

学 位 論 文 題 名

Model Selection with Empirical Bayes Criteria in L_1 -Regularization

(L_1 正則化法における経験ベイズ規準を用いたモデル選択に関する研究)

学位論文内容の要旨

自然科学や社会科学の諸分野において、不確实现象の解明と予測・制御、そして新たな知識発見のための手段として、いわゆる“モデル”が利用されている。ここで、現象を特徴付ける真の確率分布を近似するために仮定した、未知の母数(パラメータ)を含むパラメトリックな密度関数族のことを“モデル族”、モデル族に含まれる未知パラメータを観測データに基づく推定値に置き換えた一つの密度関数を“統計モデル”、現象を適切に表現する統計モデルを構築する過程のことを“統計的モデリング”と呼ぶことにする。統計理論においては、一般に何らかの統計モデルを一つ設定した上で推測の構造、性質、良さ等を議論する。一方、観測されたデータに対して、唯一の統計モデルしか考えられないといった状況はむしろ稀であり、一つの統計モデルに絞り切れないことが多い。したがって、現象発生メカニズムを捉えるためには、当該分野の専門的知識を集約して構築された複数の統計モデルを何らかの基準に基づいて評価し、適切な統計モデルを選択する必要性が常に生じる。これが統計的モデル選択の問題である。

統計的モデル選択における重要な課題の一つが、回帰分析における変数選択の問題である。回帰分析とは、応答変数 Y と説明変数の組 $X_1, X_2, \dots, X_p$ との関係を経験モデルで表現する手法である。この場合、説明変数の組 $X_1, X_2, \dots, X_p$ を与えたときの応答変数 Y の確率的挙動、およびその条件付き期待値に対して、なんらかの確率構造を仮定することにより、一つの回帰モデル族が決定される。実際の回帰分析においては、以下の2つの観点が重要である。

- (1) 将来得られるデータをできるだけ正確に予測する。
- (2) 情報抽出しやすい簡潔な形で統計モデルを表現する。

近年、(1)と(2)の両方の観点から、 L_1 正則化法が注目されつつある。 L_1 正則化法の最大の利点は、パラメータの L_1 ノルムの正則化項を加えた最小二乗法や最尤法を考えることによって、モデル族の選択とパラメータ推定を同時に行う点にある。 L_1 正則化法における統計モデルの選択に際しては、正則化パラメータと呼ばれるハイパーパラメータの選択が本質的であり、これらの選択が最終的な統計モデルの予測精度やモデルの複雑さを決定する。したがって、いくつかの正則化パラメータを用意し、各々の正則化パラメータの下で推定された複数の統計モデルを、何らかの基準に基づいて評価し、その中から適切な統計モデルを選択する必要性が生じる。しかし、これら統計モデルの選択には、従来提案された情報量規準 AIC やベイズ情報量規準 BIC などの最尤法におけるモデル評価規準を直接適用できないという問題がある。よって、現在、 L_1 正則化法におけるモデルの評価規準が必要とされている。

以上の背景により、本論文では、経験ベイズ法の枠組みから、 L_1 正則化法におけるモデル評価規準を構成し、この規準に基づくモデル選択の理論を構築することを目的とする。経験ベイズ法とは、データの周辺尤度を最大化するように、事前分布のハイパーパラメータを設定する手法であり、近年さまざまな分野で活用されている。評価規準の導出の際に問題となるのは、全パラメータ空間における周辺尤度が微分不可能な領域を含むために近似できない点である。そこで、微分不可能な領域の測度が 0 であること等に着目し、微分可能な部分空間における周辺尤度の近似としてモデル評価規準を与える。これらの提案手法の有効性を数値実験、実データの解析を通じて検証する。

本論文は、全 6 章から構成される。各章の概要は以下の通りである。

第 1 章では、本論文の背景、目的、および構成について述べている。

第 2 章では、本論文に関係する先行研究に加え、本論文で提案するモデル評価規準の導出、および非線形モデリング手法に必要な数学的な準備を与えている。具体的には、未知パラメータがグループ構造を有しない場合、未知パラメータがグループ構造を有する場合の二つの仮定の下で、従来までに提案された L_1 正則化法とその定式化について述べている。また、経験ベイズ法に基づく一般的なモデル選択の枠組み、および従来までに提案されたベイズ型モデル評価規準についても概括している。さらに、第 5 章で使用する動径基底関数ネットワークに関して、その数学的な性質、およびハイパーパラメータ推定法などについて述べている。

第 3 章では、パラメータがグループ構造を有しない場合の L_1 正則化法における経験ベイズ型モデル評価規準を提案している。まず始めに、 L_1 正則化法を経験ベイズ推定の枠組みで定式化した下で、その周辺尤度の近似として、Lasso 型ベイズ情報量規準 (LBIC)、Adaptive Lasso 型ベイズ情報量規準 (aLBIC) を導出している。また、導出されたモデル評価規準の例として、 L_1 正則化法における一般化線形モデル族に対する経験ベイズ規準を与えている。さらに、導出したモデル評価規準について、数値実験、実データの解析を通してその有効性を検証している。検証においては、提案された評価規準に基づくモデル選択法を、予測精度とパラメータ数の観点から、従来の簡便法によるモデル選択法と比較している。

第 4 章では、パラメータがグループ構造を有する場合の L_1 正則化法におけるベイズ型モデル評価規準を提案している。前章と同様に、パラメータがグループ構造を有する場合についての L_1 正則化法のベイズ的解釈を与えた上で、周辺尤度の近似として、Group Lasso 型ベイズ情報量規準 (GLBIC)、Adaptive Group Lasso 型ベイズ情報量規準 (aGLBIC) を導出している。さらに、導出されたベイズ型モデル評価規準の例として、一般化線形モデル族に対する経験ベイズ規準を与えている。また、導出したモデル評価規準については、ANOVA モデル、多項回帰モデルの下での数値実験、実データの解析を通して、従来の簡便法によるモデル選択法と本提案手法を比較している。

