空間疫学のp値計算のための逐次計算

栗木　哲　　数理・推論研究系　教授

1　スキャン統計量

ある国にn個の地域$V=\{1,\dots,n\}$があるとする．各地域毎に，サイズ (例えば患者の期待度数) を表すパラメータλ_iが利用可能とする．X_iを地域iにおけるイベント (例えば患者) の発生数とし，確率モデル

$$X_i \sim \mathrm{Poisson}(\theta_i\lambda_i) \quad (\text{独立に})$$

を想定する．ここでθ_iは SMR (Standardized Mortality Ratio) とよばれる．空間疫学では，ある地域クラスターではそれ以外の地域よりも SMR が大きいという状況を想定し，その地域クラスターをホットスポットとよぶ．

ホットスポットを検出するためには，考慮すべき地域クラスタの候補 (スキャンウィンドウ) $\mathcal{B}\subset 2^V$を考え，各候補クラスタがホットスポットであるか否かの多重検定を行う．

$\mathcal{B}$の取り方として多くの提案がある．例えば，各地域間の距離 (各地域の中核都市間の距離) を$d_{ij}, i,j\in V$とおき，

$$\mathcal{B}=\bigcup_{0\le d\le d_0}\mathcal{B}(d) \quad (d_0:\text{閾値})$$
$$\mathcal{B}(d)=\big\{B\subset V:\text{極大}\mid d_{ij}\le d,\ \forall i,j\in B,\ B\neq\emptyset\big\}$$

とする提案がある．本発表の方法はこれ以外の場合でも適用可能である．

尤度比検定統計量は

$$\max_{B\in\mathcal{B}}\varphi_N(X_B,p_B),\qquad X_B=\sum_{i\in B}X_i,\quad p_B=\frac{\sum_{i\in B}\lambda_i}{\sum_{i\in V}\lambda_i}$$

ただし $N=\sum_{i\in V}X_i$,

$$\varphi_N(X_B,p_B)=\begin{cases}N\Big\{p_B\left(\frac{X_B/N}{p_B}\log\frac{X_B/N}{p_B}-\frac{X_B/N}{p_B}+1\right)+(1-p_B)\\ \quad\times\left(\frac{1-X_B/N}{1-p_B}\log\frac{1-X_B/N}{1-p_B}-\frac{1-X_B/N}{1-p_B}+1\right)\Big\} & \text{if } \frac{X_B/N}{p_B}\ge\frac{1-X_B/N}{1-p_B}\\ 0 & \text{otherwise}\end{cases}$$

である．最大値$\max_{B\in\mathcal{B}}\varphi_N(X_B,p_B)$ の帰無仮説 ($H_0:\theta_i\equiv$一定) の下での分布から多重性調整p値が定義される．それを求めるためには

$$P\big(X_B\le x_B,\ \forall B\in\mathcal{B}\,|\,N\big)=E\Big[\textstyle\prod_{B\in\mathcal{B}}\chi(B)\,|\,N\Big] \tag{1}$$

ただし$\chi(B)=\mathbf{1}\{X_B\le x_B\}$, の形の積分を行えばよい．本発表では多重性調整$p$値を求めるために，この数値積分を高速に行うためのアルゴリズムを提案する．

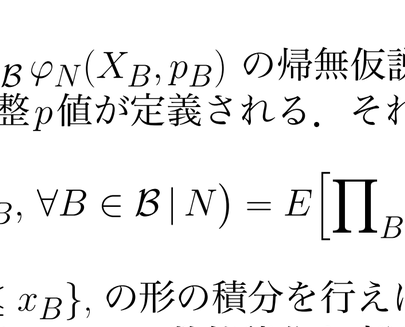
2　統計量の依存関係のグラフ表示

スキャン統計量の依存関係を反映させる形で，無向グラフ$G=(V,E)$, $V=\{1,\dots,n\}$ (頂点集合),

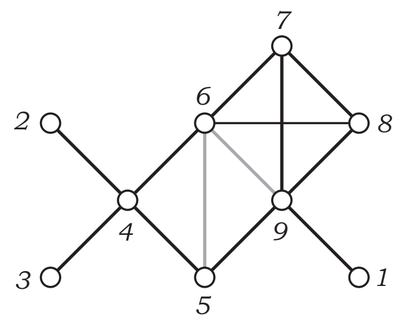
$$E=\{(i,j)\in V\times V\mid d_{ij}\le d_0\}\quad(\text{辺集合})$$

