様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　　６　日現在

機関番号：　１５３０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３００４２

研究課題名（和文）空間データの潜在構造を表現する統計モデルの効率的な推測・選択

研究課題名（英文）Effective inference and selection of statistical models to represent latent structure in spatial data

研究代表者

坂本　亘（Sakamoto, Wataru）

岡山大学・環境生命科学研究科・教授

研究者番号：７０３０４０２９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：地図・空間上のデータに潜む複雑な構造を明らかにするために，高次元の潜在変数を伴う統計モデルを考え，効率的な推測・選択を行う方法を研究した．疾病地図データへの適用で提案した，推定空間効果を用いた領域同定の方法は，従来の方法に比べて高リスク領域を適切に同定する可能性が高いことが示された．また，年齢・時代・コホート（APC）モデルによるがん死亡率データの解析で検討されたモデル選択の方法は，各効果を適切に推定し，解釈上の新たな知見を与えることが示唆された．

研究成果の概要（英文）：Some methods of making effective inference and selection in statistical models with high-dimensional latent variables are considered to reveal complicated latent structure in spacial and geographical data. It was shown that the method of detecting regions using estimated spatial effect, proposed for application to disease mapping data, had higher possibility of detecting regions with high risk than existing methods. Also it was suggested that the model selection method considered in the analysis of cancer mortality data with age-period-cohort (APC) models could estimate each effect appropriately, and give a new knowledge for interpretation.

研究分野：　統計科学（計算統計学，数理統計学，医学統計学）

キーワード：　階層 Bayes モデル　Markov 確率場　モデル選択　情報量規準　領域同定　疾病地図データ　APC モデル

１．研究開始当初の背景

環境・生命科学の諸問題を科学的根拠に基づいて解決するための手段として，地図上・空間上のデータに潜在する複雑な構造を明らかにすることが要請される．観測されたデータに基づいて，その構造を表現する統計モデルの上で推測を行ったり，最適なモデルを選択したりする方法が必要となる．

このような複雑な構造を表現するモデルとして，高次元の潜在変数を伴う統計モデルが研究されている．潜在 Gauss 型モデルがその一つである．空間上の相関構造を表現する潜在変数の分布構造に Gauss 型 Markov 確率場（GMRF）を取り入れ，Bayes 流接近法を用いて推測を行う．潜在変数やパラメータの事前分布を階層構造により導入する．そして，観測を得たときの事後分布，とくに，関心のない変数に関して周辺化した，関心のある変数の周辺事後分布を得る．

ここで問題となるのは周辺化のための高次元の積分計算である．しかしながら，Markov 連鎖モンテカルロ（MCMC）法をはじめとする既存の方法は，膨大な計算時間を要する上に，サンプリングの調整方法やその妥当性の点で課題がある．他方，周辺化の近似計算は，積分変数の次元が大きい場合に，精度が悪くなることが問題であった．最近提案され注目を集めている積分入れ子型 Laplace 近似（INLA）法（Rue *et al.*, 2009）は，潜在変数の正規性と疎な構造を利用することで，高速計算を実現し，なおかつ良好な近似精度を与えることが示されている．

最適な潜在 Gauss 型モデルの選択の方法として，所与の構造に対して偏分情報量規準（DIC）などの評点を計算し，最適化を行うことが提案されている．しかしながら，潜在変数が高次元の場合，考慮すべき変数の組み合わせが膨大になるため，より効率的な選択方法の研究が課題となっていた．

２．研究の目的

本研究では，潜在 Gauss 型モデルなどの高次元の潜在変数を伴う統計モデルにおいて，変数間の関係を効率良く推測・選択する方法を検討し，観測された現象や結果に対する原因を明確に提示するための方法などを提案することが目的であった．観測された現象・結果をもとに，複雑に絡み合う原因を解明し，環境・生命科学の諸問題の解決に寄与することを目指した．具体的には以下の三つの目的を掲げた．

