氏　　名	川島 孝行
学位(専攻分野)	博士(統計科学)
学 位 記 番 号	総研大甲第 2072 号
学位授与の日付	平成 31年 3 月22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	Robust Regression Modeling with Sparsity
論 文 審 査 委 員	主　査　教授　栗木　哲 教授　藤澤　洋徳 教授　二宮　嘉行 准教授　川野　秀一　電気通信大学大学院情報理工学研究科

（様式３）

博士論文の要旨

氏　　名　川島　孝行

論文題目　Robust Regression Modeling with Sparsity

This thesis considers robust regression modeling with sparsity. In this study, we specifically focus on robust regression modeling based on γ-divergence with sparse regularization. The γ-divergence has been investigated for the i.i.d. problem and is renowned for exhibiting strong robustness. This implies that the latent bias can be sufficiently small even under heavy contamination. In this thesis, the γ-divergence is extended to the regression problem. The parameters in regression models are estimated by minimizing the objective function which is the empirical estimation of the γ-divergence with sparse regularization. We propose an efficient parameter estimation algorithm which has a monotone decreasing property for the objective function. In particular, we discuss a linear regression with the L1 regularization in detail. Further, we consider generalized linear models, which are natural extensions of linear regression. However, the parameter estimation algorithm obtained here is not always applicable to generalized linear models. Some models require a higher computational cost as the sample size becomes larger. To reduce this computational cost, we adopt a stochastic optimization approach which can largely reduce the computational cost per iteration. Further, two types of γ-divergence are compared under homogeneous and heterogeneous contaminations. We reveal the distinct difference between two types of γ-divergence in terms of robustness. One γ-divergence can exhibit the strong robustness for any parametric model under heterogeneous contamination. The other cannot in general except under homogeneous contamination or when the parametric model of the response variable belongs to a location-scale family in which the scale does not depend on the explanatory variables. Numerical experiments and real data analyses are performed for illustrating the effectiveness of the proposed methods and for supporting the theoretical properties which we proved.

Outline of the thesis is as follows. Chapter 1 is the introduction of this thesis. In Chapter 2, we briefly describes the robust regression and sparse regression focusing on the contents related to the subsequent discussion. In Chapter 3, we discuss the robust linear regression modeling with sparsity. First, the γ-divergence is extend to the regression problem. The loss function is constructed using the empirical estimation of the γ-divergence. The estimator is defined by the minimizer of the loss function with sparse regularization. To obtain the estimator, an efficient parameter estimation algorithm is proposed via the MM algorithm. In particular, we discuss a linear

regression with the L1 regularization in detail. A tuning parameter selection method is proposed using a robust cross-validation. We additionally illustrate the strong robustness of the proposed method under heavy contamination even when outliers are heterogeneous. Finally, in numerical experiments and real data analyses, we show that our method outperformed existing robust and sparse linear regression methods in terms of predictive performance, variable selection, and computational cost. Chapter 3 is based on the following journal paper:

・Kawashima, T. and Fujisawa, H. Robust and Sparse Regression via γ-Divergence. *Entropy*, Volume 19, No. 608, 2017.

In Chapter 4, we discuss the robust and sparse Generalized Linear Modeling using a stochastic optimization approach. In Chapter 3, we proposed an efficient parameter estimation algorithm using the MM algorithm; however, the proposed one is not always applicable to the GLMs. In the Poisson regression, we need to compute the approximate value of hypergeometric series for all samples per iteration, and a huge computational cost can be required when the sample size is large, e.g., n=10^5. To overcome this problem, a new parameter estimation algorithm is proposed based on the stochastic optimization approach that can significantly reduce the computational cost per iteration and that can be easily applied to GLMs. We can see that the stochastic optimization approach can overcome the difficulty that can be observed when a Poisson regression with L1 regularization is considered. Among stochastic optimization approaches, the randomized stochastic projected gradient descent (RSPG) has been adopted. The RSPG ensures the convergence of our methods. Finally, in numerical experiments and real data analyses, we illustrate that our methods showed better performances than comparative methods in terms of predictive performance and computational cost. Chapter 4 is based on the following preprint paper:

・Kawashima, T. and Fujisawa, H. Robust and Sparse Regression in GLM by Stochastic Optimization. *arXiv*, 2018.

In Chapter 5, we reveal differences between two types of γ-divergence for the regression problem in terms of strong robustness. Fujisawa and Eguchi (2008) investigated the robustness of the γ-divergence for the i.i.d. problem under the contamination model in detail. The contamination model differs between the i.i.d. problem and the regression problem. In the regression problem, the outlier ratio in the contaminated model may depend on the explanatory variable or not. In such situations, they are referred to as the heterogeneous and homogeneous contamination, respectively. In addition to the difference between contamination models, there are two types of γ-divergence for the regression problem in which the treatments of base measure are different. We compare two types of γ-divergence for the regression problem under both homogeneous and heterogeneous contaminations in detail. One γ-divergence can exhibit the strong robustness for any parametric

model under heterogeneous contamination. The other cannot in general except under homogeneous contamination or when the parametric model of the response variable belongs to a location-scale family in which the scale does not depend on the explanatory variables. Finally, numerical experiments are performed for supporting the theoretical properties which we proved. Chapter 5 is based on the following preprint paper:

・Kawashima, T. and Fujisawa, H. On Difference Between Two Types of γ-divergence for Regression. *arXiv*, 2018.

