Robust Regression Modeling with Sparsity

川島　孝行

博士（統計科学）

総合研究大学院大学
複合科学研究科
統計科学専攻

平成３０（２０１８）年度

Robust Regression Modeling with Sparsity

Takayuki Kawashima

Department of Statistical Science
School of Multidisciplinary Sciences
The Graduate University for Advanced Studies

This dissertation is submitted for the degree of
Doctor of Philosophy

January 2019

Acknowledgements

I would like to express my great gratitude to my supervisor, Hironori Fujisawa. He has always encourage my Ph.D. life. My research life started when I first met him in the Open House at the ISM. Thanks to him, I have had enjoyable researches and discussions. Further, I am really grateful to the members of the dissertation committee, Satoshi Kuriki, Yoshiyuki Ninomiya, and Shuichi Kawano for giving valuable comments and judging my Ph.D. degree. I would like to express my special thanks to my colleagues in the ISM and NIPR. Finally, I would like to thank my family for their constant support.

Abstract

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 L_1 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.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Regression analysis is a fundamental tool in statistical data analysis. The ordinary least squares (OLS) is the most popular regression method. In modern regression analysis, we often treat high-dimensional data, so that the number of explanatory variables may become larger than the sample size. Modern data are often referred to as high-dimensional data. Typical examples of such high-dimensional data are the microarray data, social data, and biological data. Sparse regression methods with a sparsity-inducing regularization are extensively used in high-dimensional data [103, 89, 101, 104, 28, 100]. One renowned example of the sparse regression method is the least absolute shrinkage and selection operator (LASSO) [87], which employs the OLS with the L_1 regularization. However, because the LASSO is based on the OLS, it is sensitive to outliers in the explanatory variables or the response variable. To handle high-dimensional noisy data, we consider a regression method that exhibits both sparsity and robustness to outliers. In this chapter, we give some historical reviews on robust and sparse regression modelings and describe how we need a new method with both sparsity and robustness.

1.1 Robust Estimation for i.i.d. problem

M-estimation: M-estimation: The sample mean is a typical estimator of mean parameter. However, the sample mean can be adversely influenced by outliers. Detecting and deleting outliers from the data is a simple way to robustify the sample mean, but it is not always easy due to masking effect [78]. Hence, the median and trimmed mean were proposed as robust versions of the sample mean. The M-estimation [50], which is a generalization of the maximum likelihood estimation, is a more sophisticated way to obtain an robust estimator. The sample mean and median are examples of the M-estimator. The M-estimation is based on the approach that minimizes a robust loss function. The Huber's loss [50] and Tukey's bisquare loss are typical examples of the robust loss function. To compute the M-estimate, standard parameter estimation methods, such as the Newton-Raphson method, can be applied. However, methods based on derivatives may be unstable because of the assumption to attain robustness in the M-estimation. Hence, the iteratively reweighted least squares (IRLS) method has been preferably used. The theoretical properties of the M-estimator, including consistency and asymptotic normality, have been investigated (see, Sect. 10 of [66]). By virtue of these results, robust confidence intervals and statistical hypothesis testing can be derived. For the estimation on the scale of the data, the M-estimation has been studied in a similar manner. Readers may refer to Chapter 2 of [66] for details.

Measure of Robustness: The sensitivity curve (SC) is defined by the difference between the estimate for the sample and the estimate for the sample and a single outlier, so that the SC illustrates the adverse effect of a single outlier. The bounded SC means that the adverse effect of a single outlier is bounded. Aforementioned location M-estimators, e.g., based on Huber's loss and Tukey's bisquare loss, have the bound SC. The influence function (IF) [43] may be considered as an asymptotic version of the SC. The IF can approximate the bias caused by outliers under the assumption that a fraction of outliers is small. The bias is expected to be small when the IF is small. The breakdown point (BP) [66] is the proportion of outliers that can ensure the estimate to be finite even if outliers goes to infinity. For other measures of robustness, maximum asymptotic bias and gross-error sensitivity were proposed.

Outlier Detection: There are many methods for outlier detection. The most simple method is based on the 3σ distance [84]. Another simple method is the box plot rule. This uses the difference between the third quartile and first quartile, and this quantity is called the inter quartile range. The box plot rule closely related to the 3σ distance method when the data is generated from the Gaussian distribution. More enhanced methods are based on statistical hypothesis testing. The Grubb's test [42] (the maximum normed residual test) was proposed for the case of univariate data. Some variants of the Grubb's test for multivariate data were also proposed [63, 1]. The student's t-test was also applied to outlier detection. The multivariate version of the student's t-test is called the Hotelling t-squared test and has been used in the field of the quality control for a long time [49]. Further, χ-squared test was applied to outlier detection for operating system call data [97]. Nonparametric approaches such as histogram were also discussed (system call intrusion detection: [27], fraud detection [29], and network intrusion detection [47]).

1.2 Robust Regression Modeling

Case of Linear Regression: Robust linear regression methods have been studied in robust statistics for a long time. A classical idea of the robust linear regression was based on the M-estimation. This class of robust linear regression methods includes many common methods. The least absolute deviation (LAD) [9] adopted sum of the absolute values of the residuals instead of squared residuals in the OLS. The regression methods, using Huber's loss and with Tukey's bisquare loss, belong to the M-estimation. Further, the IRLS method can be used for a parameter estimation in a similar way to in the i.i.d. problem. Their robustness has been investigated by the BP. However, the M-estimation is not robust against outliers in the explanatory variables, which is referred to as leverage point, although it is robust against outliers in the response variable (see, Sect. 4.6 of [66]). To enhance the robustness to leverage point by the BP, the least median of squares (LMS) [75], least trimmed squares (LTS) [75], and S-estimation [80] were proposed. The LMS and LTS adopted respectively the median of squared residuals and trimmed sum of squared residuals instead of squared residuals in the OLS. The S-estimation is defined by minimizing the variance of residuals. However, their estimators resulted in a low asymptotic efficiency (LMS: [76], LTS: [85], S-estimation: [66]), and there were no simple parameter estimation algorithms such as the IRLS. To obtain a high asymptotic efficiency, the robust and efficient weighted least squares estimator (REWLSE) was proposed by Gervini and Yohai [36]. The REWLSE is a type of weighted least squares estimator whose weights are adaptively calculated from an initial robust estimator. Moreover, the REWLSE attains asymptotically efficient.

Case of Generalized Linear Models: Robust generalized linear models (GLMs) [67], which include the linear regression, logistic regression and Poisson regression models, have been also studied in robust statistics. Künsch, Stefanski, and Carroll [61] considered the M-estimation of GLMs and proposed conditionally unbiased bounded influence estimation. Carroll and Pederson [18] focused on the logistic regression and used downweighting scheme based on the Mahalanobis's distance in order to give a small weight to terms related to outliers. This estimation is regarded as a special case of the weighted maximum likelihood estimation. In other robust logistic regression methods, Pregibon [73] proposed the M-estimation based on the deviance of the logistic regression, and Bianco and Yohai [7] improved this estimator to attain the Fisher consistent. Croux, Flandre, and Haesbroeck [20] showed that the leverage point cause a different behaviour of the estimator in the logistic regression unlike the case of the linear regression. Generally, the non-robust estimator tends to infinity as the influence of the leverage point become large. On the other hand, in the logistic regression, the non-robust estimator tends to zero in such a situation. This behaviour is called implosion breakdown by Chi and Scott [19].

1.3 Sparse Regression Modeling

The LASSO is the most popular sparse linear regression method over the last two decades. The LASSO minimizes the squared loss with the L_1 regularization, so that some regression estimates are zero. Hence, the LASSO can perform estimation and variable selection simultaneously. Variable selection in the regression problem is not a new concept. Various variable selection methods have been proposed. Typical examples are forward/backward stepwise method and the best-subset selection method. The former tends to be a local optimizer. The latter is not computationally feasible in high-dimensional data. The LASSO can overcome these problems.

Statistical Property of LASSO: Greenshtein and Ritov [41] studied the predictive performance of the LASSO. They proved that the expected squared prediction error of the LASSO approximates the Bayes error under mild regularity conditions. Further, results under more refined conditions have been developed, e.g., the compatibility condition [90], the restricted eigenvalue condition [8], the coherence condition [16], and restricted isometry condition [17]. Zhao and Yu [102] studied a sufficient and necessary condition to recover the true sparsity pattern. Their study was based on the irrepresentable condition [68]. For the consistent parameter estimation in the L_2 sense, Meinshausen and Yu [69] investigated the behavior of the LASSO estimator when only a relaxed version of the irrepresentable condition is met.

Computation of LASSO: The LASSO is a convex optimization problem, so that the solutions, which satisfy the stationary condition, are global minimum points. On the other hand, aforementioned variable selection methods are non-convex optimization problems. It is difficult to obtain a global minimum point. In such a situation, the least angle regression (LARS) algorithm [26] was proposed using the stationary condition of the optimization problem on the LASSO. The LARS can efficiently compute the solution path which is the entire solution set of optimization problems for each value of the tuning parameter for the sparsity. This algorithm has the same computational cost as the OLS. Some variants of the LARS were also proposed [91, 58, 31]. Pathwise coordinate descent optimization algorithms [32, 95] can be more efficient in high-dimensional data. Unlike the LARS, they are easily applied to other regression methods, e.g., the logistic regression and Poisson regression methods [33]. Further, traditional convex optimization methods, such as the proximal gradient method and alternating direction method of multipliers, can be applied to the LASSO.

Beyond LASSO: Modifying sparse regularization has been investigated to improve the LASSO estimator's statistical properties. Fan and Li [28] and Zhang [100] introduced non-convex sparse regularizations, called smoothly clipped absolute deviation (SCAD) and minimax concave penalty (MCP), respectively, instead of the L_1 regularization. Fan and Li [28] defined the oracle property and showed that the LASSO estimator does not satisfy this desirable property. Zou [103] proposed the adaptive LASSO which adaptively controls the strength of the regularization and showed that the adaptive LASSO estimator satisfies the oracle property. Moreover, the adaptive LASSO can be regarded as convex approximation of the L_q $(0 < q < 1)$ regularization, which was proved to satisfy the oracle property [60]. For some specific sparse structures, the group LASSO [98] and the fused LASSO [89] were proposed. The LASSO was applied to the Cox's proportional hazards models for modeling survival data [88].

1.4 Robust Linear Regression Modeling with Sparsity

Classical Approach: A robust linear regression modeling with sparsity can be obtained by combining the classical robust linear regression with the sparse regularization [92, 58, 2]. Wang, Li, and Jiang [92] proposed the least absolute deviation LASSO (LAD-LASSO), which is constructed using the LAD with the L_1 regularization. They ensured the asymptotic consistency of the LAD-LASSO estimator. Alfons, Croux, and Gelper [2] proposed the sparse least trimmed squares (sLTS), which is constructed using the LTS with the L_1 regularization. They investigated the robustness of the sLTS estimator by virtue of its BP. The robust least angle regression (RLARS) [58] is a robust version the LARS and can be constructed by replacing the sample correlation with a robust estimate of the correlation in its parameter estimation algorithm. However, the aforementioned methods are limited to a linear regression and are not applicable to other regression methods such as GLMs. Furthermore, most of robust linear regression methods are based on a non-convex loss function for achieving robustness [79], and the L_1 regularization, which is a non-differentiable function. Hence, we need to solve a non-convex and non-differentiable optimization problem to obtain the estimator. Generally, a high computational cost can be required when such a problem is solved by standard optimization methods.

