様式 C-19

科学研究費補助金研究成果報告書

平成２１年　４月３０日現在

研究種目：基盤研究（C）
研究期間：２００５～２００８
課題番号：１７５００１７７
研究課題名（和文）特異な統計モデルに対するリサンプリング法の適用に関する研究

研究課題名（英文）Resampling method for singular statistical model

研究代表者
今井　英幸（IMAI HIDEYUKI）
北海道大学・大学院情報科学研究科・准教授
研究者番号：１０２１３２１６

研究成果の概要：観測されたデータがどのような集団に属しているかがはっきりしている場合には、どの方法を使って解析すればよいか、また、その手法がどのような性質を持っているかがよく判っている。一方、データの属する集団の形状によっては、今までに使われている手法がそのままでは使えない場合があることも判ってきた。そのような場合に、データを繰返し使って解析をする手法がどのくらい有効であるかを、数値実験と、理論の両面から考察した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００５年度	８００，０００	０	８００，０００
２００６年度	８００，０００	０	８００，０００
２００７年度	８００，０００	２４０，０００	１，０４０，０００
２００８年度	９００，０００	２７０，０００	１，１７０，０００
総　計	３，３００，０００	５１０，０００	３，８１０，０００

研究分野：情報解析学
科研費の分科・細目：総合領域・統計科学
キーワード：ブートストラップ法、モデル選択、最近隣法、正則化、非加法的測度

１．研究開始当初の背景

リサンプリング手法の中でも広く用いられているブートストラップ法は、標本からランダムにブートストラップ標本を生成し、これらを新たな標本とみなして、平均、分散を求めるなどの各種の統計手法を適用することにより、統計的推測を行うものである。ブートストラップ法は、実装が容易であることに加え、近年、計算機の計算速度が著しく向上していることから、多くの分野で応用されている。また、その理論的な解析も進められ、多くの成果が公表されている。

研究開始当初においては、ブートストラップ法が適用されるモデルは、いくつかの正則条件を満足するような、「非特異」なモデルであり、また、理論的な結果も、いくつかの正則条件を満足することが仮定されていた。一方、近年のデータマイニング方の発展に伴い、混合正規分布など、必ずしも正則条件を満足しない、「特異」なモデルに対する統計的推測理論の構築が望まれていた。また、ウェブ上のテキストマイニングなどにおいては、説明変量（特徴量）が数千あるいはそれ以上のデータ解析であり、このような場合には、説明変数の個数に対して十分なサンプルサイズの標本を得ることが難しいケースも増えてきた。

このように、モデルが正則条件を満足しなかったり、一般的な推測理論で満たされる仮定が満たされない場合のブートストラップ

法の性質については、実験的研究、理論的研究ともほとんど行われていなかった。このため、ブートストラップ法を中心とするリサンプリング法の適用に関して、その有効性や、適用の限界を明らかにすることが求められていた。

２．研究の目的

本研究の目的は、リサンプリング手法を特異な統計モデルの母数推定をはじめとする統計的手法に適用した場合の性能を、数値的側面および理論的側面の両面から評価し、明らかにすることが目的である。具体的には、以下の三つの場合を中心に研究を進める。

（１）非加法的な測度を用いた統計モデルにおいては、柔軟なモデル化が可能である反面、推定すべきパラメータの個数が説明変数の増加と共に膨大な数となることが問題点として指摘されている。推定すべきパラメータを少なくするために、パラメータやモデルにいくつかの制約条件を課すことで、パラメータ推定が行われることが多い。そのため、パラメータの信頼領域を構成することが非常に難しく、推定手法の安定性や精度を評価することができない。こうした場合でも、ブートストラップ法を用いることで、推定手法の安定性や精度を評価し、その適用限界を明らかにする。

