンス社) の記号法を使い $\eta_i = E[U_i]$, $\eta = (\eta_i)$ とおく. 情報幾何の考え方では ξ-空間と η-空間は双対空間でありこれらを結ぶ写像が *moment map* $E[U]$ である.
MLE 問題とは: $E[U]$ の逆像計算 (β は fix).
$E[U]$ の順像計算も計算量の壁があり近似が主に研究されてきたのが before HGM.

歴史

1. M.Saito, B.Sturmfels, N.Takayama, Hypergeometric polynomials and Integer Programming, Compositio Mathematica, 115 (1999), 185–204. 隣接関係式. 特性多項式の多面体による特徴づけ.
2. Holonomic Gradient Method (HGM) の登場 (2011–). 正規化定数 Z およびその微分の近似計算、または正確計算. HGM=漸化式 (差分方程式) または微分方程式による Z やその微分の数値評価.
3. M.Ogawa, A.Takemura, N.Takayama, An Application of A-hypergeometric Equations to Conditional Maximal Likelihood Estimation of $2 \times m$ Contingency Tables, in preparation. 小川 D 論 (2015/3 月) より.
4. Y.Goto, Contiguity Relations of Lauricella's F_D Revisited, arxiv:1412.3256. F_D と $2 \times m$ 分割表が対応.
5. K.Ohara, N.Takayama, Pfaffian Systems of A-Hypergeometric Systems II — Holonomic Gradient Method, arxiv:1505.02947. A-超幾何多項式の数値評価アルゴリズム. Macaulay 型行列.

問題: (1) MLE の可解条件 (2) MLE の効率的解法 (分割表, 一般の場合).

R ($\mathbb{Z}^n$ の unimodular 行列), S ($\mathbb{Z}^d$ の unimodular 行列) をうまく選べば

$$SAR = \begin{pmatrix} \alpha_1 & & O \\ & \ddots & & O \\ O & & \alpha_d \end{pmatrix}, \quad \alpha_i \neq 0,\ \alpha_i | \alpha_{i+1}.$$

$\mathrm{Ker}(A : \mathbb{Z}^n \to \mathbb{Z}^d)$ の $\mathbb{Z}$-加群としての基底は $\{Re_{d+1}, \ldots, Re_n\}$, $(Re_{d+i})^T$ を $\bar{a}_i$ と書いて,

$$\bar{A} = \begin{pmatrix} \bar{a}_1 \\ \vdots \\ \bar{a}_{n-d} \end{pmatrix}_{(n-d)\times n},$$

$\lambda = \bar{A}\xi$ とおく. $\exp(\lambda_i) = p^{\bar{a}_i} = \prod_{j=1}^n p_j^{\bar{a}_{ij}}$ を一般化 odds 比とよぶ.

例 22: $\bar{A} = (-1, 1, 1, -1)$, $\frac{p_2 p_3}{p_1 p_4}$ (元祖 odds 比=一般化 odds 比)

Theorem

Newton 多面体 $\mathrm{New}(Z)$ *(*ただし Z は ξ の式でなく p の多項式とみなす*)* の次元が $n-d$ と仮定する. このとき, *(*一般化 *odds* 比の log の空間で考えた*) moment map* は次の同型を与える.

$$E[U] \ : \ \mathbb{R}^n/\mathrm{Im}\, A^T \longrightarrow \mathrm{relint}(\mathrm{New}(Z)) \subset \mathbb{R}^n$$

なお "relint" は相対的内部を意味する.
証明は,

$$f(\xi) = \eta \cdot \xi - \log Z(\beta, p(\xi))$$

の最大化.
例 22: $25 < E[U_{21}] < 36$.
(1) この定理は MLE の可解条件の答え. 境界の対応は複雑.

