様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年６月８日現在

機関番号：６２６０３
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：２１５００２８８
研究課題名（和文）組合せ構造を持つ多次元分布の高速計算法の開発と空間疫学への応用
研究課題名（英文）Fast numerical computation for multivariate distributions with combinatorial structure and its application to spatial epidemiology
研究代表者
栗木　哲（KURIKI SATOSHI）
統計数理研究所・数理･推論研究系・教授
研究者番号：90195545

研究成果の概要（和文）：空間疫学においては，地理情報を反映したスキャン統計量の多重性調整 p 値の評価が必要である．しかしこの計算には高次元積分が必要なため，正確に数値計算を行うことは困難とされている．またモンテカルロシミュレーションでは精度良くおこなうことができない．本研究では，スキャン統計量の地理情報に基づく相関構造をグラフで表し，そのグラフからマルコフ性を抽出することによって，数値計算を逐次的な計算に帰着させる方法を開発した．提案法は計算量を大きく低減することが確認できたが，一方で実データに対しては実用的な計算時間で計算できるものではなく，今後実用化に向けた研究が必要であることが分った．

研究成果の概要（英文）：In spatial epidemiology, it is important to evaluate the multiplicity-adjusted p-value of scan statistics by taking their spatial correlations into account. However, it is difficult to conduct exact numerical calculation because high-dimensional integrations are required. On the other hand, Monte Carlo simulations cannot yield precise results. In our proposed method, we represent the spatial correlations of scan statistics as a graph, and by extracting its Markov structure, rewrite the high-dimensional integration as a successive numerical integration. Although, the proposed method reduces computational time of integration drastically, it is not powerful enough for the real datasets. More research is needed for practical use.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,200,000	360,000	1,560,000
２０１０年度	1,000,000	300,000	1,300,000
２０１１年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,200,000	960,000	4,160,000

研究分野：総合領域
科研費の分科・細目：統計科学
キーワード：空間疫学・スキャン統計・コーダルグラフ・多重検定

１．研究開始当初の背景

ある広領域に含まれる n 個の地域 $i=1,\ldots,n$ について，事故数や疾病数などのデータ X_i が，付随する共変量より予測される期待度数 λ_i

とともに得られているとする．事故や疾病が集積する地域では，X_i が λ_i よりも有意に大きい値をとると考えられる．そのような地域はホットスポットとよばれ，そのような地域を統計量（スキャン統計量）を用いて統計的に検出することがここでの問題である．これらのホットスポットは，単一地域にとどまらないでクラスターをなす場合が多い．そのため，隣接する地域を併合した値もスキャンの対象とし，対応するスキャン統計量を求め，多重性調整 p 値によって有意性を判定することが行われている．

２．研究の目的

ところでこの多重性調整 p 値は数式としては多重積分（和文）で表されるものであり，その厳密な数値計算を定義式通りに行うのは計算量の観点から現実的ではない．現状の空間疫学のデータ解析ではモンテカルロシミュレーションを行っているが，ホットスポットの判定が必要となるのは分布の裾領域に対応するため，素朴な乱数シミュレーションではその乱数の個数を相当増やしても，なかなか精度を出すことができないという問題がある．本研究の目的は，マルコフ性に基づく逐次数値積分を用いてこのような多次元数値計算を高速に行うための方法論を開発することである．

３．研究の方法

本研究では，このような多次元数値計算を高速に行うために，マルコフ性に基づく逐次数値積分の利用を試みた．地域の隣接関係や地域間距離により無向グラフを定義することができる．そのグラフを出発として，統計量間のマルコフ性を抽出し，それをもとにする漸化式によって，数値積分の計算量を低減するというものが基本的な考え方である．具体的にはグラフをコーダル拡大し，その極大クリークの完全列（ジャンクション木）を求めることによって，統計量間のマルコフ性を導く方法を用いる．

４．研究成果

上記に述べた方針で，地理情報から無向グラフを生成し，そのグラフのコーダル拡大を行い，その極大クリークの完全列を抽出し，p 値を求めるための漸化式（アルゴリズム）を開発した．また簡単なプロトタイププログラムを作成した．結果として，逐次数値積分が原理的に可能であることを見いだした．特にプログラム言語 R を用いてプロトタイププログラムを作成し，提案する逐次数値積分が正しいものであることを確認した．さらにいくつかのテストデータについて計算時間の評価を行った．提案法は計算量を大きく低減することが確認できたが，一方で実データに対しては実用的な計算時間で計算できるものではなく，今後理論・アルゴリズム，インプリメンテーションの両方の観点から実用化に向けたさらなる研究が必要であることが分った．

以下にここで用いた (1) スキャン統計量，(2) 極大クリークの完全列の導出，(3) 逐次計算方式，(4) 数値例について例題を用いて簡単に説明する．

(1) スキャン統計量

n 個の地域 $V=\{1,\ldots,n\}$ の地域 i におけるイベントの発生数 X_i の期待度数を λ_i とする．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

況を想定する．その地域クラスターがホットスポットである．

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）$\beta \subset 2^V$ 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．

β の取り方として多くの提案がある．ここでは，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，

$$\beta = \bigcup_{0 \le d \le d_0} \beta(d) \quad (d_0 : \text{閾値})$$

$$\mathcal{B}(d) = \{B \subset V : \text{極大} \mid d_{ij} \le d, \forall i, j \in B, B \neq \varnothing\}$$