(1) 環境科学・疫学のデータに潜在 Gauss 型モデルなどの高次元の潜在変数を伴うモデルを適合させる．その際に，変数間の関係を効率良く推測・選択することにより，観測された現象や結果に対する原因を明確に提示するための方法を提案する．モデルの複雑さ，具体的には潜在変数間の従属関係を制御するパラメータを，得られた観測からどのように推定するかを検討する．さらに，変数および変数間の関係の重要度を算出・提示し，容易に解釈できるような方法を提案する．

(2) 正常な状態と異常な状態を統計的に識別するための方法を提案する．潜在変数の構造が，低次元のパラメータをもつ単純なモデルで表される状態を正常とし，これに複雑な成分（ランダム効果あるいは非線形項）が追加されたモデルで表される状態を異常とする．このとき，複雑さを制御するパラメータを適切に推定することにより，両者の識別を可能にする．

(3) 提案する方法を，地理・空間情報を含む環境科学データや疾病地図データなどの解析に応用する．例えば，汚染や疾病などの特異な状況が起きている地点（いわゆるホットスポット）を早期にかつ客観的に特定するといったことが可能かどうかを検討する．

３．研究の方法

(1) 高次元の潜在変数を伴うモデルの推測方法の性能評価

Rue *et al*.（2009）などの文献，および R-INLA パッケージ（http://www.r-inla.org/）を通じて，INLA 法の理論，周辺事後分布の近似・計算の方法，諸種のモデルへの拡張の方法を調査した．さらに，一般化線形混合効果モデルの場合に，従来の近似方法（PQL：罰則付き擬似尤度法）や MCMC 法などとの近似精度の比較を行うためのシミュレーションを実施した．

(2) 高次元の潜在変数を伴うモデルの選択方法の性能評価

潜在変数を伴うモデルの選択に必要となる種々のモデル選択の方法について，ソフトウェア R のパッケージに実装されているプログラムを用いて，疾病地図データへの適用の中で検討を行った．

(3) 高リスク領域同定方法の提案

正常な状態と異常な状態を統計的に識別するための方法の提案につなげるものとして，高リスク領域（ホットスポット）を同定する方法を検討した．具体的には，疾病地図データに対して，隣接領域間の相関をもつ GMRF を空間効果に対する事前分布として導入し，INLA 法で得られた隣接空間効果の事後平均に対してエシェロン・スキャンという効率的な領域探索法を用いることを提案した．実データへの適用とシミュレーションを通じて，従来の方法（相対リスクの経験 Bayes 推定値に基づくエシェロン・スキャン）との比較を行った．

(4) ランダム・グラフ利用の検討

潜在変数間の従属関係を表すモデルに対する事前分布として，ネットワーク理論で用いられるランダム・グラフの導入を検討した．ランダム・グラフでは，変数間の関連，すなわちグラフの頂点同士が辺で結ばれているかどうかが確率的に選択される．そこで，変数間の関連の強さを制御するためのパラメータを導入し，平滑化法などで研究されている複雑度パラメータの推定方法を応用することを検討した．

(5) 年齢・時代・コホート（APC）モデルへの応用

APC モデルの中で各効果の事前分布に GMRF をおくことにより同定可能性の問題を回避した階層 Bayes モデルを構築し，INLA 法による各効果の事後分布の推定を行う方法を健闘した．GMRF の次数選択に偏分情報量規準 (DIC) を用いることを提案し，日本のがん死亡率データに適用した．

４．研究成果

(1) 高次元の潜在変数を伴うモデルの性能評価

一般化線形混合効果モデルに対してシミュレーションにより性能評価を行った．その結果，INLA 法によって推定されたパラメータ（固定効果の係数，ランダム効果の分散）は，最小二乗誤差や偏りの意味で，従来の近似方法よりも小さくなり，MCMC 法と同等の精度をもつことがわかった．とくに対象数および繰り返し観測数が小さい場合にその傾向が顕著であった．

(2) 高次元の潜在変数を伴うモデルの選択方法の性能評価

ドイツの地区別喉頭がん発生率のデータ (Rue and Held, 2005) を用いて検討した．潜在変数を含むモデルの選択では，AIC などの従来のモデル選択基準ではうまく行かない場合がある．他方，偏分情報量規準 (DIC) などの Bayes 流選択基準は，従来の規準よりも妥当なモデルを選択し，その結果，推定された潜在構造に基づいて，地図上の疾病の集積性や，喫煙のような共変量効果を提示するのに有用であることが分かった．今後，様々な実データやシミュレーションを通じて，さらなる調査が必要である．

