氏　　名　髙田 正彬

学位(専攻分野)　博士(統計科学)

学 位 記 番 号　総研大甲第 2185 号

学位授与の日付　2020 年 9 月 28日

学位授与の要件　複合科学研究科 統計科学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Sparse Regression for Correlated Variables

論 文 審 査 委 員　主　査　教授 二宮 嘉行

教授 藤澤 洋徳

教授 日野 英逸

准教授 川野 秀一
電気通信大学 大学院情報理工学研究科

准教授 今泉 允聡
東京大学 大学院総合文化研究科

（様式３）

博士論文の要旨

氏　　名　高田　正彬

論文題目　Sparse Regression for Correlated Variables

High dimensional data appear in many fields such as biology, economy, and industry. A common approach for high dimensional regression is sparse regularization such as l_1 regularization (Lasso (Tibshirani, 1996)). Since the sparse regularization performs both parameter estimation and variable selection simultaneously, it offers interpretable results by identifying informative variables and effectively avoid overfitting by discarding redundant variables. Its effectiveness has been supported empirically and theoretically by several studies.

However, one of the most significant issues in l_1 regularization is that its performance is quite sensitive to correlations among variables. Typical theoretical supports are based on a kind of "small correlation" assumptions such as the incoherence condition (Fuchs, 2005; Tropp, 2006; Wainwright, 2009) and the restricted eigenvalue condition (Bickel, Ritov, and Tsybakov, 2009; Bühlmann and Van De Geer, 2011; Hastie, Tibshirani, and Wainwright, 2015). Besides, l_1 regularization empirically incurs many false-positive variables in the presence of correlated variables, resulting in large estimation errors.

Correlated variables also hinder the interpretation of models. A standard interpretation of a (generalized) linear model comes from *regression coefficients* and *effects*. A single coefficient represents the degree to which an increase of its variable by one unit increases the prediction "when all other active variables remain fixed." When an output model contains correlated variables, we must consider the influence of all other correlated active variables simultaneously, which can be intractable. A single effect represents the degree to which its variable affects the prediction "on average." However, the value of the effect may become misleading in the presence of strong correlations because it is unlikely to increase a variable by one standard deviation but not changing other correlated variables. Therefore, it is beneficial for interpretability to construct a model with uncorrelated variables.

We address the above issues of sparse regression for correlated variables. We theoretically relax the small correlation assumptions among variables; instead, we impose small correlation assumptions among true active variables. Based on our assumptions, we propose a new regularization method, "Independently Interpretable

Lasso" (IILasso). Our proposed regularization incurs a large penalty for selecting correlated variables; hence the output models have low correlations among active variables. We can intuitively interpret the output models through regression coefficients and effects because each active variable affects the response independently in our model. We provide a coordinate descent algorithm to find a solution of our objective function, which is guaranteed to converge to a stationary point.

Some previous studies have taken correlations among variables into account, and some of them have proposed sparse regression methods that exclude correlated variables. The Uncorrelated Lasso (Chen et al., 2013) intends to construct a model with uncorrelated variables. However, it still tends to select "negatively" correlated variables so that the correlation problem is not resolved. The Exclusive Group Lasso (Kong et al., 2014) is also in this line, but it is necessary to group correlated variables beforehand, and practically it causes unstable results. Additionally, they have no theoretical guarantees for sign recovery and estimation errors. Our method, in contrast, does not select negatively correlated variables, is free from a specific pre-processing such as grouping, and has favorable theoretical guarantees.

In our theoretical analyses, we reveal that the proposed method is advantageous for its sign recovery and estimation error. We define the generalized incoherence condition and the generalized restricted eigenvalue condition for our analyses. Then, we show that our method achieves correct sign recovery under the generalized incoherence condition. This condition is milder than the ordinary incoherence condition for ordinary l_1 regularization for correlated variables. We also show that our method is also beneficial for the estimation errors for correlated variables. This is because the generalized restricted eigenvalue condition for our method is milder than the ordinary restricted eigenvalue condition for ordinary l_1 regularization. Additionally, we show that every local optimal solution achieves the same statistical error rate as the global optimal solution, and thus is almost minimax optimal.

We extend our method to generalized linear models and analyze it theoretically. We show that its estimation errors are almost minimax optimal. As an example of generalized linear models, we provide a coordinate descent algorithm for logistic regression and its estimation error bounds.