（２）判別分析において、説明変量の数が非常に多い場合、あるいは、説明変数が少なくても、その中に共変量がある場合には、標本分散共分散行列が特異になったり、あるいは、特異に非常に近くなる。こうした場合、線形判別分析をはじめとする従来手法をそのまま適応すると、小数の観測値が判別結果に非常に大きな影響を与え、判別手法が不安定になる可能性がある。こうしたことを避けるためには、何らかの正則化を用いることが必要となる。分散共分散行列が特異あるいは特異に近い場合でも優れた判別性能を有する手法として正則化判別分析が提案されている。正則化判別分析の判別性能はハイパーパラメータの設定に大きく依存することが指摘されており、適切なパラメータを選択することが問題となる。ここでは、リサンプリング法を応用した情報量規準を用いることで、正則化判別分析のハイパーパラメータ選択法に関する考察を行う。

（３）パターン認識において広く用いられる最近隣法においては，特に説明変量の数が多い場合、探索に非常に時間がかかることが問題である。こうした高次元のデータを適切な低次元空間に射影してから分析することで、計算量や計算時間を大幅に削減することができる。一方で、低次元に射影することによる情報の欠損から、ある確率で最近隣を求めることに失敗する事がある。したがって、低次元に射影することによる計算量の減少と、最近隣を求めることに失敗する確率のトレードオフを評価する必要がある。そのためには、射影したデータの最近隣までの距離の分布を求めなければならない。ブートストラップ法を適用することによる、射影されたデータの最近隣までの距離の分布の近似精度を明らかにする。

３．研究の方法

リサンプリング法を特異な統計モデルに適応する場合の性質と適用限界を正則モデルに適用したときとも比較しながら、数値的に検討する。また、正則モデルの理論的結果を検討し、特異なモデルに対する理論的な検討を行う。

（１）非加法的なモデルにおいて、リサンプリングによってパラメータの信頼領域を構成する。また、非加法的なモデルにおいては、平均、分散といった基本的な統計量の分布も求められていないため、これらを解析的な手法により導出することで、漸近的な信頼領域を得ることができる。信頼領域を比較するために、数値シミュレーションにより、真の信頼領域をもとめ、ブートストラップ法との比較を行う。

（２）正則化判別分析のハイパーパラメータの選択として、ベイズ法によるものがある。しかし、この方法は、事前分布として逆ウィシャーと分布を仮定することにより求められる。この手法と、ブートストラップ法による推定量のバイアスを求めるブートストラップ情報量規準との比較を行う。

（３）最近隣の分布を求めるために、標本からランダムに小数のサンプルを選択して、その標本分布から平均、分散などの統計量を推定する方法が提案されている。これをブートストラップ法を用いる手法に拡張し、また、極値統計に基づく推定量のバイアスを補正することにより、分布の近似精度の向上の図る。

４．研究成果

数値実験により、特異なモデルの統計的推論などにおけるブートストラップ法の有効性を評価した。特に、高次元データの最近隣を高速に探索する場合には、バイアス補正を行うことで、最近隣までの距離の分布を効率よく、かつ高い精度で近似することが可能であることが示された。

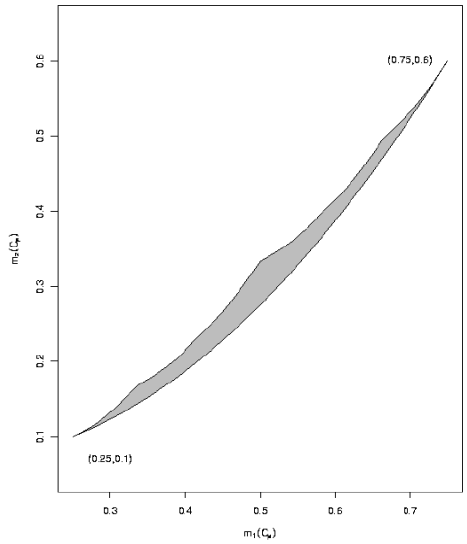

図 1　平均と二次の原点周りのモーメント

（1）非加法的測度を用いた統計モデルのパラメータの安定性と、モデルの評価に関する基礎的な検討を行った。

①モデルの当てはまりの評価をするための基本的な統計量である合計得点の平均と分散を解析的に求め、その存在範囲を明らかにした。非加法的測度を用いた統計モデルに関する事例研究は数多く報告されているが、モデルの当てはまりなど、理論的な性質はほとんど明らかになっていない。この結果を基に非可能的測度を用いた統計モデルに関する推測理論の構築など、多方面への応用が可能

