論文の内容の要旨

論文題目　Variable selection problem in mixed effects models with application to small area estimation
（混合効果モデルにおける変数選択問題と小地域推定への応用）

氏　　名　川久保　友超

本論文は，混合効果モデル（mixed effects models）における変数選択問題に取り組んだものである．同問題に対するアプローチは様々なものがあるが，特に情報量規準による手法に焦点をあてている．Vaida and Blanchard (2005) は条件付赤池情報量（conditional Akaike information, cAI）という概念を導入したが，これは変量効果（random effect）を所与とした条件付尤度にもとづいた期待カルバック・ライブラー・ダイバージェンスに関連した，モデルのリスクとも言うべきものである．cAI とそこから導かれる条件付赤池情報量規準（conditional AIC, cAIC）は，クラスターごとの推測に関心がある場合は，変量効果を積分消去した周辺尤度にもとづいた従来の AIC よりも適切である．このことは，cAIC を用いた変数選択が，小地域推定（small area estimation）のような変量効果の予測を含んだ問題に有用であることを示唆している．そこで本論文では，cAIC を中心とした混合効果モデルにおける変数選択規準に取り組むとともに，小地域推定への応用問題を考察した．

まず第 1 章において introduction として，変量効果モデルの変数選択に関する先行研究を cAIC を中心にレビューした．第 2 章においては，変量効果モデルの一般論とその最も基本的なモデルのクラスとして線形混合モデル（linear mixed model）を説明している．また，変数選択の問題設定と，Vaida and Blanchard (2005) の提案した conditional AIC の説明も行っている．そして，第 3 章から続く 5 つの章で，研究成果がまとめられている．

第 3 章では，線形混合モデルにおける Vaida and Blanchard (2005) の cAIC の問題点を指摘した上で，オリジナルの cAIC を修正した規準として modified conditional AIC (McAIC) を提案した．Vaida and Blanchard (2005) の問題点は，候補モデルが overspecified，すなわち真のモデルを含んでいるという仮定のもとで，cAI の不偏推定量として導出されている点である．その結果，underspecified モデル，すなわち真のモデルを含んでいないモデルにおいて，cAIC は cAI の推定量として大きなバイアスを生じてしまっている．そこで，overspecified モデルと underspecified モデルの双方で cAI の漸近不偏推定量となるような規準として，McAIC を導出した．また，赤池型の情報量規準にもとづいたモデル平均化推定量に McAIC を用いると，cAIC にもとづいたモデル平均化推定量よりも予測精度が良くなることを数値実験で示した．

第 4 章では，推定に用いるモデルの共変量の値と予測に用いるモデルの共変量の値が異なる状況（共変量シフト）における，線形混合モデルの変数選択問題を考えた．共変量シフト下での cAI を定義した上で，まずは他の赤池型の情報量規準と同様，候補モデルが overspecified であるという仮定のもとで，cAI の不偏推定量として covariate shift cAIC (CScAIC) を導出した．第 3 章で overspecified の仮定の欠点を指摘したが，共変量シフトの状況下ではいっそう問題となる．というのも，共変量シフトを考慮しない通常の情報量規準では，overspecified モデル，underspecified モデルの双方で，規準の尤度部分がリスク関数（cAIC では cAI）のナイーブな推定量となっており，underspecified モデルでバイアス補正が妥当でなくとも，cAIC が cAI の推定量としてそれなりに機能するためである．一方，CScAIC の尤度部分が cAI のあまり良い推定量とは言えず，overspecified の仮定が崩れると，非常に大きなバイアスを生んでしまう．そこで，overspecified, underspecified 双方で cAI の漸近不偏推定量となっているものを導出し，これを変数選択規準として提案した．さらに，共変量シフトは小地域推定における有限母集団の平均の推定問題で重要となることに言及し，数値実験で提案手法の有用性を示した．

第 5 章では，自然指数分布族（natural exponential family, NEF）にもとづいた混合効果モデルにおける cAIC 導出した．NEF にもとづいた混合効果モデルのクラスは，カウントデータや二値データのモデリングに有用なポアソン・ガンマモデル，二項ベータモデルといった非線形混合モデルを含んでいる．これ

