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　 ４ 月 ２０ 日現在

機関番号：１４４０１
研究種目：基盤研究(B)（一般）
研究期間：2012～2015
課題番号：２４３００１０６

研究課題名（和文）高次元データにおける多数の仮説の信頼度計算

研究課題名（英文）Computing confidence levels of many hypotheses for high-dimensional data

研究代表者
下平　英寿（Shimodaira, Hidetoshi）

大阪大学・基礎工学研究科・教授

研究者番号：００２９０８６７

交付決定額（研究期間全体）：（直接経費）　　13,700,000円

研究成果の概要（和文）：データからのリサンプリングによって信頼度を計算するブートストラップ法は近似誤差が大きい．高精度な信頼度を計算するために，データのサンプルサイズが変化するときの確率のスケーリング則を利用したマルチスケール・ブートストラップ法や，リサンプリングによって近似誤差を修正するダブルブートストラップ法が提案されている．本研究ではこの二つの方法を同時に適用するマルチスケール・ダブルブートストラップ法を提案して精度がさらに改善することを証明した．確率分布空間の幾何学を用いて，近似誤差は仮説境界の平均曲率や「平均曲率の平均曲率」として表され，これらをマルチスケール・ブートストラップ法が解消することが分かった．

研究成果の概要（英文）：Bootstrap method has a large approximation error for computing a confidence level by resampling from data. For computing confidence levels with higher accuracy, multiscale bootstrap method utilizes a scaling-law of probability when changing the sample size of data, and double bootstrap method adjusts the approximation error by resampling. In this research, multiscale double bootstrap method has been proposed by using these two existing approaches together. We proved that the new method further improves the accuracy. In terms of the geometry of the space of probability distributions, the approximation error of bootstrap is expressed as a “mean curvature” of the boundary surface of hypothesis, and that of double bootstrap is expressed as a “mean curvature of the mean curvature”, and they are removed by multiscale bootstrap.

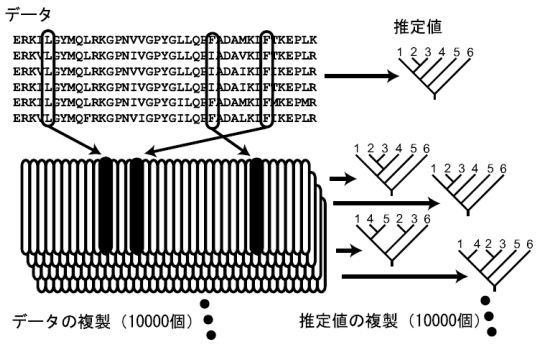
研究分野：統計科学

キーワード：ブートストラップ　リサンプリング　スケーリング則　仮説検定　モデル選択　情報幾何　高次漸近理論　多変量解析

１．研究開始当初の背景

（1）膨大なデータから知識発見を行うために，データマイニングによって非常に多くの仮説が同時に探索されることがある．このような状況では，データに内在するランダムネスの影響が増幅されてバイアスを生じ，誤った発見に導かれやすくなる．これは仮説検定の多重性と呼ばれる効果である．分子進化学を例にして説明する．生物進化の分岐順序を表すラベル付き木は系統樹とよばれ，DNA配列の生物間差異から推定される．比較する生物種の個数が増えると系統樹の個数は指数的に増える．結果として，多数の誤った系統樹のうち一つが「まぐれ」で過度にデータへ適合し，新たな発見をしたように見えてしまう確率が高くなる．

（2）そこでデータのランダムネスが推定値に与える影響を正しく評価することが重要になる．このための一般的な確率シミュレーション技法がブートストラップ法であり，アルゴリズムはきわめて単純で応用が容易である．データからのリサンプリング，すなわちランダムに要素を取り出し複製データを１万回程度生成する．これらに通常のデータ解析を繰り返し適用して得られた系統樹の集合を調べ，データが仮説を支持する頻度を信頼度として利用する．これがブートストラップ確率である．ところがブートストラップ法にはバイアスがあり信頼度の精度が不十分であることが分かってきた．