次の (k を増やして β 部をどんどん大きくしていく) 確率分布の列を考える.

$$P_k(u,\xi) = \frac{\exp(u\cdot\xi)}{u!Z_k(\xi)}, \quad u \in S_k, \quad k = 1,2,\ldots,$$

ここで

$$S_k = \{u \in \mathbb{N}^n \,|\, Au = k\beta\}, \quad Z_k(\xi) = \sum_{u\in S_k} \frac{\exp(u\cdot\xi)}{u!},$$

問題: $k \to \infty$ の時これはどのような分布になるか?

1. J.Cornfield, A Statistical Problem Arising from Retrospective Studies, Proceedings of 3rd Berkeley Symposium on Mathematical Statistics and Probability 4 (1956), 135–148.
2. 広津, 離散データ解析, 教育出版, 1982.
3. R.L.Plackett, Analysis of Categorical Data, 2nd ed, Griffin, 1981. (pp. 41 (2×2 table), pp. 65–66 ($r\times s$ table))

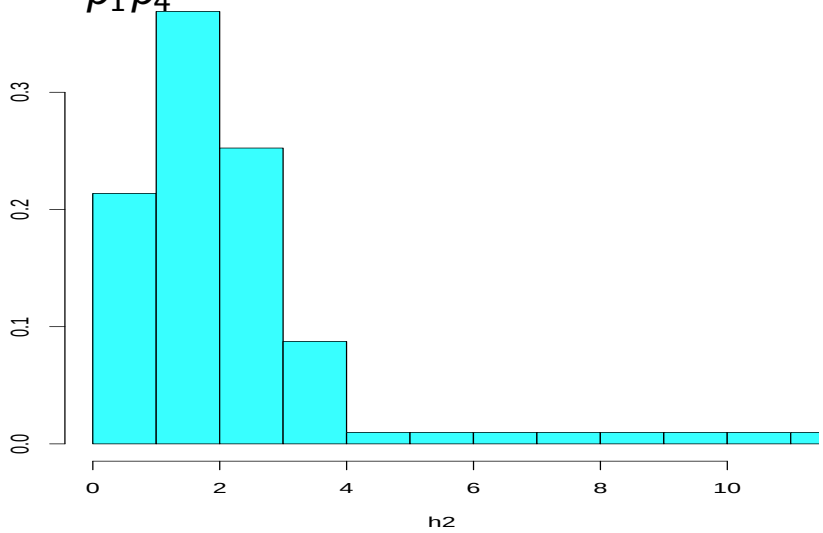
問題: $k \to \infty$ の時 $\frac{p^u/u!}{Z(k\beta;p)}$ はどのような分布になるか?

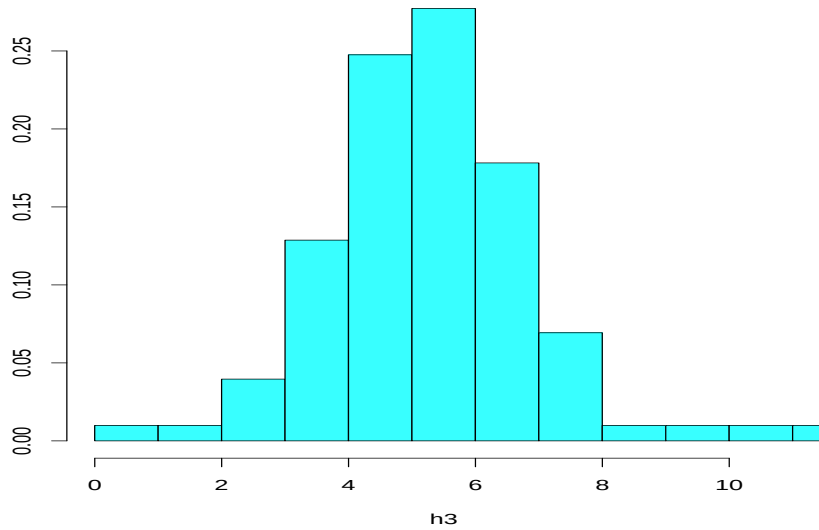
$\lambda = \bar{A}\xi$, $\exp(\xi) = p$ とおく. $m = m(\lambda)$ を次の連立代数方程式系の (一意的な正の実数) 解とする (解法は伝統的に IPS (iterative proportional scaling) と呼ばれている).

$$\begin{cases} \beta = Am, \\ \lambda = \bar{A}\log m. \end{cases} \tag{3}$$

答: km を平均とする正規分布に近づく.