を定義する．以下，$n=9$の例で説明する．

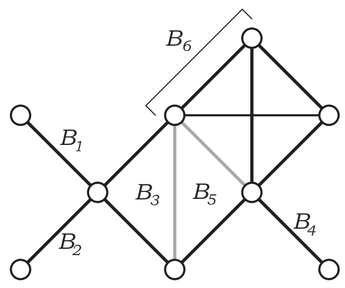
Step 1: 頂点ラベルを最小次数順序で付けかえ



Step 2: 三角化 (コーダル拡張)



Step 3, 4: 極大クリークの探索と順序付け



ここまでの手順で，極大クリークの完全列$B_m,\dots,B_1$が得られる．すなわち，各iについて，添字$k(i)>i$が存在し$H_i\cap B_i\subset B_{k(i)}$.

$$H_i:=B_{i+1}\cup\cdots\cup B_m$$

$$C_i:=H_i\cap B_i=B_{k(i)}\cap B_i,\quad R_i:=B_i-C_i$$

とおく．このとき$V=B_m\sqcup R_{m-1}\sqcup\cdots\sqcup R_1$ である．

3　逐次計算公式

極大クリークB_iに対して

$$\widetilde{\chi}(B_i)=\prod_{C\in\mathcal{B},C\subset B_i,C\not\subseteq B_{k(i)}}\chi(C)$$

とおく．また

$$M_i:=\sum_{j\in R_i}X_j,\quad T_i:=R_i\sqcup\bigsqcup_{j\in k^{-1}(i)}T_j,\quad N_i:=M_i+\sum_{j\in k^{-1}(i)}N_j$$

これらの記法の下で,

$$\begin{aligned}\text{式 (1)}&=E^{|N_6}\big[\widetilde{\chi}(B_6)\widetilde{\chi}(B_5)\widetilde{\chi}(B_4)\widetilde{\chi}(B_3)\widetilde{\chi}(B_2)\widetilde{\chi}(B_1)\big]\\&=E^{(M_6,N_5,N_4)|N_6}E^{B_6|M_6}\big[\widetilde{\chi}(B_6)\xi(C_5,N_5)\xi(C_4,N_4)\big]\end{aligned}$$

ただし

$$\begin{aligned}\xi(C_5,N_5)&=E^{(M_5,N_3)|N_5}E^{R_5|M_5}\big[\widetilde{\chi}(B_5)\xi(C_3,N_3)\big]\\\xi(C_4,N_4)&=E^{R_4|N_4}\big[\widetilde{\chi}(B_4)\big]\\\xi(C_3,N_3)&=E^{(M_3,N_1,N_2)|N_3}E^{R_3|M_3}\big[\widetilde{\chi}(B_3)\xi(C_1,N_1)\xi(C_2,N_2)\big]\\\xi(C_1,N_1)&=E^{R_1|N_1}\big[\widetilde{\chi}(B_1)\big],\quad\xi(C_2,N_2)=E^{R_2|N_2}\big[\widetilde{\chi}(B_2)\big]\end{aligned}$$

ここで$E^{R_5|M_5}$は多項分布$(X_i)_{i\in R_5}\ \sim\ \mathrm{Mult}\big(M_5;(\lambda_i/\sum_{j\in R_5}\lambda_j)_{i\in R_5}\big)$, $E^{(M_5,N_3)|N_5}$は多項分布$(M_5,N_3)\sim\mathrm{Mult}\big(N_5;(\sum_{i\in R_5}\lambda_i,\sum_{i\in T_3}\lambda_i)/\sum_{i\in T_5}\lambda_i\big)$による期待値である．

数値例：$(\lambda_i)=(3,3,3,3,3,3,3,3,3)$, $(X_i)=(2,2,2,7,2,7,2,2,2)$ のとき，スキャン統計量の最大値は$B=\{4,6\}$ときの5.167364で与えられる．p値は0.0151605，その計算時間は2分44秒 (ThinkPad T60p, Vine Linux 5.2, 言語R)．一方，素朴な数え上げによる計算時間は7時間0分45秒で，計算時間は1/154に改善される．

本発表は，原尚幸客員准教授（新潟大学），高橋邦彦氏（国立保健医療科学院）との共同研究に基づいている．