（3）下平は頻度論の立場でブートストラップ確率のバイアス補正を行い，信頼度を高精度で計算するアルゴリズムを開発してきた．キーとなるアイデアは，データのスケールを変化させたときのブートストラップ確率の変化率から，仮説領域の幾何学的な情報（データ点までの距離や仮説境界の曲率）を引きだすことである．これらの情報が得られれば，直ちに信頼度を精確に計算できる．ここでデータのスケールというのはランダムネスの程度を表すスカラー量であり，サンプルサイズの平方根に反比例する．

（4）観測データのサンプルサイズを n，ブートストラップリサンプリングでランダムに生成する複製データのサンプルサイズを m とする．通常のブートストラップ法では $m=n$ である．マルチスケール・ブートストラップ法では m を変化させる．このような方法は一般に m-out-of-n bootstrap 法と呼ばれる．データからランダムに要素を取り出して複製データを生成するときに，通常は n 回反復して n 個の要素を取り出すが，それを m 回反復して m 個の要素を取り出すようにブートストラップ法のプログラムを修正するだけであり，容易に実装できる．ブートストラップ法は重複を許したリサンプリング，すなわち，同じ要素を複数回取り出すことを許す方法である．従って任意の整数 $m>0$ に変更可能である．m を変更すると，複製データのランダムネス（バラツキ）が変化する．大きな m の複製データから計算した推定値は分散が小さくなり，小さな m の複製データから計算した推定値は分散が大きくなる．この分散は m に反比例するので，観測データから計算した推定値の分散を調べるために本来は $m=n$ にする必要がある．

（5）ところが下平の提案したマルチスケール・ブートストラップ法の理論では，精度の高い信頼度を計算するためには，負のサンプルサイズ $m=-n$ とすることが証明される．m は取り出す要素の個数であるから，正の整数と考えるのが普通であるが，理論上は負の値が最適と証明される．この常識外れなところがマルチスケール・ブートストラップ法のオリジナリティといえる．実際の計算では，複数の m（もちろん正）においてブートストラップ法を実行してその結果を $m=-n$ へ外挿する．このときブートストラップ確率をそのまま外挿するのではなく，

$$p(\sigma^2)=\Phi\left(\sigma\,\Phi^{-1}\left(BP(\sigma^2)\right)\right)$$

を形式的に $\sigma^2=-1$ へ外挿する．ただし $BP(\sigma^2)$ はブートストラップ確率，σ はデータの相対的なスケールを表す．$m=n$ のとき $\sigma^2=1$ とおけば相対的な分散は

$$\sigma^2=\frac{n}{m}$$

である．$m=n$ とおいた通常のブートストラップ確率は $p(1)$ と表され，ベイズの事後確率に相当する．一方，頻度論的信頼度（p-値）は $m=-n$ とおいて $p(-1)$ で与えられる．

２．研究の目的

下平の提案したマルチスケール・ブートストラップ法はデータのサンプルサイズが変化するときの確率のスケーリング則を利用して高精度な信頼度を計算する手法であり，主に生命科学，とくに分子進化系統樹推定や遺伝子発現解析で用いられている．ところが高次元データにおける多数の仮説の場合は

必ずしも有効に機能しない場合もある．そこで理論研究と応用研究の両面からリサンプリングによる信頼度計算法を発展させる．

３．研究の方法

（１）マルチスケール・ブートストラップ法ではサンプルサイズ n が十分に大きいことを仮定した漸近理論を用いている．近似誤差が $O(n^{-k/2})$ のとき，その漸近精度は k 次という．通常のブートストラップ法は１次の精度しかないが，マルチスケール・ブートストラップ法は３次の精度がある．一方，ブートストラップ法を使ってブートストラップ確率の誤差を修正するダブルブートストラップ法も３次の精度であることが以前から知られている．そこで，この両者を組み合わせてより高い精度が得られるかを検討する．

（２）マルチスケール・ブートストラップ法を並列計算できるように実装して高速化する．CPU のマルチコア並列実装だけでなく GPGPU による並列計算を試みる．

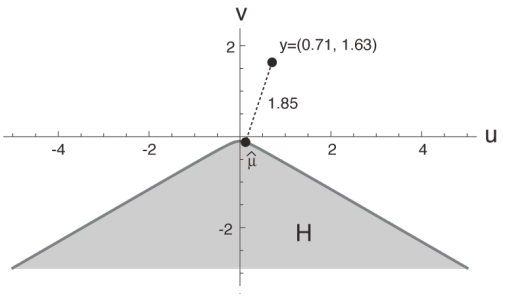
（３）統計的因果推測，機械学習，ネットワーク解析，L1 正則化などの応用にも取り組んで，新たな理論研究へ結びつける．

