様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ５ 月 ３１ 日現在

機関番号：１５４０１
研究種目：基盤研究(C)（一般）
研究期間：2012～2015
課題番号：２４５００３４３

研究課題名（和文）経時データ解析におけるモデル選択法の開発

研究課題名（英文）Development of model selection criteria for longitudinal data

研究代表者
若木　宏文（Wakaki, Hirofumi）

広島大学・理学（系）研究科（研究院）・教授

研究者番号：９０２１０８５６

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：ランダム効果モデルについて、カルバックライブラー擬距離に基づく予測分布のリスクの、最大対数尤度によるナイーブな推定量のバイアスの漸近展開公式をラプラス近似手法を用いて導出した。標本数を大きくするときの誤差項のオーダーは未知母数に関して一様であることを示し、バイアスを標本数の逆数のオーダーまで修正した変数選択規準を導出した。
ランダム係数が２個の場合の線形混合モデルの最尤推定量を導出し、予測分布のリスクを独立な３つのベータ変量のある関数の期待値として記述し、ラプラス近似の見通しを得たが、誤差項のオーダーの未知母数に関する一様性に関する問題は未解決である。

研究成果の概要（英文）：We derived an asymptotic expansion formula of the bias of the naive estimator by the maximum log-likelihood function for the risk of the predicted distribution based on Kullback-Leibler divergence for a random coefficient model with using the Laplace's method. We prove that the order of the error term of this approximation formula is uniform with respect to the unknown parameters. We proposed a bias-modified AIC criterion of which the order of bias is o(1/n) where n is the sample size.
We also treated a mixed effects model with two random coefficients. the maximum likelihood estimators of the unknown parameters are derived. We represent the bias of the predicted distribution so that we can apply the Laplace's method. However, it is not clear whether the order of the error term is uniform with respect to the unknown parameters.

研究分野：数理統計学

キーワード：経時データ　ランダム係数　ラプラス近似　変数選択　AIC

１．研究開始当初の背景

経時測定データとは、興味の対象である目的変数 y_{ij} と、目的変数に影響を与えると考えられる共変量ベクトル x_{ij} からなる。ここで、i は個体番号を表わし、j は測定時点を表わす。このようなデータの基本的な解析手法は回帰分析であり、回帰モデル

$$Y_{ij} = \alpha_i + x_{ij}'\beta + \varepsilon_{ij} \quad (i=1,\cdots,n;\ j=1,\cdots,p)$$

を仮定して、データから未知母数α、βを推定することにより、共変量の与える影響を分析する。ε_{ij} は誤差を表わす平均 0 の確率変数であり、典型的な分析では正規分布が仮定される。α_i は個体 i 固有のパラメータであるが、分析の本来の目的は個体が抽出された大きな母集団における目的変量と共変量との関連性を明らかにすることである。α_i を個体に応じて変動する確率変数として扱うモデル

$$Y_{ij} = A_i + x_{ij}'\beta + \varepsilon_{ij}$$

をランダム効果モデルと呼ぶ。ここで、A_i は平均α、分散 τ^2 を持つ確率変数であり。この項によって、$Y_{i1}, \cdots, Y_{im}$ 間の相関が表現される。回帰係数ベクトルの成分の一部も個体に応じて変動する確率変数と扱うモデルを線形混合モデルと呼ぶ。また、回帰モデルは、目的変量の期待値を説明変数の一次式で表現するモデルであるが、目的変量の期待値の関数を目的変数の一次式で表現するモデルを、一般化線形モデルと呼び、一般化線形モデルの切片項や回帰係数の一部を確率変数として扱うモデルを一般化線形混合モデルと呼ぶ。