第 5 章では、本論文で提案されたモデル評価規準の応用として、 L_1 正則化法に基づく非線形回帰モデルのパラメータ推定法、およびモデル選択法を提案している。特に、モデルの期待値の非線形構造を表現するために動径基底関数ネットワークモデルを仮定し、そのパラメータ推定を L_1 正則化法により行なっている。その際、正則化パラメータの選択、動径基底関数のハイパーパラメータの選択がモデルの構築において重要である。この選択法について、前章までに提案されたモデル評価規準を適用した手法を提案している。また、数値実験、実データの解析を通じて、従来提案された手法との比較を行い、その有用性を検証している。

最後に第 6 章で、本論文の総括とともに、本論文で提案する理論についての有用性を示している。

学位論文審査の要旨

主査	教授	水田正弘
副査	教授	大内東
副査	教授	高井昌彰
副査	助教授	南弘征

学位論文題名

Model Selection with Empirical Bayes Criteria in L_1 -Regularization

(L_1 正則化法における経験ベイズ規準を用いたモデル選択に関する研究)

昨今の情報技術の発展により、自然科学系、人文社会科学系を問わず、情報に関わる幅広い分野において収集されるデータの形式が多様化し、膨大なデータが蓄積されつつある。なかでも、ゲノム科学におけるマイクロアレイデータなどは、従来までの多変量解析法が想定していたような、標本数 n が変数の次元 p よりも十分大きいという前提を大きく逸脱しており、情報科学や統計科学における新たな課題を与えている。

一般に、変数が増えると、次元の呪いが生じ、データの希薄化、考えられるモデル候補の急増、劣決定性、多重共線性などの問題が生じる。これら多種多様なデータの量は年々増加の一途を辿っており、大量の変数の中から現象に関わる変数のみを特定し選択する変数選択問題は極めて重要な課題の一つとなっている。変数選択問題は、現象に関する観測データに基づき、それを特徴づける真の確率分布の存在を仮定し、複数の統計モデルの中から、一定の基準に則り、もっとも適切なものを選択する“統計的モデル選択問題”の一つと考えることができる。

この変数選択問題に対するアプローチとして、推定すべきパラメータに L_1 ノルムの正則化項を加えた罰則付き最小二乗問題、あるいは罰則付き最尤推定問題を解くことによって、パラメータ推定と変数選択を同時に行う L_1 正則化法が注目されている。特に、正則化パラメータと呼ばれるハイパーパラメータを適切に選択することができれば、モデルの予測精度と解釈可能性を向上させることが可能となる。従って、 L_1 正則化法によるモデル構築の際には、正則化パラメータの選択が本質的な問題となるものの、これを選択するためのモデル評価規準に関しては、十分な研究がなされていない。

以上の背景により、本論文において、著者は、 L_1 正則化法における統計的モデル選択問題を論じ、経験ベイズ法に基づくモデル評価規準の導出、およびそれを用いた L_1 正則化法におけるモデル選択法の理論構築を行っている。

L_1 正則化法におけるモデル評価規準の導出に際しては、評価関数において、 L_1 ノルムの微分不可

能な領域が存在することを考慮する必要がある。この問題に対し、著者は、部分周辺尤度を定義し、その領域に含まれる微分不可能域の測度が0であることを考慮した近似をモデル評価規準として導出している。

本論文の主要な成果は以下の三点に集約される。

1. 変数にグループ構造がない場合の L_1 正則化法によって推定されるモデルを評価するための経験ベイズ規準を導出している。すなわち、グループ構造がない場合の L_1 正則化法を経験ベイズ推定の枠組みで再構築したうえで、その枠組みでの部分周辺尤度の近似を経験ベイズ型モデル評価規準として定義している。これまでに提案された L_1 正則化法である Lasso(Tibshirani, 1996), Adaptive Lasso(Zou, 2005) に基づく一般化線形モデルを用い、数値実験および実データの解析を通じて、既存の簡便法と比較し、その結果、予測精度と変数選択の観点から提案手法が優位に働くことを示している。
2. 変数にグループ構造がある場合の L_1 正則化法に対し、それによって推定されるモデルを評価するための経験ベイズ規準を導出している。この検証では、変数にグループ構造がある場合の L_1 正則化法としてこれまでに提案されている Group Lasso(Yuan and Lin, 2006), Adaptive Group Lasso(Wang and Leng, 2006) に基づく回帰モデルを用い、数値実験および実データの解析を通じて比較を行い、1. と同様に、予測精度と変数選択の観点から提案手法が優位に働くことを示している。
3. 動径基底関数ネットワーク非線形回帰モデルのパラメータ推定問題への L_1 正則化法の適用可能性について理論的な考察を行っている。特に、 L_1 正則化法に基づく動径基底関数ネットワーク非線形回帰モデルの構築の際に本質的である正則化パラメータの選択問題に対し、1. で導出した規準に基づく選択法を提案している。また、数値実験と実データ解析から、従来提案された非線形回帰モデルの推定法よりも提案手法が優れた汎化性能を有することが示されている。

これを要するに、著者は、 L_1 正則化法を経験ベイズ推定の枠組みで再構築し、その部分周辺尤度の近似をモデル評価規準として定式化することにより、統計的モデル選択法に関して新知見を得たものであり、情報科学および統計科学の進歩に寄与するところ大である。よって、著者は、北海道大学博士(情報科学)の学位を授与される資格あるものと認める。