例 22, $\frac{p_2p_3}{p_1p_4} = 1/3$ と 3 の時の分布.





Theorem

$\beta \in \mathbb{N}_0 A \cap \mathrm{int}(\mathbb{R}_{\geq 0} A)$ と仮定する. $M = \mathrm{diag}(m_i)$.

$$\widehat{P}_k(u,\xi) = \frac{\det(\bar{A}M^{-1}\bar{A}^T)^{1/2}}{(2\pi k)^{(n-d)/2}} \exp\left(-\sum_{i=1}^{n} \frac{(u_i - km_i)^2}{2km_i}\right)$$

と置くとき,

$$\sup_{\forall i\, |u_i - km_i| < \varphi(k)} \left| \frac{P_k(u,\xi)}{\widehat{P}_k(u,\xi)} - 1 \right| \to 0 \quad (k \to \infty),$$

ここで $\varphi(k)$ は次を満たす正の値をもつ関数: $\varphi(k) = o(k^{2/3})$, $k/\varphi(k)^2 = o(1)$.

証明のスケッチ: u の値が集中するはずの km との差を $v = u - km$ とおく. 次の確率の比を考える. 分母が一番頻度が多いと予想される km の確率.

$$\begin{aligned}\log \frac{P_k(u,\xi)}{P_k(km,\xi)} &= (u-km)\cdot\xi - \log\frac{u!}{(km)!} \\ &= v\cdot\xi - \log\frac{(km+v)!}{(km)!}.\end{aligned}$$

$(km)! = \prod_{i=1}^{n}\Gamma(km_i+1)$. 次のスターリングの公式で差を評価.

$$\log u! = u(\log u - 1) + \frac{1}{2}\log(2\pi u) + R(u), \tag{4}$$

ここで $u\to\infty$ の時 $R(u) = o(1)$. $H(v) = (1+v)\log(1+v) - v$ とおく.

$$\begin{aligned}\log\frac{(km_i+v_i)!}{(km_i)!} =& v_i\log(km_i) + (km_i)H\left(\frac{v_i}{km_i}\right) + \frac{1}{2}\log\left(1+\frac{v_i}{km_i}\right) \\ &+ R(km_i+v_i) - R(km_i)\end{aligned} \tag{5}$$

for $km_i + v_i \geq 1$,

$$\sum_{i=1}^{n} v_i = 0. \tag{6}$$

$$\sum_{i=1}^{n} v_i \log m_i = \sum_{i=1}^{n} v_i \xi_i. \tag{7}$$

$H(v)$ の Taylor 展開

$$H(v) = \frac{1}{2}v^2 - \frac{v^3}{6(1+\theta v)^2}, \quad 0 < \theta < 1. \tag{8}$$

(6), (7), (8) を (5) に代入し i について和をとると

$$\log \frac{P_k(u,\xi)}{P_k(km,\xi)} = -\sum_{i=1}^{n} \frac{v_i^2}{2km_i} + \sum_{i=1}^{n} \left\{ \frac{v_i^3/(km_i)^2}{6(1+\theta_i \frac{v_i}{km_i})^2} - \frac{1}{2}\log\left(1+\frac{v_i}{km_i}\right) + o(1) \right\}$$

ここで $0 < \theta_i < 1$. $v_i = o(k^{2/3})$ の時 3 次以降は消える.