Synthetic and real data analyses indicate the effectiveness of our method. Synthetic simulations for linear models and logistic regression models showed that our method achieved accurate prediction and estimation with a few active variables. Real data experiments using ten gene expression datasets also showed that our method could estimate accurate models with small correlations.

Form8・Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　高田　正彬

Title
論文題目
Sparse Regression for Correlated Variables

高次元データに対する回帰手法であるLassoは，高相関特徴量を含むときに問題点がある．そこで，提出された論文では，高相関特徴量が同時にモデルに入りづらくなる罰則を通常のL1罰則に加えることで，その問題点に対処する手法の開発を目的としている．理論解析でも数値解析でも，提案手法の良さを示している．英文で書かれており，全6章と引用文献の計75頁からなる．

第1章は，本論文の序章である．高次元データに対する回帰手法であるLassoの利点と欠点を述べ，高相関特徴量を含むときに問題点があることを，実用上の観点と理論上の観点の両方から議論している．次にその問題点への対処を目的とした過去の手法を説明している．その後に提案手法の概要と利点を述べている．第2章は，本論文に必要な知識として，Lassoと関連手法を詳しく述べている．Lassoに関しては，問題の定式化，その凸性や疎性，幾つかのパラメータ推定アルゴリズム，理論的性質について述べられている．特に，後の議論で重要となる理論的性質として，符号一致性と推定誤差の理論に関しては，丁寧に説明されている．第3章は，Lassoの拡張として，問題点を克服する手法を提案している．まずは，高相関特徴量が同時にモデルに入りづらくなる罰則の設計の仕方が述べられている．2つの特徴量の相関係数の絶対値の単調増加関数と，その2つの特徴量の回帰係数の絶対値の積を用いて，その総和を罰則としているのが，提案手法の大きな特徴である．相関係数の絶対値が大きいと，回帰係数の少なくとも一方は0になりやすくなる点がポイントである．罰則項のデザインによって，相関係数の絶対値が大きいと，Lassoよりもスパース性が起こりやすい形になる．問題点に対処する手法として，過去にはUncorrelated Lasso と Exclusive Group Lasso (EGLasso) が提案されている．しかし，前者はそもそも相関が負の時は妥当に働かず,後者は前処理が必要である.それに対して，提案法はそのような制約や前処理が不要のため，適用しやすい．符号一致性と推定誤差に関しては，提案手法の理論的性質を明らかにしている．そこでは，これまでにLassoで用いられてきたIncoherence Condition と Restricted Eigenvalue Conditionの一般化も行っている．特に，真の回帰係数が非ゼロである特徴量同士の相関が低い時は，Lassoよりも提案手法が通常良くなることが理論解析から見て取れる．得られた手法はRpackageでiilassoとして配布されている．第4章では，第3章で得られた結果を，線形モデルに対して二乗損失を用いる場合から，一般化線形モデルに対して一般の凸損失を用いる場合に拡張し，真のモデルがモデル空間に含まれない場合にまで考察を進めている．推定誤差は，モデル誤差を考慮した形で導出されている．第5章は数値実験である．比較手法とし

ては，通常の Lasso 以外に，高相関特徴量の対処を目的とはしていないが対応できそうな手法である SCAD と MCP，上述した EGLasso と比較した．シミュレーションモデルとしては，真の回帰係数が非ゼロの特徴量同士の相関は 0 で，その他の相関は高い状態のモデルを利用した．提案手法は，予測誤差・推定誤差・モデルサイズの意味で，他の手法よりも良好な結果を示した．特に推定誤差の意味でははっきりと優れていた．また，R package の datamicroarray から 10 個のデータセットを選んで，手法を適用した．マイクロアレイデータの特徴量は高相関が想定されるため，適当な例だと考えられた．提案手法は最も良い性能を示した．第 6 章は結論である．

特徴量に高相関が存在するときに起こる問題点を，実用的な観点からも理論的な観点からも明確にしている．その問題点を克服するために，シンプルかつ含蓄深い手法を提案している．理論解析を丁寧に行い，Lasso の欠点が現れる状況で，提案手法が Lasso よりも良いことを示している．また，数値実験でも，シミュレーション実験で他手法より非常に優れた性能を発揮することが確認できるのみならず，10 セットの実データでも良い性能を示している．問題の動機づけから理論展開・数値実験に至るまで，丁寧な議論が尽くされ，非常に高いレベルにある．第 3 章から第 5 章の内容は査読付国際学術雑誌 Neural Computation に採択されている．以上の理由により，審査委員会は，本論文が学位の授与に値すると判断した．