Form8・Separate Sheet
（様式８・別紙）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏　　名　川島　孝行

Title
論文題目　Robust Regression Modeling with Sparsity

出願者は,回帰モデルに対して,外れ値に悪影響を受けないロバスト回帰モデリングと,高次元説明変数に対応できるスパース回帰モデリングを同時に達成する手法の開発を行った．提出論文は英文で書かれており，全 6 章と付録で計 87 頁からなる．

第 1 章は，本論文の序章である．研究の背景として，ロバスト推定・回帰，スパース性，スパース性を持つロバスト線形回帰・一般化線形回帰モデルの過去の研究を概観するとともに，本論文全体の概略が述べられている．第 2 章は，本論文の前提知識として，ロバスト回帰モデリングとスパース回帰モデリングを詳しく説明している．続く第 3〜5 章が本論文の主要部である．第 3 章は，ロバスト性とスパース性を同時に併せもつ線形回帰モデリングを述べている．過去の研究では，外れ値に強いとされる絶対偏差・フーバー型・刈り込み二乗法などにスパース罰則を組み込んだ罰則付きロス関数が提案されてきた．パラメータはそのロス関数の最小化を通して推定される．しかしそのような罰則付きロス関数は最適化の観点で扱いやすい形ではないため，効率的なパラメータ推定アルゴリズムの構築が難しい．本章では，ロス関数としてガンマ・ダイバージェンスに基づいた経験ロス関数を使うことで，この問題を克服している．特に Majorization-Minimization (MM) アルゴリズムを利用して補助関数を作ることで，効率的なパラメータ推定アルゴリズムを構築することに成功している.ロス関数は非凸であるが,罰則項が凸な L1 罰則に限らず,SCAD や MCP のような非凸な場合でも，大域的収束性は保証される．また交差確認法でチューニングパラメータを自動選択することも可能にする．数値実験では，提案法は過去の手法を刈り込み平均に基づく予測二乗誤差の意味ではっきりと上回る.また計算量に関しても,ロバスト性とスパース性を同時に併せもつ手法として有名な sLTS と比較しても，ある状況では計算時間が 1/100 倍以下に短縮される.なお提案手法は R パッケージ “gamreg” として公開されている．第 4 章は，対象とするモデルを線形回帰モデルから一般化線形モデルに拡張することが試みられている．この場合，線形回帰モデルで構築されたアイデアはそのまま拡張できない．たとえば，ポアソン回帰モデルでは，パラメータ推定アルゴリズムの途中で，非常に大きな回数の近似計算が発生する．この問題を克服するために，近年急速に発展している確率的勾配降下法を導入し,パラメータ推定アルゴリズムを構築した.オンラインニュースのデータに対してポアソン回帰モデルを用いて提案手法を適用したところ，通常のスパース手法よりも，刈り込み平均に基づく予測二乗誤差の意味で良いパフォーマンスを示している．第 5 章は，回帰モデルに対する二つのガンマ・ダイバージェンスが提案され，その比較がなされている．外れ値の割合が説明変数に依存する場合には，二つのうちの一方はいわゆるスーパーロバストネスが成立するが，もう一方は特殊な状況でのみ成立することを述べている．たとえば，ロジスティック回帰モデルは，その特殊な

状況に入らない．実際に，ロジスティック回帰モデルに対しては，二つのダイバージェンスに基づいてパラメータ推定を行うと，結果が大きく異なる．加えて，ある種の拡張されたピタゴリアン関係を証明している．その結果を通しても上述のスーパーロバストネスを確認することができる．第 6 章は，本研究に関連した今後の研究課題である．付録は証明の詳細にあてられている．

ロバスト統計とスパース・モデリングは独立に発展を遂げてきた統計的手法である．それらを同時に達成する手法を提案するのは一見簡単に思えるが，安定的で使いやすいパラメータ推定アルゴリズムの構築が困難であるという点が長らく問題であった．本論文ではその問題を克服する手法を提案し，数値的な良さを示している．加えて，回帰モデル特有の外れ値の割合が説明変数に依存した場合を調べ，スーパーロバストネスに関する新しい結果を提示している．第 3 章は査読付き国際学術雑誌 Entropy に採択されている．第 4 章と第 5 章の内容はそれぞれ独立した論文として投稿中である．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．