４．研究成果

（１）ダブルブートストラップ法にマルチスケール法を適用して，マルチスケール・ダブルブートストラップ法を提案し，これが４次の精度であることを証明した．二つのアプローチを同時に適用することによって，この新しい方法は従来法の精度をさらに高めたことになる．

（２）ダブルブートストラップ法では仮説境界への射影を計算してリサンプリングを行うが，もし射影計算にズレがあると漸近精度が３次から２次に落ちてしまう．ところがマルチスケール・ダブルブートストラップ法では射影計算にズレがあっても漸近精度が４次に保たれることも証明できた．

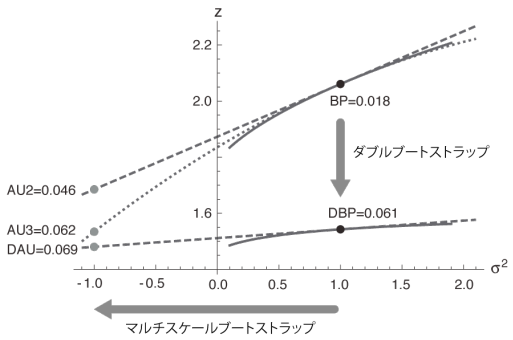
（３）簡単な数値例でアルゴリズムを説明する．下図のような２次元の仮説領域 H について仮説検定を行いたい．



各成分が分散１の正規分布(２次元)を考え，観測値 y を得たときの信頼度を計算する．この仮説境界は滑らかな曲面（2 次曲面）であるが先端の曲率が大きいため，ぴったり重なる凸錐に仮説領域を置き換えてみると，多重比較法が適用できる．このときの確率値は 0.069 であり，y が頂点（least favorable configuration）に近いため，これが信頼度の目安となる．通常のブートストラップ確率(BP)を計算すると 0.018 であり，0.069 に比べてかなり小さい．有意水準 0.05 で検定する場合，多重比較法ならば有意とならず仮説を棄却しないが，ブートストラップ確率では有意となって仮説を棄却してしまう．マルチスケール・ブートストラップ法では

$$z = -\Phi^{-1}(p(\sigma^2))$$

をプロットして $\sigma^2 = 1$ から $\sigma^2 = -1$ へ外挿して信頼度を計算する．1 次式で外挿すると AU2=0.046，2 次式で外挿すると AU3=0.062 となって，BP にくらべて多重比較法に近づくことがわかる．



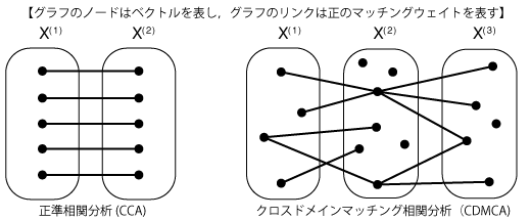
一方，ダブルブートストラップ法で信頼度を計算すると DBP=0.061 であった．マルチスケール・ダブルブートストラップ法で DBP を $\sigma^2 = 1$ から $\sigma^2 = -1$ へ外挿して信頼度を計算すると，DAU=0.069 となり，もっとも多重比較法に近い値となった．なお，y が頂点から遠くなると多重比較法は保守的で必要以上に信頼度が大きくなってしまう問題があるが，BP，AU，DBP，DAU はそのような問題無くほぼ最適な信頼度を与える．

（４）マルチスケール・ダブルブートストラップ法を漸近理論で証明する過程で，確率分布の空間におけるブートストラップ法の幾何学について理解が深まった．観測値 y から仮説境界の射影点までの符号付き距離や，射影点における境界曲面の平均曲率を考える．ブートストラップ法の誤差の原因が平均曲率であることは既存研究でも知られていたが，ダブルブートストラップ法の誤差の原因が「平均曲率の平均曲率」であることを本研究で明らかにした．マルチスケール・ブートストラップ法の適用によって，ブートストラップ法における平均曲率，およびダブルブートストラップ法における「平均曲率の平均曲率」が解消される．