例: 2×4 分割表. row sums $(4, 19)$, column sums $(9, 5, 3, 6)$.
$p = (1, 1/3, 12, 1/5001, 1, 1, 1, 1)$ (p_4 はとても小さい意地悪な例).
m (by IPS) は

$$(2.79518, 0.652785, 0.551505, 0.000540425, 6.20482, 4.34722, 2.4485, 5.99946).$$

後藤 F_D HGM による期待値の正確値

$$(2.83214, 0.627808, 0.539555, 0.000496547, 6.16786, 4.37219, 2.46044, 5.9995).$$

m と正確値の比

$$(0.98695, 1.03978, 1.02215, 1.08837, 1.00599, 0.994289, 0.995147, 0.999993).$$

k	$\log Z$	Approx by Th 2	$\lvert\text{error}\rvert$
9	-568.0127	-569.8179	1.8052
200	-26598.4556	-26598.9446	0.4890
300	-42685.5415	-42685.9149	0.3734

ただしこの近似公式で確率を計算するのは危険: Table
$u = (33, 1, 1, 1, 48, 44, 26, 53)$ を $k = 9$ の場合に考える. u を得る
正確な確率は 3.26465×10^{-7}. しかし近似公式による値は
1.98529×10^{-6} (6 倍).

(2) MLE の効率的解法.
$u \in \mathbb{N}^n$ が観測されたデータ. MLE は

$$\mathrm{maxarg}_\xi \ \frac{\exp(\xi)^u}{u!Z(\beta;\exp(\xi))}, \quad \beta = Au.$$

と解くこと. つまり, $\beta = Au$ と決めて, $E[U]^{-1}(u)$ をもとめる. (u を観測する一番もっともらしい p を求めること.)
アルゴリズム. 簡単のため,

$$\mathbb{R}_{>0} \ni y = (p_{d+1}, \ldots, p_n) \ \mapsto \ E(y) = (E[U_{d+1}], \ldots, E[U_n]) \in \mathrm{relint}(\mathrm{New}(Z)) \cap \mathbb{R}^{n-d}$$

は $p_1, \ldots, p_d$ を固定したとき 1 対 1 とする. $\eta^* = u$ の後ろの $n-d$ 個の成分 (観測値) とおく.

1. $p = u/|u|$ ($|u| = u_1 + \cdots + u_n$) とおく. (近似定理 3 より $E[U](p) \simeq m$ なのでこれは答えに近いと仮定できる.) よって y をこの p の $d+1$ 成分から n 成分をとりだしたもの.
2. 以下を十分繰り返す.

$$\text{new } y = y + \varepsilon \dot{E}(y)^{-1}(\eta^* - \eta), \tag{9}$$

ここで $\dot{E}(y) = \left(\frac{\partial E[U_{d+i}]}{\partial p_{d+j}} \right)_{i,j=1,\ldots,n-d}$, $\eta = E(y)$, ε は十分小さい.

$\dot{E}^{-1}$ の数値評価は HGM を用いる. 一般は [OT].

最近の進展 ($\dot{E}(y)$ の厳密数値計算): $r_1 \times r_2$ contingency tables は 松本-後藤. modular method による高速化実装は,

後藤-橘-高山.

例.

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

$2 \times 2 \times 2$ 分割表.

	p_1	p_2
	0	p_3

	p_4	p_5
	p_6	p_7

,

周辺和は “平面” でとる. $p_4 + p_5 + p_6 + p_7$, $p_1 + p_4 + p_6$, $p_2 + p_3 + p_5 + p_7$, $p_1 + p_2 + p_4 + p_5$.
次の観測データを与えるもっともらしい cell の確率 $p = \exp(\xi)$ は? (MLE, 逆像).

	19	132
	0	9

	11	52
	6	97

p_1, p_2, p_3, p_4 を固定. (p_5, p_6, p_7) を動かす. $\eta = (E[U_5], E[U_6], E[U_7])$

p	η
$P_0 = (19, 132, 9, 11, 52, 6, 97)/326$ 第一近似	$(51.9194, 5.99193, 97.0891)$
$P_0 + (0, 0, 0, 0, h_1, h_2, h_3)$ ここで $h = (0.000256154, -0.000152585, -0.00310983)$	$(52.0006, 6.00006, 96.9993)$