Robust Divergence Approach: Robust parameter estimation using density power weight has been intensively investigated, and the corresponding divergences have been discussed [82, 54, 5, 4, 93, 53, 35]. Lozano, Meinshausen, and Yang [64] focused on the L_2 divergence [82] and incorporated the L_1 regularization. They investigated the consistency of the estimator and the robustness based on the BP. Moreover, the proposed parameter estimation algorithm ensures convergence. Zang, Zhao, Zhang, Li, Zhang, and Ma [99] adopted the density power divergence [4] with the L_1 regularization. They proposed the parameter estimation algorithm based on the coordinate descent algorithm. As a natural extension of [99], Ghosh and Majumdar [40] considered a non-convex sparse regularization case. They studied the robustness based on the influence function that is not a classical one [44] but a fully rigorous one [3]. In contrast to classical approaches, robust divergence approaches can be easily extended to other regression models, such as GLMs, by selecting an appropriate probability density function. Especially, the γ-divergence [35] is known for having the strong robustness, which implies that the latent bias can be sufficiently small even under heavy contamination. Further, other robust divergences, including density power divergence [4], cannot achieve such robustness, and the latent bias is caused by a high outlier ratio. However, even in this approach, we need to solve a non-convex and non-differentiable optimization problem. In this thesis, we deal with such a problem.

1.5 Robust and Sparse Generalized Linear Modeling

GLMs include many important regression models such as the linear regression, logistic regression, and Poisson regression models. Recently, robust and sparse generalized linear modeling have been proposed. Chi and Scott [19] adopted a robust divergence approach and proposed robust logistic regression based on the L_2 divergence with the L_1 regularization. Moreover, they proposed an efficient parameter estimation algorithm using the majorization-minimization algorithm (MM algorithm) [52] and investigated the convergence property in a similar way to in [81]. Bootkrajang and Kabán [10] proposed the robust and sparse logistic regression modeling with mislabel probabilities on outliers. On the other hand, Hung, Jou, and Huang [51] proposed the robust logistic regression modeling based on the γ-divergence, and it does not need to model mislabel probabilities. However, they did not discuss sparse regularization methods, and the proposed method cannot be directly extended to sparse modeling. Tibshirani and Manning [86] applied a mean-shift model [83] to a sparse logistic regression method. In the linear regression, a mean-shift model gave a new characterization of the M-estimation in the form of the sparse regularization. Therefore, the sparse logistic regression with mean-shift might be expected to be robust against outliers as with the M-estimation of the linear regression. As stated above, robust and sparse generalized linear modeling has been mainly considered in the case of logistic regression. To the best of our knowledge, other robust and sparse GLMs, e.g., the Poisson regression, has not been discussed yet. In this thesis, robust and sparse GLMs including the Poisson regression are discussed.

1.6 Outline of the 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 [52]. In particular, we discuss a linear regression with the L_1 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 [57]:

- 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 L_1 regularization is considered. Among stochastic optimization approaches, the randomized stochastic projected gradient descent (RSPG) [38] 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 [56]:

- 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 [35] 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 [35, 57]. 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 [55]:

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

Chapter 2

Robust Regression and Sparse Regression

We briefly describes the robust regression and sparse regression focusing on the contents related to the subsequent discussion.

2.1 Robust Regression

The OLS method is defined by

$$\min_{\beta_0,\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2, \tag{2.1}$$

where $y \in \mathbb{R}$ is the response variable, $x \in \mathbb{R}^p$ is the explanatory vector, $\beta_0 \in \mathbb{R}$ is the intercept and $\beta \in \mathbb{R}^p$ is the regression coefficient vector. The OLS estimator is not robust against outliers in the explanatory variables or the response variable. Robust regression methods have been studied for a long time in the filed of robust statistics to alleviate an effect of outliers. Typical methods are the LAD, LMS, LTS, and regression methods with Huber's loss and with Tukey's bisquare. Recently, the robust parameter estimation using the density power weight has been intensively investigated, and the corresponding divergence has been discussed [54, 5, 4, 93, 53, 35]. The density power weight gives a small weight to the terms related to outliers; further, the estimator becomes robust against outliers. Additionally, robust regression methods that are based on such divergences have been proposed [19, 51, 39, 64, 57, 56].

In this section, we briefly discuss some classical representable robust regression methods along with the robust regression method that is based on the γ-divergence. In the subsequent sections, we consider a situation where the response variable contains outliers and refer to the regression based on the γ-divergence as γ-regression.

2.1.1 Least Absolute Deviation

Definition: The LAD adopts L_1 loss instead of the L_2 loss in (2.1) as follows:

$$\min_{\beta_0,\beta} \frac{1}{n} \sum_{i=1}^{n} |y_i - \beta_0 - x_i^T \beta|. \tag{2.2}$$

Robustness: Here, we intuitively show the robustness of the LAD. Suppose that y_1 is an outlier. The residual $|y_1 - \beta_0 - x_1^T \beta|$ will be large, but smaller than the squared residual $(y_1 - \beta_0 - x_1^T \beta)^2$ of the OLS, so that an adverse effect of the outlier y_1 can be alleviated in the LAD. Therefore, we expect that the LAD will be more robust against outliers than the OLS.

Computation: Due to the non-differentiability of the L_1 loss, we cannot use standard optimization methods such as the gradient descent, Newton-Raphson method and quasi-Newton method, to obtain the minimizer of (2.2). Using slack variables, the optimization problem (2.2) can be reformulated into the following linear programming problem:

$$\begin{aligned} &\min_{\beta_0,\beta,u_1,\ldots,u_n} \frac{1}{n} \sum_{i=1}^{n} u_i \\ &\text{subject to } -u_i \leq y_i - \beta_0 - x_i^T \beta \leq u_i \text{ for } i = 1,\ldots,n. \end{aligned}$$

The simplex method and its alternatives can effectively solve this linear programming problem. Some software packages are available to compute the LAD in the R language; e.g., “L1pack” and “MASS”.

2.1.2 Regression with Huber’s Loss

Definition: This regression method adopts Huber’s loss instead of the L_2 loss in (2.1) as follows:

$$\arg\min_{\beta_0,\beta} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \beta_0 - x_i^T \beta), \tag{2.3}$$

where

$$\rho(z) = \begin{cases} z^2 & \text{if } |z| \leq c \\ 2c|z| - c^2 & \text{otherwise,} \end{cases}$$

and c is a tuning parameter. In practice, $c = 1.345$ is used in terms of the asymptotic statistical efficiency.

Robustness: In a similar way to in the LAD, we can show that the regression method with Huber's loss is robust. Suppose that y_1 is an outlier. The residual $|y_1 - \beta_0 - x_1^T\beta|$ will be large, but $2c|y_1 - \beta_0 - x_1^T\beta| - c^2$ will be smaller than the squared residual $(y_1 - \beta_0 - x_1^T\beta)^2$ of the OLS, so that an adverse effect of the outlier y_1 can be alleviated in the regression method with Huber's loss. Therefore, we expect the regression method with Huber's loss to be more robust against outliers than the OLS.

Computation: The Huber's loss is differentiable unlike the L_1 loss in the LAD. Therefore, we can apply standard optimization methods to (2.3). Here, we exhibit the following IRLS method which has been used in robust statistics. For simplicity, we assume that $\beta_0 = 0$.

$$\beta^{(t+1)} = \left[X^T W^{(t)} X\right]^{-1} X^T W^{(t)} Y,$$

where $X = (x_1^T, \ldots, x_n^T) \in \mathbb{R}^{n\times p}$, $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, $\beta^{(t)}$ is the regression coefficient vector at the t-th iterative step, and

$$W^{(t)} = \mathrm{diag}\left\{\rho\left(y_1 - x_1^T\beta^{(t)}\right)/(y_1 - x_1^T\beta^{(t)}), \ldots, \rho\left(y_n - x_n^T\beta^{(t)}\right)/(y_n - x_n^T\beta^{(t)})\right\}.$$

The IRLS for the regression method with Huber's loss is closely related to the MM algorithm [52] which has been recently studied intensively in the machine learning community [71]. Interested readers may refer to Chapter 8 of [34] for details. Some software packages are available to compute the regression method with Huber's loss in the R language; e.g., "robustreg" and "MASS".

2.1.3 Least Trimmed Squares

Definition: The LTS adopts the trimmed loss instead of the L_2 loss in (2.1) as follows:

$$\min_{\beta_0,\beta} \frac{1}{m}\sum_{i=1}^{m} e_{[i]}, \tag{2.4}$$

where $e_i = (y_i - \beta_0 - x_i^T\beta)^2$, $e_{[1]} \le \cdots \le e_{[n]}$ are the order statistics of $e_1, \cdots, e_n$ and $m \le n$.

Robustness: The trimming constant m is generally selected to satisfy $\frac{n}{2} \le m \le n$. In practice, the following value of the trimming constant $\frac{\lfloor (n+p+1) \rfloor}{2}$ is used. The value of the trimming constant determines the number of residuals required for the estimation. Large residuals, which may contain outliers, will be excluded from the estimation. Therefore, the LTS is expected to be robust against outliers.

Computation: Compared with the OLS and LAD, the optimization problem (2.4) is a non-convex optimization problem. It is difficult to obtain a global minimizer for non-convex optimization problems. Thus, a main task for non-convex optimization problems is to obtain a stationary point where the derivative of the objective function is equal to zero. To obtain a stationary point of (2.2), Rousseeuw and Driessen [77] proposed the following iterative parameter estimation algorithm referred to as the FAST-LTS algorithm.

For $t = 0, 1, 2, \ldots$:

Step 1. given the t-th iterate $(\beta_0^{(t)}, \beta^{(t)})$, compute the order statistics of $e_{[1]}^{(t)} \leq \cdots \leq e_{[n]}^{(t)}$ based on the t-th iterate $(\beta_0^{(t)}, \beta^{(t)})$;

Step 2. construct the trimmed loss with the set of indices corresponding to the m smallest residuals, $e_{[1]}^{(t)} \leq \cdots \leq e_{[m]}^{(t)}$, and compute next iterate $(\beta_0^{(t+1)}, \beta^{(t+1)})$ based on the current trimmed loss;

Step 3. If a convergence criterion, e.g., $\|\beta^{(t)} - \beta^{(t+1)}\| < 10^{-4}$, is satisfied, the algorithm is stopped; otherwise, it returns to Step 1.

The FAST-LTS algorithm has the following monotone decreasing property:

$$\frac{1}{m}\sum_{i=1}^{m} e_{[i]}^{(0)} \geq \frac{1}{m}\sum_{i=1}^{m} e_{[i]}^{(1)} \geq \frac{1}{m}\sum_{i=1}^{m} e_{[i]}^{(2)} \geq \cdots.$$

Some software packages are available to compute the LTS in the R language; e.g., “robustbase”.