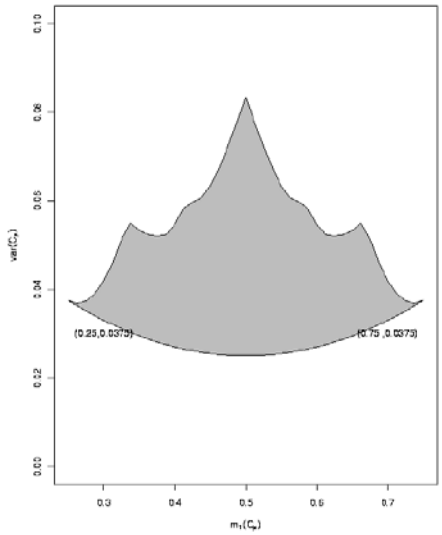

図 2　平均と分散

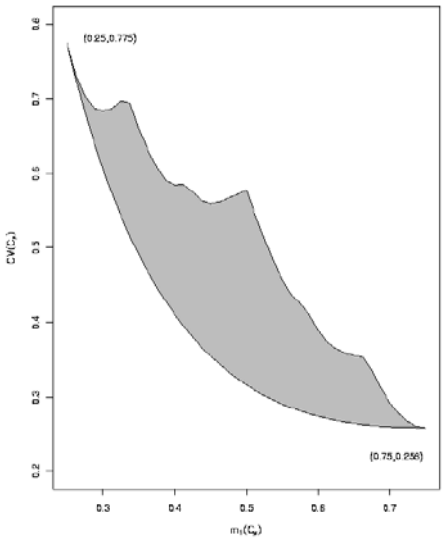

図 3　平均と変動係数

となった。
図１は変量数が３の場合の平均と二次の原点周りのモーメント、図２は平均と分散の存在範囲、図３は平均と変動係数（標準偏差を平均で割ったもの）の存在範囲である。

②パラメータの信頼領域をブートストラップ法によって構成し、シミュレーションによるパラメータの分布から求めた真の信頼領域と比較することでブートストラップ法の性能を評価した。数値実験の結果により推定されたパラメータが母数空間の内点である場合と、母数空間の境界にある場合では、真の信頼領域の広さが大きく異なる。ブートストラップ法による信頼領域の構成でも同様の傾向を示すため、ブートストラップ法による信頼領域の構成が有効であることが示された。また、母数空間のなかでも、端点においては、ブートストラップ法を用いても、満足する結果を得ることができないことも確かめられたため、今後、手法の改良が求められる。

（2）正則化判別分析のハイパーパラメータ選択に関しては、ブートストラップ法を用いた選択法は説明変数の個数によって従来法との特性が異なることがわかった。

①説明変数が少数（おおむね５０個以下）の場合には、ブートストラップ法による選択法と、事前分布による選択法がほぼ同程度の性能を示した。これは、母集団に正規分布を仮定した結果であり、母集団によってどのような違いがあるかは、更に分析が必要である。ただし、ブートストラップ法による選択法では、計算時間は 1/10 程度であり、十分な有

効性があるといえる。

②説明変数が１００を超える場合には、事前分布による選択法は膨大な計算量が必要となり、実用的な時間では計算を終えることができないのに対し、ブートストラップ法では実用的な時間内に、ある程度の精度の高いハイパーパラメータを選択すること可能であることが示された。ただし、ブートストラップ法を用いた場合でもサンプルサイズが小さい場合には選択の精度が悪いため、今後の手法の改良が必要である。テキストマイニングをはじめとするデータマイニングにおいては、説明変数が数千以上である場合も多いため、ここで示された結果は、このような大規模データにおいても正則化判別分析の適用が可能であることを示すものであり、大規模データ解析における有効な結果であるといえる。