変数選択問題とは、共変量の候補として、$X_1,\cdots, X_k$ があるとき、この中からどれをモデルに組み込むかという問題である。AIC(赤池情報量規準)は、変数選択問題において、広く用いられている方法であり、データから得られる未知母数の推定量を用いた予測分布のカルバックライブラー擬距離に基づくリスクの推定量である。共変量の候補から変数を取捨選択して得られる各モデルの AIC の値を計算し、AIC が最小となるモデルを選択することで、共変量の組み合わせを決定する方法であるが、ランダム効果モデルや線形混合モデルでは未知母数の最尤推定値が母数空間の境界上に現れる場合があり、このとき、AIC は、リスクの推定量としては偏った推定量となってしまう。

申請者は、指導大学院生とともに、パラレルプロファイルモデルにおいて、ランダム効果を導入することで得られる、ある共分散構造の仮説検定問題において、検定統計量の分布の近似公式の導出に取り組んだが、そこで用いた近似手法が、線形混合モデルのリスク評価に利用できると考え、本研究を立案した。

２．研究の目的

本研究では、(1)ランダム効果モデル、線形混合モデル、一般化線形混合モデルにおけるモデル選択規準の改良および、新たな選択規準の構築と、（２）一般化推定方程式におけるモデル選択規準の構築を目指した。

３．研究の方法

各モデルに関して、（１）最尤推定量の導出、（２）最大対数尤度によるリスク推定量のバイアス評価、（３）バイアス補正の順に研究を進めることとした。

（１）　ランダム係数の導入によって、母数間に不等式制約が課せられるが、不等式制約がない場合の最尤推定値が、不等式制約を満たしていない場合、母数空間の境界上で、尤度の最大化を行った。

（２）　ランダム効果モデルや線形混合モデルでは、バイアスがベータ分布に関する、ある関数の制限した領域での期待値と表現される。ベータ分布の自由度が標本数とともに増加する状況であるので、ラプラス近似の手法を利用して、バイアスの近似公式を導出した。

（３）　（２）で得られた近似公式と、不等式制約の閾値をずらして得られる、定義関数の期待値の展開式と比較しながら、バイアス補正を試みた。

４．研究成果

（１）　ランダム効果モデルの変数選択規準の導出

①　漸近バイアス

多標本の成長曲線モデルで、測定時点に関する多項式回帰を想定したモデルで、個体に関する観測ベクトルの相関構造が、ランダム効果の分散のみで決まるモデルを扱った。回帰係数ベクトルの最尤推定量は、通常の最小２乗推定量となる。目的変量ベクトルの分散共分散行列は一様相関構造をもち、二つのパラメータ σ^2 と λ^2 によって表される。ランダム効果由来のパラメータであることから、$\lambda^2 \geqq \sigma^2$ という制約条件が課せられる。説明変数行列をうまく直交化すると、残差平方和は、偏回帰係数ベクトルに関する偏差平方和 S_1 と、線形回帰式の切片項に関する残差平方和 S_2 に分解することができる。p を観測時点数とすると、λ^2、σ^2 の最尤推定値は、$S_1 > (p-1)S_2$ のとき、母数空間の境界 $(\lambda^2=\sigma^2)$ 上に値を取る。最大対数尤度のリスクの推定量としてのバイアスは、ベータ分布に従う確率変数 B と $\rho=\lambda^2/\sigma^2$ の関数 $G(B,\rho)$ の、$B>c$ なる領域での期待値を用いて表わされる。c は未知母数ρの関数である。ベータ分布の確率密度関数は n を標本数として、

exp(n F(b))g(b)の形で表わされるので、ラプラスの近似手法を用いて

$$\int_c^1 \exp(n\,F(b))\,g(b)\,G(b,\rho)\,db$$

の n→∞のときの漸近展開を導出すればバイアスを評価できる。ラプラス近似は、通常、F(b)の最大化点 b_0 が、積分区間の内点であることを仮定しているが、本研究においては、ρ= 1、すなわち、ランダム効果の分散が 0 である場合には、c = b_0 となり、ρ> 1 の場合には、c < b_0 となる。ラプラスの近似手法を検討した結果、最大化点 b_0 が積分区間の端点である場合には、積分の極限は、最大化点が内点である場合の丁度半分であることがわかる。したがってバイアスの極限は、ρ>1 の場合、ρの値によらず未知母数の個数の-２倍、すなわち、従来の AIC と同じ値となり、ρ=1 の場合には、-２×(未知母数の個数- 1)となることがわかった。