2.1.4 γ-Regression

Definition of γ-divergence for regression: The γ-divergence was defined for two probability density functions, and its properties were investigated by Fujisawa and Eguchi [35]. First, we review the γ-divergence for the i.i.d. problem. Let $g(u)$ and $f(u)$ be two probability density functions. The γ-cross entropy and γ-divergence were defined by

$$\begin{aligned} d_\gamma(g(u), f(u)) &= -\frac{1}{\gamma}\log\int g(u)f(u)^\gamma du + \frac{1}{1+\gamma}\log\int f(u)^{1+\gamma} du, \\ D_\gamma(g(u), f(u)) &= -d_\gamma(g(u), g(u)) + d_\gamma(g(u), f(u)), \end{aligned}$$

respectively, where γ is the positive tuning parameter that controls the trade-off between efficiency and robustness. This satisfies the following two basic properties of divergence:

(i) $D_\gamma(g(u), f(u)) \geq 0$.

(ii) $D_\gamma(g(u), f(u)) = 0 \Leftrightarrow g(u) = f(u)$ $(a.e.)$.

Let us consider the γ-divergence for regression, which is defined for two conditional probability density functions. Suppose that $g(x,y)$, $g(y|x)$, and $g(x)$ are the underlying probability density functions of (x,y), y given x and x, respectively. Let $f(y|x)$ be another parametric conditional probability density function of y given x. For the regression problem, Fujisawa and Eguchi [35] proposed the following cross entropy and divergence:

$$\begin{aligned}
&d_{\gamma,1}(g(y|x), f(y|x); g(x)) \\
&= -\frac{1}{\gamma}\log\int \exp\{-\gamma d_\gamma(g(y|x), f(y|x))\} g(x)dx \\
&= -\frac{1}{\gamma}\log\int \left\{\int g(y|x) f(y|x)^\gamma dy \Big/ \left(\int f(y|x)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}\right\} g(x)dx \\
&= -\frac{1}{\gamma}\log\int\int \left\{f(y|x)^\gamma \Big/ \left(\int f(y|x)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}\right\} g(x,y)dxdy. \qquad (2.5)
\end{aligned}$$

$$D_{\gamma,1}(g(y|x), f(y|x); g(x)) = -d_{\gamma,1}(g(y|x), g(y|x); g(x)) + d_{\gamma,1}(g(y|x), f(y|x); g(x)). \qquad (2.6)$$

The cross entropy is empirically estimable, as will be seen later, and the parameter estimation is easily defined. Further, we propose the following cross entropy and divergence, respectively:

$$\begin{aligned}
&d_{\gamma,2}(g(y|x), f(y|x); g(x)) \\
&= -\frac{1}{\gamma}\log\int\left(\int g(y|x) f(y|x)^\gamma dy\right) g(x)dx + \frac{1}{1+\gamma}\log\int\left(\int f(y|x)^{1+\gamma}dy\right) g(x)dx \\
&= -\frac{1}{\gamma}\log\int\int f(y|x)^\gamma g(x,y)dxdy + \frac{1}{1+\gamma}\log\int\left(\int f(y|x)^{1+\gamma}dy\right) g(x)dx. \qquad (2.7)
\end{aligned}$$

$$D_{\gamma,2}(g(y|x), f(y|x); g(x)) = -d_{\gamma,2}(g(y|x), g(y|x); g(x)) + d_{\gamma,2}(g(y|x), f(y|x); g(x)). \qquad (2.8)$$

The base measures on the explanatory variable are taken twice on each term of the γ-divergence for the i.i.d. problem. This extension from the i.i.d. problem to the regression problem seems to be more natural than (2.5). The cross entropy is also empirically estimable. We refer to these two types, (2.5) and (2.7), as type I and type II, respectively. Both types of γ-divergence satisfy the following two basic properties of divergence and relation to the KL-divergence for $j = 1, 2$:

Theorem 2.1.1.

$$
\begin{aligned}
&\text{(i)} \quad D_{\gamma,j}(g(y|x), f(y|x); g(x)) \geq 0,\\
&\text{(ii)} \quad D_{\gamma,j}(g(y|x), f(y|x); g(x)) = 0 \Leftrightarrow \quad g(y|x) = f(y|x) \quad \text{(a.e.)},\\
&\text{(iii)} \quad \lim_{\gamma \to 0} D_{\gamma,j}(g(y|x), f(y|x); g(x)) = \int D_{KL}(g(y|x), f(y|x)) g(x) dx,
\end{aligned}
$$

where $D_{KL}(g(y|x), f(y|x)) = \int g(y|x) \log g(y|x) dy - \int g(y|x) \log f(y|x) dy.$

The proof is in Appendix A. We consider the robust and sparse regression based on type II of γ-divergence in Chapter 3 and type I of γ-divergence in Chapter 4. Further, we discuss the difference and theoretical robust properties on both types of γ-divergence in Chapter 5.

Estimation of γ-Regression: Let $f(y|x;\theta)$ be the conditional probability density function of y given x with the parameter θ. The target parameter can be considered by:

$$
\begin{aligned}
\theta^*_{\gamma,j} &= \arg\min_{\theta} D_{\gamma,j}(g(y|x), f(y|x;\theta); g(x))\\
&= \arg\min_{\theta} d_{\gamma,j}(g(y|x), f(y|x;\theta); g(x)) \quad \text{for} j = 1, 2. \qquad (2.9)
\end{aligned}
$$

When $g(y|x) = f(y|x;\theta^*)$, we have $\theta^*_{\gamma,j} = \theta^*$.

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be the observations randomly drawn from the underlying distribution $g(x, y)$. Using the formulas (2.5) and (2.7), both types of γ-cross entropy for regression, $d_{\gamma,j}(g(y|x), f(y|x;\theta); g(x))$, can be empirically estimated by:

$$
\begin{aligned}
\bar{d}_{\gamma,1}(f(y|x;\theta)) &= -\frac{1}{\gamma} \log \frac{1}{n} \sum_{i=1}^{n} \frac{f(y_i|x_i;\theta)^{\gamma}}{\left(\int f(y|x_i;\theta)^{1+\gamma} dy\right)^{\frac{\gamma}{1+\gamma}}},\\
\bar{d}_{\gamma,2}(f(y|x;\theta)) &= -\frac{1}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^{n} f(y_i|x_i;\theta)^{\gamma} \right\} + \frac{1}{1+\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^{n} \int f(y|x_i;\theta)^{1+\gamma} dy \right\}.
\end{aligned}
$$

By virtue of (2.9), we define the γ-estimator by:

$$
\hat{\theta}_{\gamma,j} = \arg\min_{\theta} \bar{d}_{\gamma,j}(f(y|x;\theta)) \quad \text{for} j = 1, 2. \qquad (2.10)
$$

In a similar way to in Fujisawa and Eguchi [35], we can show the consistency of $\hat{\theta}_{\gamma,j}$ to $\theta^*_{\gamma,j}$ under some conditions.

Robustness: Here, we briefly show why type II of γ-estimator $\hat{\theta}_{\gamma,2}$ is robust. Suppose that y_1 is an outlier. The conditional probability density $f(y_1|x_1;\theta)$ is expected to be sufficiently small. We see from $f(y_1|x_1;\theta) \approx 0$ and (2.10) that:

$$\begin{aligned}
&\arg\min_{\theta} \bar{d}_{\gamma,2}(f(y|x;\theta)) \\
&= \arg\min_{\theta} -\frac{1}{\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^{n} f(y_i|x_i;\theta)^{\gamma}\right\} + \frac{1}{1+\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^{n}\int f(y|x_i;\theta)^{1+\gamma}dy\right\} \\
&\approx \arg\min_{\theta} -\frac{1}{\gamma}\log\left\{\frac{1}{n-1}\sum_{i=2}^{n} f(y_i|x_i;\theta)^{\gamma}\right\} + \frac{1}{1+\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^{n}\int f(y|x_i;\theta)^{1+\gamma}dy\right\}.
\end{aligned}$$

Therefore, the term $f(y_1|x_1;\theta)$ is naturally ignored in (2.10). We can show the robustness of type I of γ-estimator $\hat{\theta}_{\gamma,1}$ in a similar manner.

Computation: We can apply standard optimization methods in a similar way to in the regression method with Huber's loss. However, the objective function of the γ-regression is non-convex. Therefore, another optimization method is required to achieve numerical stability and efficiency. In Chapters 3 and 4, we discuss this problem in detail.

2.2 Sparse Regression

Let us rewrite (2.1) as:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}||Y - X\beta||^2.$$

For simplicity, we assume that $\beta_0 = 0$. Then, $\hat{\beta}$ is represented as $(X^TX)^{-1}X^TY$. However, X^TX is not invertible when $n < p$, and $\hat{\beta}$ does not uniquely exist, i.e., $\hat{\beta}$ is an infinite set. The following two problems exist for the OLS in high-dimensional data.

- Overfitting: $\hat{\beta}$ makes the training loss $\frac{1}{n}||Y - X\hat{\beta}||^2$ to be zero. This is an overfitting to the training data, i.e., poor predictive performance.

- Interpretation: Typically, the elements of $\hat{\beta}$ are nonzero. Hence, we cannot determine a subset that shows strong effects among a large number of explanatory variables.

To prevent overfitting, regularization methods have been extensively incorporated into the OLS. Hoerl and Kennard [48] proposed the following regularized regression method and its estimator:

$$\text{Ridge regression: } \hat{\beta}_{ridge} = \arg\min_{\beta} \frac{1}{n}||Y - X\beta||^2 + \lambda||\beta||^2,$$

where $\lambda \geq 0$ is the tuning parameter for the regularization term $||\beta||^2$. When $\lambda = 0$, the ridge regression is equal to the OLS. Then, $\hat{\beta}_{ridge}$ is represented as $(X^TX + \lambda I_p)^{-1}X^TY$, and $X^TX + \lambda I_p$ for $\lambda > 0$ is invertible even when $n < p$. However, the ridge regression does not overcome the problem of interpretation because the elements of $\hat{\beta}_{ridge}$ are nonzero. Tibshirani [87] proposed another regularized regression method referred to as the LASSO. The difference between the ridge regression and LASSO is the regularization term. The LASSO adopts the L_1 regularization $||\beta||_1 (= |\beta_1| + \cdots + |\beta_p|)$ instead of $||\beta||^2$. By virtue of the non-differentiability at the origin of $||\beta||_1$, the LASSO yields the sparse estimator whose elements are zero. The LASSO is applied to many research fields such as signal processing, bioinformatics, machine learning, and image processing, by virtue of its effectiveness. In this section, we briefly describe the LASSO.

Definition: The LASSO and its estimator are defined by

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \frac{1}{n}||Y - X\beta||^2 + \lambda||\beta||_1, \tag{2.11}$$

where $\lambda \geq 0$ is the tuning parameter for the regularization term $|\beta|$. When $\lambda = 0$, the LASSO is equal to the OLS. Let us consider the case where $p = n$ and $X = \mathrm{I}_n$ to see the sparsity of the LASSO estimator. We can explicitly obtain $\hat{\beta}_{lasso}$ as follows:

$$\hat{\beta}_{lasso,j} = \begin{cases} \mathrm{sign}(y_i)(y_i - n\lambda/2) & (|y_i| > n\lambda/2) \\ 0 & (|y_i| \leq n\lambda/2), \end{cases}$$

where $\hat{\beta}_{lasso,j}$ is the j-th element of $\hat{\beta}_{lasso}$. The LASSO estimator $\hat{\beta}_{lasso,j}$ has the threshold function, $\mathrm{sgn}(y_i)\max(|y_i| - n\lambda/2, 0)$, that is referred to as the soft-thresholding function. By virtue of this threshold scheme, some elements become zero. Even in a general case, the soft-thresholding function appears in optimization methods as will be shown later.