（3）多次元の標本を英治言に射影した場合の、最近隣までの距離の分布は、極値統計学の結果から、ワイブル分布に従うことが示される。したがって、ワイブル分布のパラメータをいかに制度よく推定するかが問題となる。リサンプリングによる標本に基づいて推定量を求める場合、推定量にはバイアスがあるため、これを補正して推定することによる、分布の近似を行った。この近似の精度がリサンプリングのサンプルサイズによってどのように変化するかを調べた。また、母集団分布が正規分布以外の分布に従う場合の数値実験も行った。その結果、リサンプリングに基づく近似は母集団が正規分布以外の分布をする場合でも有効であることが示された。最近隣法はパターン認識の中でも基本的な手法であるが、変量の次元が高い場合には計算量が多くなるため、大規模データでの最近隣を精度良く、かつ高速に求める手法が求められている。本研究により提案された近似法を用いることで、大規模なデータでも、十分実用的な計算量で最近隣の候補を選び出すことが可能となり、データマイニングなどへ応用することができる。データを低次元に射影することによる情報の損失があるため、最近隣を取り出すことに失敗することがあるが、この確率を精度良く近似できれば、計算量と誤り確率のトレードオフを評価することが可能となる。

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１２件）

① S. Ikeda and Y. Sato: “Kernel methods for regression model based on variable selection,” International Journal of Knowledge Engineering and Soft Data Paradigms, Vol. 1, 49-62 (2008).（査読有）

② N. Taneichi and Y. Sekiya: “Approximations of the distributions of test statistics for homogeneity of a product multinomial model,” Communications in Statistics, Vol. 37, 1610-1631 (2008).（査読有）

③ A. Tanaka, H. Imai, and M. Miyakoshi: “A unified framework of subspace identification for D.O.A.estimation,” IEICE Transactions on Fundamentals, Vol. E90-A, 419-428 (2007).（査読有）

④ A. Tanaka, H. Imai, M. Kudo and M. Miyakoshi: “Integrated kernels and their properties,” Pattern Recognition, Vol. 40, 2930-2938 (2007).（査読有）

⑤ N. Taneichi and Y. Sekiya: “Improved transformed statistics for the test of independent in contingency tables,” Journal of Multivariate Analysis, Vol. 98, 1630-1657 (2007).（査読有）

⑥ K. Aoki and Y. Sato: “A test statistics in canonical correlation analysis for categorical variables”, Behaviormetrika, Vol. 34, 59-74 (2007)（査読有）

⑦ Y. Sato: “Clustering mixed data using spherical representation,” Knowledge-based Intelligente Information and Engineering

Systems, Vol. 1, 94-101(2006). （査読有）

⑧ 田中章，今井英幸，宮腰政明：“線形制約の一般解によるパラメトリック部分射影フィルタの解釈とアフィン制約付き復元問題への応用”，電子情報通信学会論文誌，Vol. J89-A, 679-681 (2006). （査読有）

⑨ Y. Muto, M. Kudo and T. Murai: “Reduction of attribute values for Kansei representation,” Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 10, 666-672 (2006). （査読有）

⑩ H. Sakurai and M. Taguri: “Test of mean difference in longitudinal data based on block resampling, 2006 Proceedings in Computational Statistics, Vol. 1, 1087-1094 (2006). （査読有）

⑪ 河田 岳大，工藤 峰一，中村 篤祥，外山 淳：“両方向 N-gram 確率を用いた誤り文字検出法，”電子情報通信学会論文誌，Vol. J88-D, 629-635 (2005) （査読有）

⑫ N. Abe and M. Kudo: “Nonparametric classifier-independent feature selection,” Pattern Recognition, Vol. 39, 737-746 (2006) （査読有）

〔学会発表〕（計１２件）

① S. Ohnishi, Y. Yamanoi, and H. Imai: “A weight representation for fuzzy constraint-based AHP,” 12th International Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems, 2008年6月22日～27日,Hotel Amaragua, Malaga Spain.

