

論文審査の結果の要旨

論文題目： **Essays on Shrinkage Methods in Statistics**

題目和訳： 統計学における縮小化法に関する研究

氏名： 加藤 賢悟

本論文は、近年統計学のさまざまな場面で注目されている縮小化法に関して、著者のこれまでの研究成果をまとめたものである。論文は主に次の2つの問題から構成されており、1つは、線形回帰モデルにおける説明変数の選択と回帰係数の縮小推定法に関する問題であり、もう1つは縮小型予測分布の導出とその理論的有効性の証明に関する問題である。具体的には以下のような章立てで構成されている。

第1章 序

第2章 縮小推定における自由度について

2.1 節 序

2.2 節 予測リスクの不偏推定

2.3 節 主な結果 (ダイバージェンス形式, 自由度)

2.4 節 例 (Lasso, Fused Lasso, Group Lasso)

2.5 節 $p > n$ の場合への拡張 (p は回帰係数ベクトルの次元, n はデータ数)

2.6 節 考察

2.7 節 定理の証明

第3章 区分的線形な損失関数を用いた ℓ_1 正則化問題の解法

3.1 節 序

3.2 節 線形計画法

3.3 節 パラメトリック単体法 (記述, 解の特徴)

3.4 節 具体的な計算の方法 (線形方程式の解法, 被約コスト・ベクトルの計算, 逆行列の更新, 計算コスト)

3.5 節 解が変化する回数

3.6 節 ブロック毎の ℓ_∞ ペナルティー

3.7 節 考察

第4章 ℓ_1 ペナルティーを用いた分位点回帰の損害保険データへの応用

4.1 節 序

4.2 節 ℓ_1 ペナルティーを用いた分位点回帰 (分位点回帰, ℓ_1 ペナルティーによる分位点回帰, チューニング・パラメータの選択)

4.3 節 損害保険データへの応用 (データ, 結果)

4.4 節 考察

4.5 節 定理の証明

第5章 多変量正規分布の改良型予測分布

5.1 節 序

5.2 節 主な結果 (左及び右不変事前分布を用いた予測分布, 改良型予測分布, 数値実験)

5.3 節 定理の証明 (star と dispersive 順序とそれらを用いた証明方法)

5.4 節 正規線形モデルへの応用

5.5 節 考察

5.6 節 補足

第6章 学位論文の要約

まず, 本論文の第2章から第4章までの内容を紹介する. 線形回帰モデルにおける回帰係数の推定問題はかなり古典的で基本的な問題であるが, その問題に対して近年Lassoと呼ばれる新しいアプローチが考案され, その性質について注目されてきた. Lassoというのは, 縮小化法の1つで,

$$\text{“}2 \text{乗損失関数} + (\text{チューニングパラメータ}) \times (\ell_1 \text{ペナルティ)}\text{”}$$

の最小化問題の形で定式化される. 一般に,

$$\text{“損失関数} + (\text{チューニングパラメータ}) \times (\text{ペナルティ)}\text{”}$$

の最小化問題から導出される推定値は, ペナルティ項のない無制約な推定値に比べて何らかの意味でより小さい推定値を返すことから, このような方法は縮小化法もしくは正則化法と呼ばれる. 縮小化法を用いる利点として, 回帰係数をいくらか縮小することによってモデルの予測精度の向上が期待できる点と, 特定のペナルティを用いることで推定と変数選択を同時に実行できる点があげられる.

Lassoは, Tibshirani (1996) により提案されたもので, 2乗損失関数に ℓ_1 ペナルティを課した評価関数の最小化の解で定義される. ℓ_1 ペナルティの性質から, Lassoはいくつかの係数を正確にゼロに縮小する. すなわち, Lassoは推定と説明変数の選択を同時に実行できる手法となっている. また, Lassoに対しては解のパスを計算する効率的なアルゴリズムが知られており, 計算量の観点からも, Lassoが従来の変数選択法と比べて魅力的な手法であるといえる. 近年, 特に生物データなどは, 共変数の数が従来データと比較して極めて大きくなる傾向にある. 従って, Lassoの持つ計算上の利点は非常に重要な特徴となる.