Theoretical Property: Here, we prepare some notations and a condition to show a theoretical property of the LASSO when $n \ll p$. We consider the following true regression model:

$$Y = X\beta^* + \varepsilon,$$

where $\beta^* \in \mathbb{R}^p$ is the true regression coefficient vector and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^n$ is an i.i.d. noise vector. Let us denote the index set as $S(\beta) := \{j | \beta_j \neq 0\}$. For a set S, $|S|$ denotes the cardinality of S; further, S^c denotes the complement set of S. We consider the following assumptions:

Assumption 1. $|S(\beta^*)| = d \ (\ll n)$.

Assumption 2. $\max_{i,j} |X_{i,j}| \leq 1$, where $X_{i,j}$ is the element of the i-th row and j-th column of X.

Assumption 3. $(\varepsilon_i)_{i=1}^n$ is an i.i.d. sub-Gaussian sequence: $E\left[e^{t\varepsilon_i}\right] \leq e^{\frac{\sigma^2 t^2}{2}}$ $(\forall t \in \mathbb{R})$ for $\sigma > 0$.

We define the following condition for X

Definition of Restricted Eigenvalue Condition.

$$\phi_{RE}(a,b) := \inf_{I \subseteq \{1,\dots,n\}, v \in \mathbb{R}^p : |I| \leq a, b\|v_I\|_1 \geq \|v_{I^c}\|_1} \frac{vX^TXv}{n\|v_I\|^2} \text{ and } \phi_{RE}(a,b) > 0.$$

Then, we state the following theorem for ensuring the parameter consistency.

Theorem 2.2.1. *For $0 < \delta < 1$, λ is set to $4\sigma\sqrt{\frac{2\log(2p/\delta)}{n}}$. Then, we can obtain the following bound under the restricted eigenvalue condition, $\phi_{RE}(2d,3) > 0$:*

$$\|\beta_{lasso} - \beta^*\|^2 \leq C\frac{d\log(p/\delta)}{n} \textit{ with probability } 1-\delta,$$

where C is a positive constant.

Proof. See Chapter 11 of [46] or Chapter 6 of [15]. □

The dimension of the regression coefficient vector affects the rate by $\log p$ in high-dimensional data,. The rate intrinsically depends on d, which represents the sparsity of the true regression coefficient vector. For other theoretical properties such as the prediction error and variable selection consistency, the readers can refer to [46] and [15].

Computation: Many optimization methods have been developed for the LASSO. Here we show optimization algorithms based on the proximal gradient method (PG) [72], the alternating direction method of multipliers (ADMM) [13] and the coordinate descent method (CD) [94].

The PG for the LASSO is give by

$$\beta^{(t+1)} = \arg\min_{\beta} \left\langle -\frac{2}{n}X^T(Y - X\beta^{(t)}), \beta \right\rangle + \lambda\|\beta\|_1 + \frac{1}{2\eta_t}\|\beta - \beta^{(t)}\|^2, \tag{2.12}$$

where η_t is the t-th iterative step size. The first term of the objective function is related to the first order approximation of $\frac{1}{n}\|Y - X\beta\|^2$ at $\beta^{(t)}$ as follows:

$$\begin{aligned}\frac{1}{n}\|Y - X\beta\|^2 &\approx \frac{1}{n}\|Y - X\beta^{(t)}\|^2 + \left\langle -\frac{2}{n}X^T(Y - X\beta^{(t)}), \beta - \beta^{(t)} \right\rangle \\ &= \left\langle -\frac{2}{n}X^T(Y - X\beta^{(t)}), \beta \right\rangle + const.\end{aligned}$$

To ensure the convergence, the t-th iterative step size is chosen to be $0 < \eta_t < \frac{n}{2\|X^T X\|_2}$, where A, $\|A\|_2$ is the spectral norm for a matrix A. Then, the update formula of (2.12) can be obtained explicitly as follows:

$$\beta_j^{(t+1)} = \begin{cases} \operatorname{sign}\left(\beta_j^{(t)} - \eta_t g_j\right)\left(\beta_j^{(t)} - \eta_t g_j - \lambda\eta_t\right) & \left(|\beta_j^{(t)} - \eta_t g_j| > \lambda\eta_t\right) \\ 0 & \left(|\beta_j^{(t)} - \eta_t g_j| \leq \lambda\eta_t\right), \end{cases}$$

where $\beta_j^{(t)}$ and g_j are the j-th element of $\beta^{(t)}$ and $-\frac{2}{n}(Y - X\beta^{(t)})X$, respectively.

The ADMM for the LASSO is given by

$$\left(\theta^{(t+1)}, \xi^{(t+1)}\right) = \underset{\theta,\xi}{\arg\min}\, \frac{1}{n}\|Y - X\theta\|^2 + \lambda\|\xi\|_1 + \langle h^{(t)}, \theta - \xi\rangle + \frac{\mu}{2}\|\theta - \xi\|^2, \quad (2.13)$$

where $h^{(t)}$ is t-th iterative Lagrange multipliers, and μ is the penalty parameter for the augmented Lagrangian. Then, the update formula of (2.13) can be obtained explicitly as follows:

$$\begin{aligned}\theta^{(t+1)} &= \left(\mu I_p + \frac{2}{n}X^T X\right)^{-1}(2X^T Y/n + \mu\xi^{(t)} - h^{(t)}), \\ \xi_j^{(t+1)} &= \begin{cases} \operatorname{sign}\left(\theta_j^{(t+1)} + h_j^{(t)}/\mu\right)\left(\theta_j^{(t+1)} + h_j^{(t)}/\mu - \lambda/\mu\right) & \left(|\theta_j^{(t+1)} + h_j^{(t)}/\mu| > \lambda/\mu\right) \\ 0 & \left(|\theta_j^{(t+1)} + h_j^{(t)}/\mu| \leq \lambda/\mu\right), \end{cases} \\ h^{(t+1)} &= h^{(t)} + \mu(\theta^{(t)} - \xi^{(t)}),\end{aligned}$$

where $\xi_j^{(t+1)}$, θ_j and $h_j^{(t)}$ are the j-th elements of $\xi^{(t+1)}$, θ and $h^{(t)}$, respectively. The ADMM splits the regression coefficient vector into θ and ξ. In practice, $\theta^{(t+1)}$, $\xi^{(t+1)}$ or $\frac{\theta^{(t+1)}+\xi^{(t+1)}}{2}$ is used as the final output.

The CD for the LASSO is given by

$$\beta_j^{(t+1)} = \arg\min_{\beta_j} \frac{1}{n} \|Y - X\beta_{[-j]}^{(t)}\|^2 + \lambda \|\beta_j\|_1, \tag{2.14}$$

where $\beta_{[-j]}^{(t)}$ is $(\beta_1^{(t+1)}, \ldots, \beta_{j-1}^{(t+1)}, \beta_j, \beta_{j+1}^{(t)}, \ldots, \beta_p^{(t)})$. Then, the update formula of (2.14) can be obtained explicitly as follows:

$$\beta_j^{(t+1)} = \begin{cases} \frac{\mathrm{sign}\left(\sum_{i=1}^n (y_i - r_{i,-j}^{(t)}) x_{ij}\right) \left(\sum_{i=1}^n (y_i - r_{i,-j}^{(t)}) x_{ij} - 2\lambda/n\right)}{\sum_{i=1}^n x_{ij}^2} & \left(|\sum_{i=1}^n (y_i - r_{i,-j}^{(t)}) x_{ij}| > 2\lambda/n\right) \\ 0 & \left(|\sum_{i=1}^n (y_i - r_{i,-j}^{(t)}) x_{ij}| \le 2\lambda/n\right), \end{cases}$$

where $r_{i,-j}^{(t)} = \sum_{k \neq j} x_{ik} (\mathbb{1}_{(k<j)} \beta_k^{(t+1)} + \mathbb{1}_{(k>j)} \beta_k^{(t)})$ and x_{ik} is the k-th element of x_i.

We compare these optimization methods applied to the LASSO in terms of convergence, parallelization, and scalability.

- Convergence: The PG needs to set the step size appropriately for ensuring convergence, while the ADMM and CD do not have a parameter which needs to be iteratively adjusted. Under some assumptions, every method can guarantee the convergence for the LASSO. However, the convergence rate of each method depends on one problem, i.e., the optimal method cannot be determined in terms of the convergence rate. For the CD, the convergence rate depends on the order of the update cycle through coordinates [94].

- Parallelization: The PG and ADMM can use parallel computing for β and ξ, respectively. However, the CD does not use parallel computing since the update formulas of each coordinate are dependent on each other.

- Scalability: The ADMM needs to calculate the inverse of the matrix $\left(\mu I_p + \frac{2}{n} X^T X\right)$. It is difficult for very high-dimensional data, e.g., for $p = 10^5$. Hence, the Cholesky decomposition and a matrix inversion lemma are used to decrease the computational cost (see Sect. 4.2.4 of [13]). To update the whole parameter β, the CD has to operate the update formula for every coordinate, β_j $(j = 1, \ldots, p)$, in one cycle. Increasing the number of coordinates results in a high computational cost because the CD cannot use parallel computing, and computational cost increases linearly with the number of coordinates.

Some software packages are available to compute the LASSO in the R language; e.g., “glamlasso”, “APG”, “ADMM”, “flare”, “glmnet”, and “lassoshooting”.

Chapter 3

Robust Linear Regression with Sparsity via γ-Divergence

3.1 γ-Regression with Sparsity

We adopt the type II of γ-divergence in this chapter. We have already investigated the estimation of γ-regression with non-sparsity in Sect 2.1.4. Here, we consider the estimation of γ-regression with sparsity. In what follows, we refer to the γ-regression with sparsity as the sparse γ-regression.

Let $f(y|x;\theta)$ be the conditional probability density function of y given x with parameter θ. The target parameter can be considered by:

$$\begin{aligned}
\theta^*_{\gamma,2} &= \arg\min_{\theta} D_{\gamma,2}(g(y|x), f(y|x;\theta); g(x)) \\
&= \arg\min_{\theta} d_{\gamma,2}(g(y|x), f(y|x;\theta); g(x)) \\
&= \arg\min_{\theta} -\frac{1}{\gamma}\log E_{g(x,y)}\left[f(y|x)^{\gamma}\right] + \frac{1}{1+\gamma}\log E_{g(x)}\left[\int f(y|x)^{1+\gamma}dy\right].
\end{aligned}$$

Moreover, we can also consider the γ-cross entropy and the target parameter with a regularization term, given by

$$\begin{aligned}
\theta^*_{\gamma_2,pen} &= \arg\min_{\theta} D_{\gamma,2}(g(y|x), f(y|x;\theta); g(x)) + \lambda P(\theta)\\
&= \arg\min_{\theta} d_{\gamma,2}(g(y|x), f(y|x;\theta); g(x)) + \lambda P(\theta)\\
&= \arg\min_{\theta} -\frac{1}{\gamma}\log E_{g(x,y)}\left[f(y|x)^{\gamma}\right] + \frac{1}{1+\gamma}\log E_{g(x)}\left[\int f(y|x)^{1+\gamma}dy\right] + \lambda P(\theta),
\end{aligned} \tag{3.1}$$

where $P(\theta)$ is a regularization term for parameter θ and λ is a tuning parameter for the regularization term. As an example of the penalty term, we can consider L_1 (Lasso, [87]), elasticnet [104], group Lasso [98], fused Lasso [89], and so on.