（５）マルチスケール・ブートストラップ法の並列計算を実装して高速化に一定の効果を確認できた． 従来からの CPU マルチコア並列実装に加えて，本研究では GPGPU による並列実装を行い，CPU と GPGPU のハイブリッド並列が有効であることが分かった．

（６）応用研究において次の成果を得た．因果推論について潜在共通原因がある場合の研究を行い，脳機能イメージングデータから脳内因果ネットワーク構造のどの部分に有意な違いがあるかを調べた．機械学習について，さまざまなデータ変換に対して不変なダイバージェンスのクラスを導出して，ロバスト統計へ適用した．ガウシアングラフィカルモデルによるネットワーク推定において Lasso と一種のグループ Lasso をつなぐ正則化項のクラスを検討して，スパースなグラフの推定精度が高くなるような工夫を行った．ネットワークデータの統計解析における優先的選択関数のノンパラメトリック推定法を考案して，YouTube のフォロー関係データ等の分析を行った．

（７）ブートストラップの応用研究の過程で情報統合の高次元多変量解析の着想を得た．画像や文書など様々な情報源をドメインと呼び，各ドメインから得られるデータベクトルと，ベクトル間の関連を表すマッチング行列を多変量解析する手法である．



この多変量解析にグラフのリサンプリング法を適用したクロスバリデーションを考案して，これが予測誤差の不偏推定量になることを高次元の漸近理論によって証明した．この手法を文書ベクトルや画像特徴量ベクトルに適用した．とくに Flickr 画像のタグ検索や多言語文書ベクトルの翻訳タスクへの応用も試みて，その有効性を確認した．

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１９件）

(1) Hidetoshi Shimodaira, Cross-validation of matching correlation analysis by resampling matching weights, Neural Networks, 75, 126-140, 2016, 査読有
DOI:10.1016/j.neunet.2015.12.007

(2) Kei Hirose, Yukihiro Ogura, Hidetoshi Shimodaira, Estimating Scale-Free Networks via the Exponentiation of Minimax Concave Penalty, Journal of the Japanese Society of Computational Statistics, 28, 139-154, 2015, 査読有
DOI:10.5183/jjscs.1503001_215

(3) Thong Pham, Paul Sheridan, Hidetoshi Shimodaira, PAFit: A Statistical method for measuring preferential attachment in temporal complex networks, PLOS ONE, 10, e0137796, 2015, 査読有
DOI:10.1371/journal.pone.0137796

(4) Takafumi Kanamori, Hironori Fujisawa, Robust Estimation under Heavy Contamination using Unnormalized Models, Biometrika, 102, 559-572, 2015, 査読有
DOI:10.1093/biomet/asv014

(5) Hidetoshi Shimodaira, Higher-order accuracy of multiscale-double bootstrap for testing regions, Journal of Multivariate Analysis, 130, 208-223, 2014, 査読有
DOI: 10.1016/j.jmva.2014.05.007

(6) Shohei Shimizu and Kenneth Bollen, Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-Gaussian distributions, Journal of Machine LearningResearch, 15, 2629-2652, 2014, 査読有

(7) Takafumi Kanamori, Hironori Fujisawa, Affine Invariant Divergences associated with Proper Composite Scoring Rules and their Applications, Bernoulli, 20, 2278-2304, 2014, 査読有

(8) Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen and Takashi Washio, ParceLiNGAM: A causal ordering method robust against latent confounders, Neural Computation, 26, 57-83, 2014,査読有
DOI:10.1162/NECO_a_00533

(9) Takafumi Kanamori, Atsumi Ohara, A Bregman extension of quasi-Newton updates II: analysis of robustness properties, Journal of Computational and Applied Mathematics, 253, 104-122,2013, 査読有
DOI:10.1016/j.bbr.2011.03.031.