② A. Tanaka, H. Imai and M. Miyakoshi: “Noisy BBS based on joint diagonalization of differences of correlation matrices,” 10th IASTED international Conferences Signal and Image Processing, 2008年8月18日～20日, Kailua-Kona Hotel, Hawaii USA.

③ A. Tanaka, H. Imai, M. Kudo and M. Miyakoshi, “Optimal kernel in a class of kernels with and invariant metric,” The Joint IAPR International Workshops on Syntactical and Structural Pattern Recognition, 2008年 12 月 4 日～6 日,University of Central Florida, Orland USA.

④ M. Sera, H. Imai and Y. Sato: “Parameter estimation for events in the divided observation periods in a Poisson process,” 18th International Conference on Computational Statistics, 2008年8月24日～29日, University of Porto, Port Portugal.

⑤ A. Tanaka, H. Imai, and M. Miyakoshi: “Selection of correlation matrices for second order statistics based blind source separation”, IEEE Statistical Signal Processing Workshop 2007, 2007年8月２６日～２９日, Monona Terrace Convention Center, Madison, USA.

⑥ Y. Sato: “k-means method for binary and mixed data and related methods,” The 2007 International Association for Statistical Computing-Asian Region Section Special Conference, 2007 年 6 月 7 日～8 日, Seoul Press Center, Seoul Korea.

⑦ K. Aoki, and Y. Sato: “A method for eliminating the horseshoe effect in correspondence analysis,” The 9th Japan-China Symposium on Statistics, 2007年9月25日～28日, 北海道大学.

⑧ S. Ikeda, J. Tsuchiya and Y. Sato: “Kernel regression and variable selection problem,” The 9th Japan-China Symposium on Statistics, 2007年9月25日～28日, 北海道大学.

⑨ N. Taneichi and Y. Sekiya: “Performance of an asymptotic approximation for the distributions of statistics for multinomial homogeneity test,” The 9th Japan-China Symposium on Statistics, 2007年9月25日～28日, 北海道大学.

⑩ H. Sakurai and M. Taguri: “Test of mean difference for paired longitudinal data using circular block bootstrap,” The 56th Session of the International Statistical Institute, 2007 年 8 月 22 日～29 日, Lisboa Congress Center, Lisbon Portugal.

⑪ H. Imai and V. Torra: “Moments of

aggregation operators of Choquet integral family,” 日本計算機統計学会第 20 回大会, 2006 年 5 月 20 日～21 日, 同志社大学.

⑫ K. Aoki and Y. Sato: “A test for redundancy of some items and categories in categorical correlation analysis,” Proceedings of the 55th Session of the International Statistical Institute, 2005 年 4 月 5 日～12 日, Sydney Australia.

〔図書〕(計１件)

① 佐藤義治,「多変量データの分類」, 朝倉書店, 2009 年, 178 ページ.

６．研究組織

(1)研究代表者

今井　英幸 (IMAI HIDEYUKI)

北海道大学・大学院情報科学研究科・准教授

研究者番号：１０２１３２１６

(2)研究分担者

佐藤　義治 (SATO YOSHIHARU)

北海道大学・大学院情報科学研究科・教授

研究者番号：８００９１４６１

工藤　峰一 (KUDO MINEICHI)

北海道大学・大学院情報科学研究科・教授

研究者番号：６０２０５１０１

種市　信裕 (TANEICHI NONUHIRO)

鹿児島大学・理学部・教授

研究者番号：００２０７２００

村井　哲也 (MURAI TETSUYA)

北海道大学・大学院情報科学研究科・准教授

研究者番号：９０２０１８０５

櫻井　裕仁 (SAKURAI HIROHITO)

北海道大学・大学院情報科学研究科・助教

研究者番号：００３３３６２５

(3)連携研究者

なし