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be the observations randomly drawn from the underlying distribution $g(x, y)$. As we have seen in Sect. 2.1.4, the γ-cross entropy, $d_{\gamma,2}(g(y|x), f(y|x;\theta); g(x))$, can be empirically estimated by:

$$\bar{d}_{\gamma,2}(f(y|x;\theta)) = -\frac{1}{\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^{n} f(y_i|x_i;\theta)^{\gamma}\right\} + \frac{1}{1+\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^{n}\int f(y|x_i;\theta)^{1+\gamma}dy\right\}.$$

By virtue of (3.1), we define the sparse γ-estimator by:

$$\hat{\theta}_{\gamma_2,pen} = \arg\min_{\theta} \bar{d}_{\gamma,2}(f(y|x;\theta)) + \lambda P(\theta). \tag{3.2}$$

To obtain the minimizer, we propose the iterative algorithm by the majorization-minimization algorithm (MM algorithm) [52].

3.2 Parameter Estimation Procedure

3.2.1 MM Algorithm for Sparse γ-Regression

The MM algorithm is constructed as follows. Let $h(\eta)$ be the objective function. Let us prepare the majorization function h_{MM} satisfying:

$$\begin{aligned}
h_{MM}(\eta^{(m)}|\eta^{(m)}) &= h(\eta^{(m)}),\\
h_{MM}(\eta|\eta^{(m)}) &\geq h(\eta) \quad \text{for all } \eta,
\end{aligned}$$

where $\eta^{(m)}$ is the parameter of the m-th iterative step for $m = 0, 1, 2, \ldots$ Let us consider the iterative algorithm by:

$$\eta^{(m+1)} = \arg\min_{\eta} h_{MM}(\eta|\eta^{(m)}).$$

Then, we can show that the objective function $h(\eta)$ monotonically decreases at each step, because:

$$\begin{aligned} h(\eta^{(m)}) &= h_{MM}(\eta^{(m)}|\eta^{(m)}) \\ &\geq h_{MM}(\eta^{(m+1)}|\eta^{(m)}) \\ &\geq h(\eta^{(m+1)}). \end{aligned}$$

Note that $\eta^{(m+1)}$ does not necessarily have to be the minimizer of $h_{MM}(\eta|\eta^{(m)})$. We only need:

$$h_{MM}(\eta^{(m)}|\eta^{(m)}) \geq h_{MM}(\eta^{(m+1)}|\eta^{(m)}).$$

Here, we consider a convergence property of the MM algorithm. Let us denote the difference between h_{MM} and h as $H(\eta)$, i.e., $H(\eta) := h_{MM}(\eta) - h(\eta)$. We define the following directional derivative of h at $\eta^{(m)}$ in the direction $\eta - \eta^{(m)}$:

$$\nabla h\left(\eta^{(m)}, \eta - \eta^{(m)}\right) := \lim_{t \to +0} \frac{h(\eta^{(m)} + t(\eta - \eta^{(m)})) - h(\eta^{(m)})}{t}.$$

Then, a sequence $\left\{\eta^{(m)}\right\}_{m \geq 0}$ satisfies the following asymptotic stationary point condition under mild conditions [65]:

Proposition 3.2.1. [Asymptotic Stationary Point Condition in [65]] *If $H(\eta)$ is differentiable, its gradient is Lipschitz continuous, and $H(\eta') = \nabla H(\eta') = 0$, a sequence $\left\{\eta^{(m)}\right\}_{m \geq 0}$ satisfies*

$$\liminf_{m \to +\infty} \inf_{\eta} \frac{\nabla h\left(\eta^{(m)}, \eta - \eta^{(m)}\right)}{\|\eta - \eta^{(m)}\|} \geq 0.$$

Proof. See [65], Proposition 2.1. □

We construct the majorization function for the sparse γ-regression by the following inequality:

$$\kappa(z^T\eta) \leq \sum_i \frac{z_i\eta_i^{(m)}}{z^T\eta^{(m)}}\kappa\left[\eta_i\frac{z^T\eta^{(m)}}{\eta_i^{(m)}}\right], \tag{3.3}$$

where $\kappa(u)$ is a convex function, $z=(z_1,\dots,z_n)^T$, $\eta=(\eta_1,\dots,\eta_n)^T$, $\eta^{(m)}=(\eta_1^{(m)},\dots,\eta_n^{(m)})^T$, and z_i, η_i and $\eta_i^{(m)}$ are positive. The inequality (3.3) holds from Jensen's inequality. Here, we take $z_i=\frac{1}{n}$, $\eta_i=f(y_i|x_i;\theta)^\gamma$, $\eta_i^{(m)}=f(y_i|x_i;\theta^{(m)})^\gamma$, and $\kappa(u)=-\log u$ in (3.3). We can propose the majorization function as follows:

$$\begin{aligned}
&h(\theta)\\
&=L_\gamma(\theta;\lambda)\\
&=-\frac{1}{\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^n f(y_i|x_i;\theta)^\gamma\right\}+\frac{1}{1+\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^n\int f(y|x_i;\theta)^{1+\gamma}dy\right\}+\lambda P(\theta)\\
&\leq -\frac{1}{\gamma}\sum_{i=1}^n\alpha_i^{(m)}\log\left\{f(y_i|x_i;\theta)^\gamma\frac{\frac{1}{n}\sum_{l=1}^n f(y_l|x_l;\theta^{(m)})^\gamma}{f(y_i|x_i;\theta^{(m)})^\gamma}\right\}\\
&\qquad+\frac{1}{1+\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^n\int f(y|x_i;\theta)^{1+\gamma}dy\right\}+\lambda P(\theta)\\
&=-\sum_{i=1}^n\alpha_i^{(m)}\log f(y_i|x_i;\theta)+\frac{1}{1+\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^n\int f(y|x_i;\theta)^{1+\gamma}dy\right\}+\lambda P(\theta)\\
&\qquad+const.\\
&=h_{MM}(\theta|\theta^{(m)})+const.,
\end{aligned}$$

where

$$\alpha_i^{(m)}=\frac{f(y_i|x_i;\theta^{(m)})^\gamma}{\sum_{l=1}^n f(y_l|x_l;\theta^{(m)})^\gamma},$$

$$h_{MM}(\theta|\theta^{(m)})=-\sum_{i=1}^n\alpha_i^{(m)}\log f(y_i|x_i;\theta)+\frac{1}{1+\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^n\int f(y|x_i;\theta)^{1+\gamma}dy\right\}+\lambda P(\theta)$$

and $const.$ is a term that does not depend on the parameter θ.

The first term on the original target function $h(\theta)$ is a mixture type of densities, which is not easy to optimize, while the first term on $h_{MM}(\theta|\theta^{(m)})$ is a weighted log-likelihood, which is often easy to optimize.

3.2.2 Sparse γ-Linear Regression

Let $f(y|x;\theta)$ be the conditional density with $\theta = (\beta_0, \beta, \sigma^2)$, given by:

$$f(y|x;\theta) = \phi(y;\beta_0 + x^T\beta, \sigma^2),$$

where $\phi(y;\mu,\sigma^2)$ is the normal density with mean parameter μ and variance parameter σ^2. Suppose that $P(\theta)$ is the L_1 regularization $||\beta||_1$. After a simple calculation, we have:

$$h_{MM}(\theta|\theta^{(m)}) = \frac{1}{2(1+\gamma)}\log\sigma^2 + \frac{1}{2}\sum_{i=1}^{n}\alpha_i^{(m)}\frac{(y_i - \beta_0 - x_i^T\beta)^2}{\sigma^2} + \lambda||\beta||_1. \tag{3.4}$$

This function is easy to optimize by an update algorithm. For a fixed value of σ^2, the function h_{MM} is almost the same as Lasso except for the weight, so that it can be updated using the coordinate decent algorithm with a decreasing property of the loss function. For a fixed value of $(\beta_0, \beta^T)^T$, the function h_{MM} is easy to minimize. Consequently, we can obtain the update algorithm in Algorithm 1 with the decreasing property:

$$h_{MM}(\theta^{(m+1)}|\theta^{(m)}) \le h_{MM}(\theta^{(m)}|\theta^{(m)}).$$

It should be noted that h_{MM} is convex with respect to parameter β_0, β and has the global minimum with respect to parameter σ^2, but the original objective function h is not convex with respect to them, so that the initial points of Algorithm 1 are important. This issue is discussed in Sect. 3.4.4.

In practice, we also use the active set strategy [32] in the coordinate decent algorithm for updating $\beta^{(m)}$. The active set consists of the non-zero coordinates of $\beta^{(m)}$. Specifically, for a given $\beta^{(m)}$, we only update the non-zero coordinates of $\beta^{(m)}$, until they are converged. Then, the non-active set parameter estimates are updated once. When they remain zero, the coordinate descent algorithm stops. If some of them do not remain zero, those are added to the active set, and the coordinate descent algorithm continues.

Algorithm 1 Sparse γ-linear regression.

```
Input: $\beta_0^{(0)}, \beta^{(0)}, \sigma^{2(0)}$
  repeat $m = 0, 1, 2, \dots$
    $\alpha_i^{(m)} \leftarrow \frac{\phi(y_i;\beta_0{}^{(m)}+x_i^T\beta^{(m)},\sigma^{2(m)})^\gamma}{\sum_{l=1}^n \phi(y_l;\beta_0{}^{(m)}+x_l{}^T\beta^{(m)},\sigma^{2(m)})^\gamma} \quad (i = 1, 2, \dots, n).$
    $\beta_0{}^{(m+1)} \leftarrow \sum_{i=1}^n \alpha_i^{(m)} (y_i - x_i{}^T \beta^{(m)}).$
    for do $j = 1, \dots, p$
      $\beta_j{}^{(m+1)} \leftarrow \frac{S\left(\sum_{i=1}^n \alpha_i^{(m)}(y_i-\beta_0^{(m+1)}-r_{i,-j}^{(m)})x_{ij},\ \sigma^{2(m)}\lambda\right)}{\left(\sum_{i=1}^n \alpha_i^{(m)} x_{ij}^2\right)},$
      where $S(t,\lambda) = \mathrm{sign}(t)(|t| - \lambda)_+$ and $r_{i,-j}^{(m)} = \sum_{k \neq j} x_{ik}(\mathbb{1}_{(k<j)}\beta_k^{(m+1)} + \mathbb{1}_{(k>j)}\beta_k^{(m)}).$
    $\sigma^{2(m+1)} \leftarrow (1+\gamma)\sum_{i=1}^n \alpha_i^{(m)}(y_i-\beta_0^{(m+1)}-x_i^T\beta^{(m+1)})^2.$
  until convergence
Output: $\hat{\beta}_0, \hat{\beta}, \hat{\sigma}^2$
```

3.2.3 Robust Cross-Validation

In sparse regression, a regularization parameter is often selected via a criterion. Cross-validation is often used for selecting the regularization parameter. Ordinal cross-validation is based on the squared error, and it can also be constructed using the KL-cross entropy with the normal density. However, the ordinal cross-validation will fail due to outliers. Therefore, we propose the robust cross-validation based on the γ-cross entropy. Let $\hat{\theta}_\gamma$ be the robust estimate based on the γ-cross entropy. The cross-validation based on the γ-cross entropy can be given by:

$$\begin{aligned}&\mathrm{RoCV}(\lambda)\\&= -\frac{1}{\gamma_0}\log\left\{\frac{1}{n}\sum_{i=1}^{n} f(y_i|x_i;\hat{\theta}_\gamma^{[-i]})^{\gamma_0}\right\} + \frac{1}{1+\gamma_0}\log\left\{\frac{1}{n}\sum_{i=1}^{n}\int f(y|x_i;\hat{\theta}_\gamma^{[-i]})^{1+\gamma_0}dy\right\},\end{aligned}$$

where $\hat{\theta}_\gamma^{[-i]}$ is the γ-estimator deleting the i-th observation and γ_0 is an appropriate tuning parameter. We can also adopt the K-fold cross-validation to reduce the computational task [45].