②剰余項のオーダーと漸近展開近似

①で導出した、漸近バイアスはρ= 0 で不連続であるが、リスクそのものはρの連続関数であるから、近似公式もρの連続関数であるべきである。ラプラス近似の誤差項を詳しく調べた結果、近似誤差の 0 への収束の速さには、ρの値が関係しており、ρが 0 に近い程、収束が遅くなっていることがわかった。ラプラス近似の誤差は、前述の積分において、被積分関数を b=b_0 の周りでテーラー展開したときの剰余項と、変数変換後の積分区間が、n の増加とともに広がっているものを無限区間に置き換えることによって生じる、外側の積分値の二つからなる。誤差の収束が遅くなるのは、ρが 0 に近いほど、積分区間の食い違いが大きくなるためである。積分区間を広げずに、標準正規分布関数を用いて積分値を評価することで、バイアスをρの連続関数として表現することができた。また、被積分関数のテーラー展開を高次まで計算することで、バイアスの漸近展開近似が得られた。

③バイアス補正

最尤推定量が母数空間上の境界上に値をとる確率 $P(S_1>(p-1)S_2)$ は、ラプラス法を用いて近似計算すると、②の積分値と漸近的に同等であることが分かった。したがって、通常の AIC から、不等式 $S_1>(p-1)S_2$ の定義関数の２倍を減ずることで、リスクの漸近不偏推定量が得られ、バイアスのオーダーはρに関して一様に $O(n^{-1/2})$であることが分かった。通常の AIC では、バイアスのオーダーは $O(n^{-1})$であるが、バイアスは $n^{-1/2}$ のべきで展開されるため、n がそれほど大きくない場合には、その影響が無視できない。$S_1>(p-1)S_2$ の定義関数と、S_1、S_2 の関数の組み合わせだけでは、高次のバイアスを推定することはできないが、不等式を、$S_1/S_2>(p-1)+a\ n^{-1}$ と少しだけずらした定義関数を利用することで、$O(n^{-1})$までバイアスを修正することができた。本研究成果については、論文執筆中である。

（２） 線形混合モデルの変数選択規準

線形混合モデルとして、経時測定データの多項式回帰モデルで、切片項と１次の係数を個体変動を表わす確率変数としたモデルを扱った。説明変数行列を直交化してパラメータ変換することで、目的変量ベクトルの分散共分散行列は観測誤差の分散σ^2と、２次元ランダム係数ベクトルの分散共分散Ξによって表わされる。

①最尤推定量

回帰係数ベクトルの最尤推定量は混合効果モデルと同様に、最小２乗推定量と一致する。ランダム効果の平均ベクトルに関する２次の平方和積和行列を S_1、多項式の２次以上の係数ベクトルに関わる平方和を s_2 とすると、S_1の固有値 l_1、l_2 と s_2 に関する領域 A_1：$s_2<(p-2)l_2$、A_2：$(p-2)l_2 \leqq s_2<(p-1)l_1-l_2$、$A_3$：$(p-1)l_1-l_2 \leqq s_2$ ごとに、σ^2 とΞの最尤推定値が異なる表現を持つ。領域 A_1 ではΞは正定値行列、領域 A_2 では階数１の非負定値行列、領域 A_3 ではゼロ行列として推定される。

②リスクの漸近展開近似

S_2 をバートレット分解し、うまく変数変換することで、予測分布のカルバックライブラー擬距離に基づくリスクは、３つの独立なカイ２乗変数の、①で導出した領域 A_1、A_2、A_3 に対応する領域上の、ある関数の期待値として表現される。混合効果モデルで用いたラプラス近似の手法を、多次元に拡張することで、リスクの漸近展開近似を導出できることが分かったが、近似誤差のオーダーが、母集団パラメータΞに依存する可能性が出てきた。Ξの固有値が重根である場合、l_1、l_2 が、前述のベータ変量の関数として、密度関数の最大化点の近傍でテーラー展開可能でないからである。リスクを表現する関数自体は、テーラー展開可能であるかも知れず、この点は検討中である。