(10) Takafumi Kanamori, Akiko Takeda, Taiji Suzuki, Conjugate Relation between Loss Functions and Uncertainty Sets in

Classification Problems, Journal of Machine Learning Research, 14, 1461-1504, 2013, 査読有

(11) Akiko Takeda, Hiroyuki Mitsugi, Takafumi Kanamori, A Unified Classification Model Based on Robust Optimization, Neural Computation, 25, 759-804, 2013, 査読有
DOI: 10.1162/NECO_a_00412

(12) Paul Sheridan, Yuichi Yagahara, Hidetoshi Shimodaira, Measuring preferential attachment in growing networks with missing-timelines using Markov chain Monte Carlo, Physica A: Statistical Mechanics and its Applications, 391, 5031-5040, 2012, 査読有
DOI:10.1016/j.physa.2012.05.041

(15) Shohei Shimizu, Joint estimation of linear non-Gaussian acyclic models, Neurocomputing, 81, 104-107, 2012, 査読有
DOI: 10.1016/j.neucom.2011.11.005

〔学会発表〕(計３９件)

(1) 下平英寿, データベクトル間マッチングの多変量解析とそのクロスバリデーション, 統計関連学会連合大会, 2015/9/8, 岡山大学 (岡山)

(2) 奥野彰文, 福井一輝, 下平英寿, 多ドメインでのマッチング相関分析のロバスト化とその応用, 統計関連学会連合大会, 2015/9/8, 岡山大学 (岡山)

(3) 永田晴久, 下平英寿, グラフ構造のブートストラップ法とクラスター係数についての漸近評価, 統計関連学会連合大会, 2015/9/9, 岡山大学 (岡山)

(4) 福井一輝, 奥野彰文, 下平英寿, マッチング相関分析を用いた画像-マルチタグ間の相互検索, 第 18 回画像の認識・理解シンポジウム MIRU 2015, 2015/7/29, ホテル阪急エキスポパーク (大阪府吹田市)

(5) Haruhisa Nagata, Hidetoshi Shimodaira, Bootstrap method for networks and its properties in random graphs, The 7th International Conference of the ERCIM WG on Computational and Methodological Statistics (ERCIM 2014), 2014/12/7,Pisa (Italy)

(6) 永田晴久,下平英寿, ランダムグラフにおけるグラフ構造のブートストラップの性質, 2014/9/15, 統計関連学会連合大会, 東京大学(東京都文京区)

(7) 小倉 幸弘, 廣瀬 慧, 下平 英寿, ガウシアングラフィカルモデルのスパース推 定における正則化項の検討, 日本計算機統計学会 第 27 回シンポジウム, 2013/11/16,崇城大学ホール(熊本県熊本市)

(8) Thong Pham,Paul Sheridan,下平 英寿, 複雑ネットワークの生成モデルにおける優先的選択関数のノンパラメトリック推定, 統計関連学会連合大会, 2013/9/9, 大阪大学(大阪府豊中市)

(9) 下平英寿, マルチスケール・ダブルブートストラップ法による領域の検定, 統計関連学会連合大会, 2013/9/11, 大阪大学(大阪府豊中市)

(10) 永田 晴久,下平 英寿, 階層型クラスタリングにおけるブートストラップ法の GPGPU 化と CPU/GPU 負荷分散, 統計関連学会連合大会, 2013/9/10, 大阪大学(大阪府豊中市)

(11) Hidetoshi Shimodaira, Higher-order accuracy of multiscale double-bootstrap resampling for testing regions, Bernoulli Society Satellite Meeting to the ISI World Statistics Congress (招待講演), 2013/9/2, 東京大学(東京都文京区)

(12) 永田 晴久, 下平 英寿, GPGPU を用いた階層型クラスタリングの信頼度計算, 統計関連学会連合大会, 2012/09/11, 北海道大学 (北海道)

(13) Hidetoshi Shimodaira, Converting a Bayesian confidence value into a frequentist by reversing the sign of the data length, IMS Asia Pacific Rim Meeting (IMS-APRM)(招待講演), 2012/07/03, つくば国際会議場 (茨城県つくば市)

６．研究組織
(1)研究代表者
下平　英寿 (SHIMODAIRA, Hidetoshi)
大阪大学・大学院基礎工学研究科・教授
研究者番号：００２９０８６７

(2)研究分担者
清水昌平 (SHIMIZU, Shohei)
大阪大学・産業科学研究所・准教授
研究者番号：１０５０９８７１

金森敬文 (KANAMORI, Takafumi)
名古屋大学・大学院情報科学研究科・准教授
研究者番号：６０３３４５４６