Here, we give a small modification of the above. We often focus only on the mean structure for prediction, not on the variance parameter. Therefore, in this paper, $\hat{\theta}_\gamma^{[-i]} = \left(\hat{\beta}_\gamma^{[-i]}, \hat{\sigma^2}_\gamma^{[-i]}\right)$ is replaced by $\left(\hat{\beta}_\gamma^{[-i]}, \hat{\sigma^2}_{fix}\right)$. In numerical experiments and real data analyses, we used $\sigma^{2^{(0)}}$ as σ^2_{fix}.

3.3 Robust Properties

In this section, the robust properties are presented from two viewpoints of latent bias and Pythagorean relation. This section is closely related to Sect. 5.2.3. However, the discussion here is more detailed, and in addition we show the redescending property on the sparse γ linear regression.

The latent bias was discussed in Fujisawa and Eguchi [35] and Kanamori and Fujisawa [54], which is described later. Using the results obtained there, the Pythagorean relation is shown in Theorems 3.3.1 and 3.3.2.

Let $f^*(y|x) = f_{\theta^*}(y|x) = f(y|x;\theta^*)$ and $\delta(y|x)$ be the target conditional probability density function and the contamination conditional probability density function related to outliers, respectively. Let ε and $\varepsilon(x)$ denote the outlier ratios, which are independent of and dependent on x, respectively. Under homogeneous and heterogeneous contaminations, we suppose that the underlying conditional probability density function can be expressed as:

$$\begin{aligned} g(y|x) &= (1-\varepsilon)f(y|x;\theta^*) + \varepsilon\delta(y|x), \\ g(y|x) &= (1-\varepsilon(x))f(y|x;\theta^*) + \varepsilon(x)\delta(y|x). \end{aligned}$$

Let:

$$\nu_{f,\gamma}(x) = \left\{ \int \delta(y|x) f(y|x)^{\gamma} dy \right\}^{\frac{1}{\gamma}} \qquad (\gamma > 0),$$

and let:

$$\nu_{f,\gamma} = \left\{ \int \nu_{f,\gamma}(x)^{\gamma} g(x) dx \right\}^{\frac{1}{\gamma}}.$$

Here, we assume that:

$$\nu_{f_{\theta^*},\gamma} \approx 0,$$

which implies that $\nu_{f_{\theta^*},\gamma}(x) \approx 0$ for any x (a.e.) and illustrates that the contamination conditional probability density function $\delta(y|x)$ lies on the tail of the target conditional probability density function $f(y|x;\theta^*)$. For example, if $\delta(y|x)$ is the Dirac function at the outlier $y_{\dagger}(x)$ given x, then we have $\nu_{f_{\theta^*},\gamma}(x) = f(y_{\dagger}(x)|x;\theta^*)$, which should be sufficiently small because $y_{\dagger}(x)$ is an outlier. In this section, we show that $\theta^*_{\gamma} - \theta^*$ is expected to be small even if ε or $\varepsilon(x)$ is not small. To make the discussion easier, we prepare the monotone transformation of the γ-cross entropy for regression by:

$$\begin{aligned} &\tilde{d}_{\gamma}(g(y|x), f(y|x;\theta); g(x)) \\ &= -\exp\left\{-\gamma d_{\gamma}(g(y|x), f(y|x;\theta); g(x))\right\} \\ &= -\frac{\int \left(\int g(y|x) f(y|x;\theta)^{\gamma} dy\right) g(x) dx}{\left\{\int \left(\int f(y|x;\theta)^{1+\gamma} dy\right) g(x) dx\right\}^{\frac{\gamma}{1+\gamma}}}. \end{aligned}$$

3.3.1 Case of Homogeneous Contamination

Here, we provide the following proposition, which was given in Kanamori and Fujisawa [54].

Proposition 3.3.1. [Kanamori and Fujisawa [54], Section 4]

$$\begin{aligned}&\tilde{d}_\gamma(g(y|x), f(y|x;\theta); g(x))\\&= (1-\varepsilon)\tilde{d}_\gamma(f(y|x;\theta^*), f(y|x;\theta); g(x)) - \frac{\varepsilon \nu_{f_\theta,\gamma}^{\gamma}}{\{\int\left(\int f(y|x;\theta)^{1+\gamma}dy\right)g(x)dx\}^{\frac{\gamma}{1+\gamma}}}.\end{aligned}$$

Recall that θ_γ^* and θ^* are also the minimizers of $\tilde{d}_\gamma(g(y|x), f(y|x;\theta); g(x))$ and $\tilde{d}_\gamma(f(y|x;\theta^*), f(y|x;\theta); g(x))$, respectively. We can expect $\nu_{f_\theta,\gamma} \approx 0$ from the assumption $\nu_{f_{\theta^*},\gamma} \approx 0$ if the tail behavior of $f(y|x;\theta)$ is close to that of $f(y|x;\theta^*)$. We see from Proposition 3.3.1 and the condition $\nu_{f_\theta,\gamma} \approx 0$ that:

$$\begin{aligned}\theta_\gamma^* &= \arg\min_\theta \tilde{d}_\gamma(g(y|x), f(y|x;\theta); g(x))\\&= \arg\min_\theta \Big[(1-\varepsilon)\tilde{d}_\gamma(f(y|x;\theta^*), f(y|x;\theta); g(x))\\&\qquad - \frac{\varepsilon \nu_{f_\theta,\gamma}^{\gamma}}{\{\int\left(\int f(y|x;\theta)^{1+\gamma}dy\right)g(x)dx\}^{\frac{\gamma}{1+\gamma}}}\Big]\\&\approx \arg\min_\theta (1-\varepsilon)\tilde{d}_\gamma(f(y|x;\theta^*), f(y|x;\theta); g(x))\\&= \theta^*.\end{aligned}$$

Therefore, under homogeneous contamination, it can be expected that the latent bias $\theta_\gamma^* - \theta^*$ is small even if ε is not small. Moreover, we can show the following theorem, using Proposition 3.3.1.

Theorem 3.3.1. *Let $\nu = max\{\nu_{f_\theta,\gamma}, \nu_{f_{\theta^*},\gamma}\}$. Then, the Pythagorean relation among $g(y|x)$, $f(y|x;\theta^*)$, $f(y|x;\theta)$ approximately holds:*

$$\begin{aligned}&D_\gamma(g(y|x), f(y|x;\theta); g(x)) - D_\gamma(g(y|x), f(y|x;\theta^*); g(x))\\&= D_\gamma(f(y|x;\theta^*), f(y|x;\theta); g(x)) + O(\nu^\gamma).\end{aligned}$$

The proof is in Appendix B. The Pythagorean relation implies that the minimization of the divergence from $f(y|x;\theta)$ to the underlying conditional probability density function $g(y|x)$ is approximately the same as that to the target conditional probability density function $f(y|x;\theta^*)$. Therefore, under homogeneous contamination, we can see why our proposed method works well in terms of the minimization of the γ-divergence.

3.3.2 Case of Heterogeneous Contamination

Under heterogeneous contamination, we assume that the parametric conditional probability density function $f(y|x;\theta)$ is a location-scale family given by:

$$f(y|x;\theta) = \frac{1}{\sigma} s\left(\frac{y-q(x;\xi)}{\sigma}\right),$$

where $s(y)$ is a probability density function, σ is a scale parameter and $q(x;\xi)$ is a location function with a regression parameter ξ, e.g., $q(x;\xi) = \xi^T x$. Then, we can obtain:

$$\begin{aligned}\int f(y|x;\theta)^{1+\gamma}dy &= \int \frac{1}{\sigma^{1+\gamma}} s\left(\frac{y-q(x;\xi)}{\sigma}\right)^{1+\gamma} dy \\ &= \sigma^{-\gamma}\int s(z)^{1+\gamma}dz.\end{aligned}$$

That does not depend on the explanatory variable x. Here, we provide the following proposition, which was given in Kanamori and Fujisawa [54].

Proposition 3.3.2. [Kanamori and Fujisawa [54], Section 4]

$$\begin{aligned}&\tilde{d}_\gamma(g(y|x), f(y|x;\theta); g(x)) \\ &= c\tilde{d}_\gamma(f(y|x;\theta^*), f(y|x;\theta); \tilde{g}(x)) - \frac{\int \nu_{f_\theta,\gamma}(x)^\gamma \varepsilon(x) g(x)dx}{\left\{\sigma^{-\gamma}\int s(z)^{1+\gamma}dz\right\}^{\frac{\gamma}{1+\gamma}}},\end{aligned}$$

where $c = (1-\int \varepsilon(x)g(x)dx)^{\frac{\gamma}{1+\gamma}}$ *and* $\tilde{g}(x) = (1-\varepsilon(x))g(x)$.

The second term $\frac{\int \nu_{f_\theta,\gamma}(x)^\gamma \varepsilon(x) g(x)dx}{\left\{\sigma^{-\gamma}\int s(z)^{1+\gamma}dz\right\}^{\frac{\gamma}{1+\gamma}}}$ can be approximated to be zero from the condition $\nu_{f_\theta,\gamma} \approx 0$ and $\varepsilon(x) < 1$ as follows:

$$\begin{aligned}\frac{\int \nu_{f_\theta,\gamma}(x)^\gamma \varepsilon(x) g(x)dx}{\left\{\sigma^{-\gamma}\int s(z)^{1+\gamma}dz\right\}^{\frac{\gamma}{1+\gamma}}} &< \frac{\int \nu_{f_\theta,\gamma}(x)^\gamma g(x)dx}{\left\{\sigma^{-\gamma}\int s(z)^{1+\gamma}dz\right\}^{\frac{\gamma}{1+\gamma}}} \\ &= \frac{\nu_{f_\theta,\gamma}^\gamma}{\left\{\sigma^{-\gamma}\int s(z)^{1+\gamma}dz\right\}^{\frac{\gamma}{1+\gamma}}} \\ &\approx 0. \qquad (3.5)\end{aligned}$$

We see from Proposition 3.3.2 and (3.5) that:

$$\begin{aligned}
\theta^*_\gamma &= \arg\min_\theta \tilde{d}_\gamma(g(y|x), f(y|x;\theta); g(x)) \\
&= \arg\min_\theta \Big[c\tilde{d}_\gamma(f(y|x;\theta^*), f(y|x;\theta); \tilde{g}(x)) \\
&\qquad\qquad - \frac{\int \nu_{f_\theta,\gamma}(x)^\gamma \varepsilon(x) g(x) dx}{\{\sigma^{-\gamma}\int s(z)^{1+\gamma} dz\}^{\frac{\gamma}{1+\gamma}}} \Big] \\
&\approx \arg\min_\theta c\tilde{d}_\gamma(f(y|x;\theta^*), f(y|x;\theta); \tilde{g}(x)) \\
&= \theta^*.
\end{aligned}$$

Therefore, under heterogeneous contamination in a location-scale family, it can be expected that the latent bias $\theta^*_\gamma - \theta^*$ is small even if $\varepsilon(x)$ is not small. Moreover, we can show the following theorem, using Proposition 3.3.2.