Lassoが提案されて以降, 用途に応じた様々なペナルティが多くの研究者によって提案されている. また, 2乗損失以外の損失関数に対して ℓ_1 ペナルティなどを課す試みも多く見受けられる. このような縮小化法を実際に行う際, 以下の点が課題となる:

- ・ チューニングパラメータの選択
- ・ 解のパスの計算
- ・ 実データに対するパフォーマンス

本論文の第2章から第4章はそれぞれ上記の課題に焦点を当てて議論している.

まず第2章では、凸ペナルティ付き最小2乗法に対して、自由度と呼ばれる量の不偏推定量を微分幾何的なアプローチを用いて導出している。導出された推定量を用いることによって、マロウズの C_p 規準や赤池情報量規準AICといったチューニングパラメータの選択規準が導出される。Lassoに対しては、Zou et al. (2007) が $p \leq n$ の下で自由度の不偏推定量を導出しているが、第2章の内容は彼らの結果を一般の凸ペナルティの空間へ拡張したものである。また、共変量の数 p が標本数 n よりも多いケース(すなわち、 $p > n$ のケース)において、Lassoの自由度の不偏推定値を明示的に計算する方法を示している。

第3章は区分的に線形な損失関数と ℓ_1 ペナルティまたはブロック ℓ_1 ペナルティを持つ正則化問題の計算上の側面を扱っている。区分的に線形な損失関数としては、分位点回帰において用いられる“check”関数や、SVMにおいて用いられる“hinge”関数などがあげられる。第3章では、この正則化問題を線形計画問題の枠組みで記述し、パラメトリック改訂単体法を用いて解のパスを計算する安定的かつ効率的な手法を与えている。また、シミュレーション実験を通して、正則化パスの変化点の個数が標本数・共変量の数を変えたときにどのように変化するかについて調べている。

第4章では分位点回帰と ℓ_1 ペナルティ付き分位点回帰を用いて自動車保険のデータ解析を行っている。 ℓ_1 ペナルティ付き分位点回帰を用いると、例えば高い保険請求額に影響を与えるリスク要因を特定化することができるなど、実用上非常に役に立つことがわかる。また、第4章ではデータ解析に加えて、分位点回帰と ℓ_1 ペナルティ付き分位点回帰の理論的な性質もまとめている。

第5章の内容は、第2章から第4章までの内容とは異なり、縮小型予測分布の構成とその有効性の証明を行い、縮小化法が予測分布の構成においても有効な手法であることを示している。具体的には、平均及び分散が未知の d 次元の多変量正規分布について将来のデータが従う分布(予測分布)を予測するという問題で、Kullback-Leibler情報量を損失関数として用いると、統計的決定理論の枠組みで議論することができる。通常の間推定や点予測の問題をもう1段階難しくした問題として捉えることができ近年活発に研究されている分野の1つである。分散が既知のときには、事前分布として平均に一様分布を仮定したときのベイズ予測分布がミニマックス解となる。これに対して、Komaki (2001)は縮小型事前分布に基づいたベイズ縮小予測分布を考え、それが $d \geq 3$ のとき上述のミニマックス解を改良することを示した。ここでいう“縮小型事前分布”とは一様分布に比べて原点方向により大きな重みをおいた分布であり、その結果生ずる縮小型ベイズ予測分布はミニマックス解を原点の方向へ縮小した形をしている。点推定の世界では、 $d \geq 3$ のときにミニマックス解が改良されることは“スタイン現象”と呼ばれて多くの研究成果が出されてきたが、Komaki(2001)の結果はスタイン現象が分布予測においても現れることを示していて興味深い。

第5章は、分散が既知のときのKomakiの結果が分散未知の場合へ拡張できることを示している。まず、平均と分散に関して右不変な事前分布に基づくベイズ予測分布が定数リスクを持つミニマックスな予測分布となることを示している。次に平均と分散に関して縮小型事前分布を想定し、それに基づくベイズ予測分布が $d \geq 3$ のときミニマックス予測分布を改良することを証明した。この証明では、従来の点推定の文脈では用いられてこなかった確率順序のテクニックが取り入れられて、斬新なアイデアに基づいて証明に成功している。