とする．尤度比検定統計量は

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

ただし $N = \sum_{i \in V} X_i$，

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

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

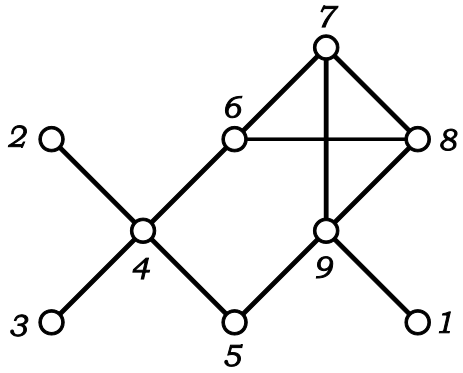
$$P\left(X_B \le x_B, \forall B \in \mathcal{B} | N\right) = E\left[\prod_{B \in \mathcal{B}} \chi(B) | N\right] \quad (1)$$

ただし $\chi(B) = \mathbf{1}\{X_B \le x_B\}$,の形の積分を行えばよい．

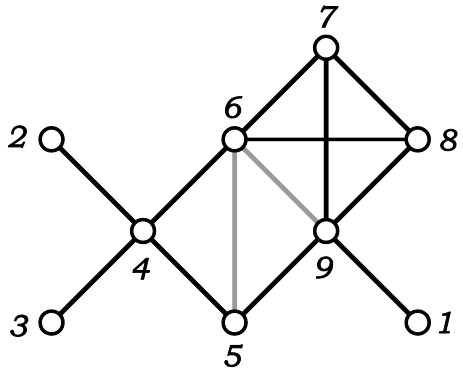
(2) 統計量の依存関係のグラフ表示

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 $G = (V, E)$，$V = \{1, \ldots, n\}$（頂点集合），$E = \{(i, j) \in V \times V \mid d_{ij} \le d_0\}$（辺集合）を定義する．以下，$n = 9$ の例で説明する．

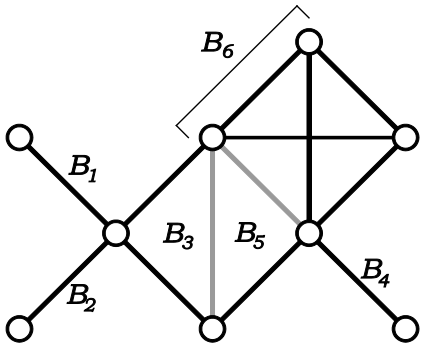
Step 1: 無向グラフを定義する．



Step 2: グラフの三角化（コーダル拡張）を行う．



Step 3: 極大クリークを探索し，その順序付けを行う．



ここまでの手順で，極大クリークの完全列 $B_m, \ldots, B_1$ 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

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

とおく．また

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

$$N_i := M_i + \sum_{j \in k^{-1}(i)} N_j$$

これらの記法の下で，

式(1) $= E^{|N_6}\left[\tilde{\chi}(B_6)\tilde{\chi}(B_5)\tilde{\chi}(B_4)\tilde{\chi}(B_3)\tilde{\chi}(B_2)\tilde{\chi}(B_1)\right]$

$= E^{(M_6,N_5,N_4)|N_6} E^{B_6|M_6}\left[\tilde{\chi}(B_6)\xi(C_5,N_5)\xi(C_4,N_4)\right]$ (2)

ただし

$\xi(C_5,N_5) = E^{(M_5,N_3)|N_5} E^{R_5|M_5}\left[\tilde{\chi}(B_5)\xi(C_3,N_3)\right]$

$\xi(C_4,N_4) = E^{R_4|N_4}\left[\tilde{\chi}(B_4)\right]$

$\xi(C_3,N_3)$

$= E^{(M_3,N_1,N_2)|N_3} E^{R_3|M_3}\left[\tilde{\chi}(B_3)\xi(C_1,N_1)\xi(C_2,N_2)\right]$

$\xi(C_1,N_1) = E^{R_1|N_1}\left[\tilde{\chi}(B_1)\right], \xi(C_2,N_2) = E^{R_2|N_2}\left[\tilde{\chi}(B_2)\right]$

ここで $E^{R_5|M_5}$ は多項分布

$$(X_i)_{i \in R_5} \sim \mathrm{Mult}\left(M_5; (\lambda_i / \textstyle\sum_{j \in R_5} \lambda_j)_{i \in R_5}\right),$$

$E^{(M_5,N_3)|N_5}$ は多項分布

$$(M_5, N_3) \sim \mathrm{Mult}\left(N_5; (\textstyle\sum_{i \in R_5} \lambda_i, \sum_{i \in T_3} \lambda_i) / \sum_{i \in T_5} \lambda_i\right)$$

による期待値である．式 (2) は逐次数値計算が可能な式になっている．

(4) 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）であった．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 に改善された．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

① Satoshi Kuriki, Akimichi Takemura, Volume of tubes and the distribution of the maximum of a Gaussian random field, Selected Papers on Probability and Statistics, American Mathematical Society Translations Series 2, 査読無, Vol. 227, 2009, 25-28

② Satoshi Kuriki and Yasuhide Numata, Graph presentations for moments of noncentral Wishart distributions and their applications, Annals of the Institute of Statistical Mathematics, 査読有, Vol. 62, 2010, 645-672

③ 加藤直広、栗木哲、2 次多項式回帰曲線の正値性検定、応用統計学、査読有、41 巻、2012、1－15

〔学会発表〕（計 1 件）

Satoshi Kuriki, Tube method --- An integral-geometric approach to statistical distribution theory, GEOMETRY, TOPOLOGY, ALGEBRA and NUMBER THEORY, APPLICATIONS" dedicated to the 120th anniversary of Boris Delone, 2010.08.16, Steklov Mathematical Institute, Moscow, Russia

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕

ホームページ等

６．研究組織

(1)研究代表者

栗木　哲（KURIKI SATOSHI）

統計数理研究所・数理･推論研究系・教授

研究者番号：90195545

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者
高橋邦彦（TAKAHASHI KUNIHIKO）
国立保健医療 科学院・技術評価部
・研究員
研究者番号：50323259