Theorem 3.3.2. *Let $\nu = max\{\nu_{f_\theta,\gamma}, \nu_{f_{\theta^*},\gamma}\}$. Then, the following relation among $g(y|x)$, $f(y|x;\theta^*)$, $f(y|x;\theta)$ approximately holds:*

$$\begin{aligned}
&D_\gamma(g(y|x), f(y|x;\theta); g(x)) - D_\gamma(g(y|x), f(y|x;\theta^*); g(x)) \\
&= D_\gamma(f(y|x;\theta^*), f(y|x;\theta); \tilde{g}(x)) + O(\nu^\gamma).
\end{aligned}$$

The proof is in Appendix B. The above is slightly different from a conventional Pythagorean relation, because the base measure changes from $g(x)$ to $\tilde{g}(x)$ in part. However, it also implies that the minimization of the divergence from $f(y|x;\theta)$ to the underlying conditional probability density function $g(y|x)$ is approximately the same as that to the target conditional probability density function $f(y|x;\theta^*)$. Therefore, under heterogeneous contamination in a location-scale family, we can see why our proposed method works well in terms of the minimization of the γ-divergence.

3.3.3 Redescending Property

First, we review a redescending property on M-estimation (see, e.g., [66]), which is often used in robust statistics. Suppose that the estimating equation is given by $\sum_{i=1}^n \zeta(z_i;\theta) = 0$. Let $\hat{\theta}$ be a solution of the estimating equation. The bias caused by outlier z_o is expressed as $\hat{\theta}_{n=\infty} - \theta^*$, where $\hat{\theta}_{n=\infty}$ is the limiting value of $\hat{\theta}$ and θ^* is the true parameter. We hope the bias is small even if the outlier z_o exists. Under some conditions, the bias can be approximated to $\varepsilon\mathrm{IF}(z_o;\theta^*)$, where ε is a small outlier ratio and $\mathrm{IF}(z;\theta^*)$ is the influence function. The bias is expected to be small when the influence function is small. The influence function can be expressed as $\mathrm{IF}(z;\theta^*) = A\zeta(z;\theta^*)$, where A is a matrix independent of z, so that the bias is also expected to be small when $\zeta(z_o;\theta^*)$ is small. In particular, the estimating equation is said to have a redescending property if $\zeta(z;\theta^*)$ goes to zero as $||z||$ goes to infinity. This property is favorable in robust statistics, because the bias is expected to be sufficiently small when z_o is very large.

Here, we prove a redescending property on the sparse γ-linear regression, i.e., when $f(y|x;\theta) = \phi(y;\beta_0 + x^T\beta, \sigma^2)$ with $\theta = (\beta_0, \beta, \sigma^2)$ for fixed x. Recall that the estimate of the sparse γ-linear regression is the minimizer of the loss function:

$$L_\gamma(\theta;\lambda) = -\frac{1}{\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^{n}\phi(y_i;\beta_0 + {x_i}^T\beta, \sigma^2)^\gamma\right\} + b_\gamma(\theta;\lambda),$$

where $b_\gamma(\theta;\lambda) = \frac{1}{1+\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^{n}\int\phi(y;\beta_0 + {x_i}^T\beta, \sigma^2)^{1+\gamma}dy\right\} + \lambda||\beta||_1$. Then, the estimating equation is given by:

$$\begin{aligned}0 &= \frac{\partial}{\partial\theta}L_\gamma(\theta;\lambda)\\ &= -\frac{\sum_{i=1}^{n}\phi(y_i;\beta_0 + {x_i}^T\beta, \sigma^2)^\gamma s(y_i|x_i;\theta)}{\sum_{i=1}^{n}\phi(y_i;\beta_0 + {x_i}^T\beta, \sigma^2)^\gamma} + \frac{\partial}{\partial\theta}b_\gamma(\theta;\lambda),\end{aligned}$$

where $s(y|x;\theta) = \frac{\partial\log\phi(y;\beta_0 + x^T\beta, \sigma^2)}{\partial\theta}$. This can be expressed by the M-estimation formula given by:

$$0 = \sum_{i=1}^{n}\psi(y_i|x_i;\theta),$$

where $\psi(y|x;\theta) = \phi(y;\beta_0 + x^T\beta, \sigma^2)^\gamma s(y|x;\theta) - \phi(y;\beta_0 + x^T\beta, \sigma^2)^\gamma\frac{\partial}{\partial\theta}b_\gamma(\theta;\lambda)$. We can easily show that as $||y||$ goes to infinity, $\phi(y;\beta_0 + x^T\beta, \sigma^2)$ goes to zero and $\phi(y;\beta_0 + x^T\beta, \sigma^2)s(y|x;\theta)$ also goes to zero. Therefore, the function $\psi(y|x;\theta)$ goes to zero as $||y||$ goes to infinity, so that the estimating equation has a redescending property.

3.4 Numerical Experiment

In this section, we compare our method (sparse γ-linear regression) with the representative sparse linear regression method, the least absolute shrinkage and selection operator (Lasso) [87], and the robust and sparse regression methods, sparse least trimmed squares (sLTS) [2] and robust least angle regression (RLARS) [58].

3.4.1 Simulation Model

We used the simulation model given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e, \quad e \sim N(0, 0.5^2).$$

The sample size and the number of explanatory variables were set to be $n = 100$ and $p = 100, 200$, respectively. The true coefficients were given by:

$$\beta_1 = 1,\ \beta_2 = 2,\ \beta_4 = 4,\ \beta_7 = 7,\ \beta_{11} = 11,$$
$$\beta_j = 0 \text{ for } j \in \{0, \ldots, p\} \backslash \{1, 2, 4, 7, 11\}.$$

We arranged a broad range of regression coefficients to observe sparsity for various degrees of regression coefficients. The explanatory variables were generated from a normal distribution $N(0, \Sigma)$ with $\Sigma = (\rho^{|i-j|})_{1 \le i,j \le p}$. We generated 100 random samples.

Outliers were incorporated into simulations. We investigated two outlier ratios ($\varepsilon =$ 0.1 and 0.3) and two outlier patterns: (a) the outliers were generated around the middle part of the explanatory variable, where the explanatory variables were generated from $N(0, 0.5^2)$ and the error terms were generated from $N(20, 0.5^2)$; (b) the outliers were generated around the edge part of the explanatory variable, where the explanatory variables were generated from $N(-1.5, 0.5^2)$ and the error terms were generated from $N(20, 0.5^2)$.

3.4.2 Performance Measure

The root mean squared prediction error (RMSPE) and mean squared error (MSE) were examined to verify the predictive performance and fitness of regression coefficient:

$$\text{RMSPE}(\hat{\beta}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i^* - x_i^{*T} \hat{\beta})^2},$$
$$\text{MSE} = \frac{1}{p+1} \sum_{j=0}^{p} (\beta_j^* - \hat{\beta}_j)^2,$$

where (x_i^*, y_i^*) $(i = 1, \ldots, n)$ is the test sample generated from the simulation model without outliers and β_j^*'s are the true coefficients. The true positive rate (TPR) and true negative rate

(TNR) were also reported to verify the sparsity:

$$\mathrm{TPR}(\hat{\beta}) = \frac{|\{j \in \{1,\dots,p\} : \hat{\beta}_j \neq 0 \wedge \beta_j^* \neq 0\}|}{|\{j \in \{1,\dots,p\} : \beta_j^* \neq 0\}|},$$

$$\mathrm{TNR}(\hat{\beta}) = \frac{|\{j \in \{1,\dots,p\} : \hat{\beta}_j = 0 \wedge \beta_j^* = 0\}|}{|\{j \in \{1,\dots,p\} : \beta_j^* = 0\}|}.$$

3.4.3 Comparative Methods

In this subsection, we explain three comparative methods: Lasso, RLARS and sLTS.

Lasso is performed by the R-package “glmnet”. The regularization parameter λ_{Lasso} is selected by grid search via cross-validation in “glmnet”. We used “glmnet” by default.

RLARS is performed by the R-package “robustHD”. This is a robust version of LARS [26]. The optimal model is selected via BIC by default.

sLTS is performed by the R-package “robustHD”. sLTS has the regularization parameter λ_{sLTS} and the fraction parameter α of squared residuals used for trimmed squares. The regularization parameter λ_{sLTS} is selected by grid search via BIC. The number of grids is 40 by default. However, we considered that this would be small under heavy contamination. Therefore, we used 80 grids under heavy contamination to obtain a good performance. The fraction parameter α is 0.75 by default. In the case of $\alpha = 0.75$, the ratio of outlier is less than 25%. We considered this would be small under heavy contamination and large under low contamination in terms of statistical efficiency. Therefore, we used 0.65, 0.75, 0.85 as α under low contamination and 0.50, 0.65, 0.75 under heavy contamination.

3.4.4 Details of Our Method

Initial Points

In our method, we need an initial point to obtain the estimate, because we use the iterative algorithm proposed in Sect. 3.2.2. The estimate of other conventional robust and sparse regression methods would give a good initial point. For another choice, the estimate of RANSAC (random sample consensus) algorithm would also give a good initial point. In this experiment, we used the estimate of sLTS as an initial point.

How to Choose Tuning Parameters

In our method, we have to choose some tuning parameters. The parameter γ in the γ-divergence was set to 0.1 or 0.5. The parameter γ_0 in the robust cross-validation was set to

0.5. In our experience, the result via RoCV is not sensitive to the selection of γ_0 when γ_0 is large enough, e.g., $\gamma_0 = 0.5, 1$. The parameter λ of L_1 regularization is often selected via grid search. We used 50 grids in the range $[0.05\lambda_0, \lambda_0]$ with the log scale, where λ_0 is an estimate of λ, which would shrink regression coefficients to zero. More specifically, in a similar way as in Lasso, we can derive λ_0, which shrinks the coefficients β to zero in $h_{MM}(\theta|\theta^{(0)})$ [3.4] with respect to β, and we used it. This idea was proposed by the R-package "glmnet".

3.4.5 Result

Table 3.1 is the low contamination case with Outlier Pattern (a). For the RMSPE, our method outperformed other comparative methods (the oracle value of the RMSPE is 0.5). For the TPR and TNR, sLTS showed a similar performance to our method. Lasso presented the worst performance, because it is sensitive to outliers. Table 3.2 is the heavy contamination case with Outlier Pattern (a). For the RMSPE, our method outperformed other comparative methods except in the case $(p, \varepsilon, \rho) = (100, 0.3, 0.2)$ for sLTS with $\alpha = 0.5$. Lasso also presented a worse performance, and furthermore, sLTS with $\alpha = 0.75$ showed the worst performance due to a lack of truncation. For the TPR and TNR, our method showed the best performance. Table 3.3 is the low contamination case with Outlier Pattern (b). For the RMSPE, our method outperformed other comparative methods (the oracle value of the RMSPE is 0.5). For the TPR and TNR, sLTS showed a similar performance to our method. Lasso presented the worst performance, because it is sensitive to outliers. Table 3.4 is the heavy contamination case with Outlier Pattern (b). For the RMSPE, our method outperformed other comparative methods. sLTS with $\alpha = 0.5$ showed the worst performance. For the TPR and TNR, it seems that our method showed the best performance. Table 3.5 is the no contamination case. RLARS showed the best performance, but our method presented comparable performances. In spite of no contamination case, Lasso was clearly worse than RLARS and our method. This would be because the underlying distribution can generate a large value in simulation, although it is a small probability.