「評価」

Lassoは“2乗損失関数 + (チューニングパラメータ) × (ℓ_1 ペナルティ)”の最小化問題であるが、第2章では、かなり一般の閉凸な制約集合 K を考え、“ K に入る回帰係数ベクトルに関して2乗損失関数を最小化する問題”を扱っている。もちろん、これは、Lassoの一般化になっているが、母数制約下での推定問題やLasso型の様々な縮小推定の問題 (Fused Lasso, Group Lasso) を統一的に扱うことができる。このような一般的な問題設定において予測量の自由度を計算するために、筆者は微分幾何の道具を用意し、自由度の不偏推定量を制約集合 K に関する幾何的な量を用いて表すことに成功している。Zou et al. (2007) はLassoの自由度の導出を行っているが、第2章のアプローチは彼らの方法とはかなり異なるものであり、しかも微分幾何の知識を用いて一般の制約集合 K にまで拡張できる理論を構築している。この分野はスタンフォード大学を中心に日進月歩の勢いで研究されつづけており、その最先端の様々な手法を統一的に扱える理論を築いたことは極めて高く評価される。またそうした理論を建てることのできた筆者の数学的能力の高さも評価される。

近年、特に生物データなどは、共変量の数が従来のデータと比較して極めて大きくなる傾向にあり、こうした高次元データの解析に関心が集まりつつある。筆者は、 $p > n$ のケースに対しても、Lassoに対しては自由度の明示的な不偏推定量を導出している。このことは、 $p \leq n$ のケースしか考察できなかった従来の限界を超えた大きな進歩であり、Lassoの持つ計算上の利点を考慮すると重要な貢献になると思われる。

第2章が理論的な研究成果であるのに対して、第3章と第4章は応用上の貢献を与えている。第3章では、実際計算させるため正則化問題を線形計画問題の枠組みで記述しパラメトリック改訂単体法を用いた計算アルゴリズムを提案している。特に、分位点回帰で用いられる区分的に線形な損失関数と ℓ_1 ペナルティを持つ正則化問題について、問題の特性を活かした具体的な計算上の工夫を提案している点が評価される。例えば ℓ_1 ペナルティ付き分位点回帰に対しては、Li and Zhu (2008) が正則化パスを計算するアルゴリズムを提案しているが、彼らのアルゴリズムは正則化パスの一意性といった強い仮定に依存している。これに対して、パラメトリック単体法はデータに対する仮定を必要としない点で彼らのアルゴリズム

ムより有利である。第4章では、第3章で提案した計算アルゴリズムを用いて自動車保険データの分位点回帰分析を行っており、 $n = 6113$, $p = 9$ という大きなデータにもかかわらず、短時間で計算結果を出すことができている。しかも実際の側面からも損害保険の高額請求に対するリスク要因として意味のある説明変数を選択することに成功している。第3章と第4章の内容は、計算アルゴリズムの構築とプログラミングに関する筆者の能力の高さを示しており、第2章のような数学的な理論展開にとどまらず、計算上の工夫と意味のある実証分析を行っている点も高く評価できる。

第5章で扱った問題は、この分野の専門家の間では解くのが非常に難しい問題として認識されてきた、いわゆる **open problem** であった。実際、最終的に示すべき不等式が自由度の異なる2つの関数の間の評価になっているため、従来のテクニックが利用できない。そこで、筆者は確率順序という、この分野では誰も用いてこなかった斬新な方法を導入し、いくつかの不等式の理論を利用することによって見事に問題を解くことに成功している。従って、**open** で困難な数理統計的問題の解法に成功している点で極めて優れた研究成果であると評価される。また、問題の本質を見極めて適切な解決方法の筋道を立て、大変困難な中でも挑戦し続けて証明を完成させることができたことは、研究者として大成する資質があることを感じさせる。

以上、見てきたように、本論文は、線形回帰モデルや分位点回帰モデルにおける変数選択及び回帰係数の縮小推定に関する理論と応用において優れた研究成果をまとめるとともに、縮小型予測分布の理論においても極めて優れた研究成果を与えており、審査委員会は申請論文が博士（経済学）の学位にふさわしいものであると評価する。

審査委員
久保川達也
國友直人
矢島美寛
竹村彰通
大森裕浩