らのモデルは解析的に周辺尤度を導出できるため，変数選択規準として周辺尤度にもとづいた mAIC が当然利用可能である．しかしながら，クラスターごとの推測に関心がある場合は，mAIC は適切でない．そこで，非線形混合モデルにおける cAI を定義し，その漸近不偏推定量として cAIC を導出した．漸近不偏推定量の構成法として，3 つの手法を考えた．1 つは，完全に解析的な方法で，これは超母数が Godambe and Thompson (1989) から示唆される推定方程式で推定されたもとで推定量の確率展開を行うことでなされる．さらにこの手法は非線形なリンク関数を多項式近似するため，クラスター内のサンプルサイズがクラスター数 m に対して一定のスピードで発散するという仮定（仮定 A）も必要となる．仮定はやや強いが，計算負荷のあまりかからない手法といえる．2 つ目の手法は，数値微分，数値積分を用いた手法であり，仮定 A は必要でない．これは超母数の推定量の確率展開は行い，漸近バイアス，漸近分散は解析的に求めるが，バイアス補正にモンテカルロ近似にもとづいた数値積分および数値微分を用いた手法である．3 つ目の手法は，パラメトリック・ブートストラップを用いた手法である．仮定 A も超母数の推定法の制約もないが，ブートストラップ法でモンテカルロ積分やバイアス補正を行い，ブートストラップ繰り返しのそれぞれのステップで超母数の推定が必要なため，計算負荷が他の 2 つの手法に比べ大きい．

第 6 章では，ベイズ周辺尤度にもとづいた予測密度のリスクを，頻度論の立場から測った情報量規準を提案した．具体的には，線形回帰モデルの変数選択規準を，回帰係数に事前分布を入れることで導出した．提案手法には 3 つの利点がある．1 つは，この手法は頻度論とベイジアンの折衷であり，ベイズモデルのリスクを頻度論の立場から測っているため，事前分布の特定化の誤りの影響を受けにくい．2 つ目は，規準の構築に無情報事前分布を用いることができる点である．回帰係数に一様事前分布を仮定すると，得られる規準は Shi and Tsai (2002) の residual information criterion (RIC) と一致する．3 つ目は，提案手法は変数選択の一致性を持つ点である．誤差項ベクトルの共分散の構造を様々なケースで数値実験を行い，提案手法の有用性を示した．特に，興味の対象がクラスターでなく母集団全体である場合には，線形混合モデルの変数選択規準としても機能する．

第 7 章では，線形混合モデルの変数選択問題において，変量効果を所与とした条件付尤度の期待カルバック・ライブラー・ダイバージェンスにもとづいた予測密度のリスクを，様々な予測密度で比較し，導出される情報量規準の性質を議論した．予測密度としてプラグイン予測密度を考えると，そこから得られる情報量規準は Vaida and Blanchard (2005) の cAIC である．しかし，予測密度をプラグイン予測密度に制約する必然性はなく，ある観点からプラグイン予測密度より優れた予測密度は存在する．そこで，以下の 2 つの予測密度を考えた．1 つ目は，ベイズ予測密度（Bayesian predictive density）であり，これは期待カルバック・ライブラー・ダイバージェンスを最小にするという意味で，最も優れた予測密度であることが知られている．ここから導かれる規準は，Akaike (1980) の predictive likelihood および Kitagawa (1997) の predictive information criterion (PIC) と呼ばれるものである．ここでは，線形混合モデルにおける PIC として，回帰係数を未知パラメータと考えたもの（PIC1）と，回帰係数に事前分布を仮定したものの（PIC2），2 通りを考え，比較した．2 つ目は，ベイズ周辺尤度にもとづいた予測密度であり，これは第 6 章で導出した規準の，変量効果の予測を目的とした場合の変形と見ることができる．数値実験で，ベイズ周辺尤度にもとづいた規準は変数選択の一致性があることが確認できた一方，PIC2 はサンプルサイズが小さくノイズの大きいデータに対して有用であることが分かった．

参考文献

Akaike, H. (1980). On the use of predictive likelihood of a Gaussian model. *Annals of the Institute of Statistical Mathematics*, **32**, 311–324.

Godambe, V.P. and Thompson, M.E. (1989). An extension of quasi-likelihood estimation (with Discussion). *Journal of Statistical Planning and Inference*, **22**, 137–152.

Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Communications in Statistics — Theory and Methods*, **26**, 2223–2246.

Shi, P. and Tsai, C.-L. (2002). Regression model selection—a residual likelihood approach. *Journal of the Royal Statistical Society series B*, **64**, 237–252.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351–370.