(3) 高リスク領域同定方法の提案

潜在 GMRF モデルのもとでの隣接空間効果の事後平均に対するエシェロン・スキャン法を(2)と同じ疾病地図データに適用した．その結果，従来の方法（相対リスクの経験 Bayes 推定値に基づくエシェロン・スキャン）に比べて，より凝集した解釈しやすい領域が同定され，最適なモデル選択基準の値を与えることが示された．喫煙率を共変量に加えた場合に，異なるクラスター領域が同定されることも分かった．さらに，シミュレーション実験を行い，従来の方法に比べて，高リスク領域とそうでない領域を適切に区別する可能性が高いことが示された．

今後は，偏分情報量規準 (DIC) などのモデル選択基準などを用いて同定された高リスク領域の評価方法について検討を重ねたい．

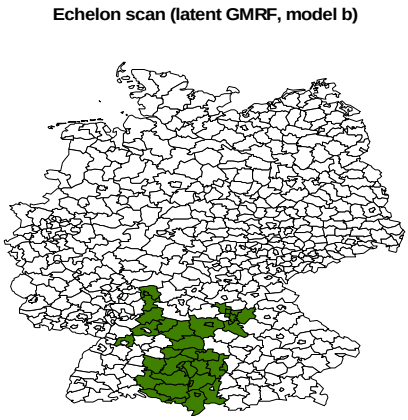

図１：提案方法によって得られた１次クラスター領域（喫煙の線形効果を考慮した場合）．

(4) ランダム・グラフ利用の検討

潜在変数間の従属関係に対する事前分布として，ランダム・グラフの導入を検討した．しかしながら，高リスク領域の同定のために，個々の候補領域に対してその都度モデルのあてはめ（事後平均の算出）の計算を行う必要があり，計算時間が予想外にかかり，効率的でないことが分かった．ランダム・グラフの導入については今後の検討課題としたい．

(5) APC モデルへの応用

日本の肝がん死亡率データを解析した結果，DIC によって次数選択された GMRF を用いて推定された各効果のうち，コホート効果について急激な変化が暗示された．適用したモデル選択方法の理論的妥当性については，シミュレーションなどによるさらなる性能評価が必要であると考える．

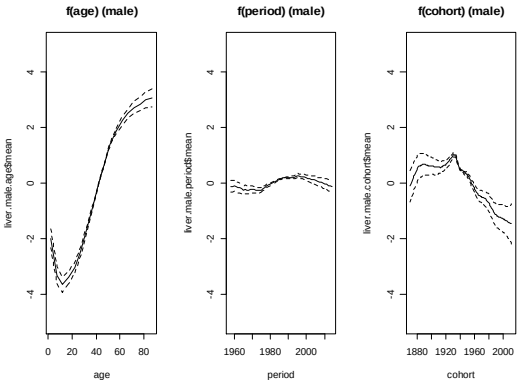

図２：選択されたモデルによって推定された年齢・時代・コホートの各効果．

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）

〔学会発表〕（計４件）

[1] W. Sakamoto : Cluster detection of disease mapping data based on latent Gaussian Markov random field models. IASC-ARS Conference, 2016/11/4-5, Daejeon (Korea)

[2] W. Sakamoto : An analysis of Japanese liver cancer mortality data with Bayesian age-period-cohort models. International Conference for JSCS 30th Anniversary, 2016/10/16-17, Seattle (USA)

[3] W. Sakamoto : Cluster detection of disease mapping data based on latent Gaussian Markov random field models. COMPSTAT 2016, 2016/8/23-26, Oviedo (Spain)

[4] S. Hagihara and W. Sakamoto : Performance of Bayesian inference with integrated nested Laplace approximation in generalized linear mixed effect models. The 24th South Taiwan Statistical Conference, 2015/6/28, Changhua (Taiwan)

６．研究組織

(1) 研究代表者

坂本　亘（SAKAOMOTO, Wataru）

岡山大学・環境生命科学研究科・教授

研究者番号：70304029