③ランダム係数の個数が３個以上の場合

ランダム係数の個数が q 個の場合は、前述の平方和積和行列が q×q 行列となり、その q 個の固有値を用いて、統計量の領域が q+1 に分割られ、それぞれの領域で、Ξの推定値が、異なる階数の行列として推定される。Ξの固有値が重根を持たない場合には、上記と同様にラプラス近似を利用できると予想されるが、領域ごとに積分をとるべき関数が陽に表現されないため、陰関数の定理を利用す

る必要がある。リスクの評価、バイアス補正を含め、今後の課題である。

（3）一般化線形混合モデルの変数選択規準

ロジスティック回帰において、線形予測子の切片項を確率変数とする場合にすいて検討した。切片項の分布として、ブリッジ分布と呼ばれる分布を想定すると目的変量の周辺分布もロジスティック回帰モデルとなるが、目的変量全体の同時分布が陽に表現されないため、最尤推定量の導出が困難である。リスク評価の見通しはたっていない。

（4）その他関連する研究成果

① ランダム係数を含まない通常の一般化線形モデルの変数選択規準について、高次までのバイアス補正を行った。一般に、AIC 等の高次のバイアス補正は、スコア統計量の高次モーメントの評価が必要であるが、一般化線形モデルについては、指数型分布族を定義する関数の微係数を用いて補正できることがわかった。

② 構造方程式モデルにおいて、共分散行列の縮小推定量を導出するときの、縮小パラメータの選択方法として、一般化情報量規準(GIC)を最小化する方法を提案した。

③ モデル選択規準については、リスクの推定とともに、真のモデルの選択確率が重要な指標である。多変量線形回帰モデルにおいて、AIC が一致性を持つための十分条件を導出した。

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 5 件）

1.Yanagihara,H. Wakaki, H. and Fujikoshi,Y. (2015) A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large Electronic Journal of Statistics, 9, 869-897. 査読あり
DOI:10.1214/15-EJS1022

2. Kamada, A., Yanagihara, H., Wakaki, H. and Fukui, K. (2014) Selecting a shrinkage parameter in structural equation modeling with a near singular covariance matrix by the GIC minimization method, Hiroshima Mathematical Journal, 44, 315-32. 査読あり
http://projecteuclid.org/euclid.hmj/1419619749

3. Imori, S., Yanagihara, H. and Wakaki, H. (2014) Simple Formula for Calculating Bias-Corrected AIC in Generalized Linear Models. Scandinavian Journal of Statistics, 41 (2), 535-555, 査読あり
DOI: 10.1111/sjos.12049

〔学会発表〕（計 15 件）

1. 若木宏文、ランダム係数を持つ GMANOVA モデルの変数選択規準、第 10 回日本統計学会春季集会、2016 年 3 月 5 日、東北大学

2. 若木宏文、A Consistency Property of the AIC for Multivariate LInear Models When the Dimension and the Sample Size are large、The 3rd Institute of Mathematical Statistics Asia Pacific Rim Meeting、2014 年 6 月 29 日～2014 年 7 月 3 日、Howard International House, Taipei, Taiwan

3. 若木宏文、On the consistency of model selection criteria under a high dimensional framework、2014 年度統計関連学会連合大会、2012 年 9 月 11 日～2012 年 9 月 11 日、北海道大学

〔その他〕
ホームページ等
http://home.hiroshima-u.ac.jp/~wakaki/

６．研究組織
(1)研究代表者
若木 宏文（WAKAKI, Hirofumi）
広島大学・理学研究科・教授
研究者番号：９０２１０８５６

(2)研究分担者
なし

(3)連携研究者
なし