Table 3.1 Outlier Pattern (a) with $p = 100,\ 200,\ \varepsilon = 0.1$ and $\rho = 0.2,\ 0.5$. RMSPE, root mean squared prediction error (RMSPE); RLARS, robust least angle regression; sLTS, sparse least trimmed squares.

	$p = 100, \varepsilon = 0.1, \rho = 0.2$				$p = 100, \varepsilon = 0.1, \rho = 0.5$			
Methods	**RMSPE**	**MSE**	**TPR**	**TNR**	**RMSPE**	**MSE**	**TPR**	**TNR**
Lasso	3.04	9.72×10^{-2}	0.936	0.909	3.1	1.05×10^{-1}	0.952	0.918
RLARS	0.806	6.46×10^{-3}	0.936	0.949	0.718	6.7×10^{-3}	0.944	0.962
sLTS ($\alpha = 0.85$, 80 grids)	0.626	1.34×10^{-3}	1.0	0.964	0.599	1.05×10^{-3}	1.0	0.966
sLTS ($\alpha = 0.75$, 80 grids)	0.651	1.71×10^{-3}	1.0	0.961	0.623	1.33×10^{-3}	1.0	0.961
sLTS ($\alpha = 0.65$, 80 grids)	0.685	2.31×10^{-3}	1.0	0.957	0.668	1.76×10^{-3}	1.0	0.961
sparse γ-linear reg ($\gamma = 0.1$)	0.557	6.71×10^{-4}	1.0	0.966	0.561	6.99×10^{-4}	1.0	0.965
sparse γ-linear reg ($\gamma = 0.5$)	0.575	8.25×10^{-4}	1.0	0.961	0.573	9.05×10^{-4}	1.0	0.959
	$p = 200, \varepsilon = 0.1, \rho = 0.2$				$p = 200, \varepsilon = 0.1, \rho = 0.5$			
Methods	**RMSPE**	**MSE**	**TPR**	**TNR**	**RMSPE**	**MSE**	**TPR**	**TNR**
Lasso	3.55	6.28×10^{-2}	0.904	0.956	3.37	6.08×10^{-2}	0.928	0.961
RLARS	0.88	3.8×10^{-3}	0.904	0.977	0.843	4.46×10^{-3}	0.9	0.986
sLTS ($\alpha = 0.85$, 80 grids)	0.631	7.48×10^{-4}	1.0	0.972	0.614	5.77×10^{-4}	1.0	0.976
sLTS ($\alpha = 0.75$, 80 grids)	0.677	1.03×10^{-3}	1.0	0.966	0.632	7.08×10^{-4}	1.0	0.973
sLTS ($\alpha = 0.65$, 80 grids)	0.823	2.34×10^{-3}	0.998	0.96	0.7	1.25×10^{-3}	1.0	0.967
sparse γ-linear reg ($\gamma = 0.1$)	0.58	4.19×10^{-4}	1.0	0.981	0.557	3.71×10^{-4}	1.0	0.977
sparse γ-linear reg ($\gamma = 0.5$)	0.589	5.15×10^{-4}	1.0	0.979	0.586	5.13×10^{-4}	1.0	0.977

Table 3.2 Outlier Pattern (a) with $p = 100,\ 200$, $\varepsilon = 0.3$ and $\rho = 0.2,\ 0.5$.

	$p = 100$, $\varepsilon = 0.3$, $\rho = 0.2$				$p = 100$, $\varepsilon = 0.3$, $\rho = 0.5$			
Methods	**RMSPE**	**MSE**	**TPR**	**TNR**	**RMSPE**	**MSE**	**TPR**	**TNR**
Lasso	8.07	6.72×10^{-1}	0.806	0.903	8.1	3.32×10^{-1}	0.8	0.952
RLARS	2.65	1.54×10^{-1}	0.75	0.963	2.09	1.17×10^{-1}	0.812	0.966
sLTS ($\alpha = 0.75$, 80 grids)	10.4	2.08	0.886	0.709	11.7	2.36	0.854	0.67
sLTS ($\alpha = 0.65$, 80 grids)	2.12	3.66×10^{-1}	0.972	0.899	2.89	5.13×10^{-1}	0.966	0.887
sLTS ($\alpha = 0.5$, 80 grids)	1.37	1.46×10^{-1}	0.984	0.896	1.53	1.97×10^{-1}	0.976	0.909
sparse γ-linear reg ($\gamma = 0.1$)	1.13	9.16×10^{-2}	0.964	0.97	0.961	5.38×10^{-2}	0.982	0.977
sparse γ-linear reg ($\gamma = 0.5$)	1.28	1.5×10^{-1}	0.986	0.952	1.00	8.48×10^{-2}	0.988	0.958
	$p = 200$, $\varepsilon = 0.3$, $\rho = 0.2$				$p = 200$, $\varepsilon = 0.3$, $\rho = 0.5$			
Methods	**RMSPE**	**MSE**	**TPR**	**TNR**	**RMSPE**	**MSE**	**TPR**	**TNR**
Lasso	8.11	3.4×10^{-1}	0.77	0.951	8.02	6.51×10^{-1}	0.81	0.91
RLARS	3.6	1.7×10^{-1}	0.71	0.978	2.67	1.02×10^{-1}	0.76	0.984
sLTS ($\alpha = 0.75$, 80 grids)	11.5	1.16	0.738	0.809	11.9	1.17	0.78	0.811
sLTS ($\alpha = 0.65$, 80 grids)	3.34	3.01×10^{-1}	0.94	0.929	4.22	4.08×10^{-1}	0.928	0.924
sLTS ($\alpha = 0.5$, 80 grids)	4.02	3.33×10^{-1}	0.892	0.903	4.94	4.44×10^{-1}	0.842	0.909
sparse γ-linear reg ($\gamma = 0.1$)	2.03	1.45×10^{-1}	0.964	0.924	3.2	2.86×10^{-1}	0.94	0.936
sparse γ-linear reg ($\gamma = 0.5$)	1.23	7.69×10^{-2}	0.988	0.942	3.13	2.98×10^{-1}	0.944	0.94

Table 3.3 Outlier Pattern (b) with $p = 100,\ 200$, $\varepsilon = 0.1$ and $\rho = 0.2,\ 0.5$.

	$p = 100$, $\varepsilon = 0.1$, $\rho = 0.2$				$p = 100$, $\varepsilon = 0.1$, $\rho = 0.5$			
Methods	**RMSPE**	**MSE**	**TPR**	**TNR**	**RMSPE**	**MSE**	**TPR**	**TNR**
Lasso	2.48	5.31×10^{-2}	0.982	0.518	2.84	5.91×10^{-2}	0.98	0.565
RLARS	0.85	6.58×10^{-3}	0.93	0.827	0.829	7.97×10^{-3}	0.91	0.885
sLTS ($\alpha = 0.85$, 80 grids)	0.734	5.21×10^{-3}	0.998	0.964	0.684	3.76×10^{-3}	1.0	0.961
sLTS ($\alpha = 0.75$, 80 grids)	0.66	1.78×10^{-3}	1.0	0.975	0.648	1.59×10^{-3}	1.0	0.961
sLTS ($\alpha = 0.65$, 80 grids)	0.734	2.9×10^{-3}	1.0	0.96	0.66	1.74×10^{-3}	1.0	0.962
sparse γ-linear reg ($\gamma = 0.1$)	0.577	8.54×10^{-4}	1.0	0.894	0.545	5.44×10^{-4}	1.0	0.975
sparse γ-linear reg ($\gamma = 0.5$)	0.581	7.96×10^{-4}	1.0	0.971	0.546	5.95×10^{-4}	1.0	0.977
	$p = 200$, $\varepsilon = 0.1$, $\rho = 0.2$				$p = 200$, $\varepsilon = 0.1$, $\rho = 0.5$			
Methods	**RMSPE**	**MSE**	**TPR**	**TNR**	**RMSPE**	**MSE**	**TPR**	**TNR**
Lasso	2.39	2.57×10^{-2}	0.988	0.696	2.57	2.54×10^{-2}	0.944	0.706
RLARS	1.01	5.44×10^{-3}	0.896	0.923	0.877	4.82×10^{-3}	0.898	0.94
sLTS ($\alpha = 0.85$, 80 grids)	0.708	1.91×10^{-3}	1.0	0.975	0.790	3.40×10^{-3}	0.994	0.97
sLTS ($\alpha = 0.75$, 80 grids)	0.683	1.06×10^{-4}	1.0	0.975	0.635	7.40×10^{-4}	1.0	0.977
sLTS ($\alpha = 0.65$, 80 grids)	1.11	1.13×10^{-2}	0.984	0.956	0.768	2.60×10^{-3}	0.998	0.968
sparse γ-linear reg ($\gamma = 0.1$)	0.603	5.71×10^{-4}	1.0	0.924	0.563	3.78×10^{-3}	1.0	0.979
sparse γ-linear reg ($\gamma = 0.5$)	0.592	5.04×10^{-4}	1.0	0.982	0.566	4.05×10^{-3}	1.0	0.981

Table 3.4 Outlier Pattern (b) with $p = 100,\ 200$, $\varepsilon = 0.3$ and $\rho = 0.2,\ 0.5$.

	$p = 100$, $\varepsilon = 0.3$, $\rho = 0.2$				$p = 100$, $\varepsilon = 0.3$, $\rho = 0.5$			
Methods	**RMSPE**	**MSE**	**TPR**	**TNR**	**RMSPE**	**MSE**	**TPR**	**TNR**
Lasso	2.81	6.88×10^{-2}	0.956	0.567	3.13	7.11×10^{-2}	0.97	0.584
RLARS	2.70	7.69×10^{-2}	0.872	0.789	2.22	6.1×10^{-2}	0.852	0.855
sLTS ($\alpha = 0.75$, 80 grids)	3.99	1.57×10^{-1}	0.856	0.757	4.18	1.54×10^{-1}	0.878	0.771
sLTS ($\alpha = 0.65$, 80 grids)	3.2	1.46×10^{-1}	0.888	0.854	2.69	1.08×10^{-1}	0.922	0.867
sLTS ($\alpha = 0.5$, 80 grids)	6.51	4.62×10^{-1}	0.77	0.772	7.14	5.11×10^{-1}	0.844	0.778
sparse γ-linear reg ($\gamma = 0.1$)	1.75	3.89×10^{-2}	0.974	0.725	1.47	2.66×10^{-2}	0.976	0.865
sparse γ-linear reg ($\gamma = 0.5$)	1.68	3.44×10^{-2}	0.98	0.782	1.65	3.58×10^{-2}	0.974	0.863
	$p = 200$, $\varepsilon = 0.3$, $\rho = 0.2$				$p = 200$, $\varepsilon = 0.3$, $\rho = 0.5$			
Methods	**RMSPE**	**MSE**	**TPR**	**TNR**	**RMSPE**	**MSE**	**TPR**	**TNR**
Lasso	2.71	3.32×10^{-2}	0.964	0.734	2.86	3.05×10^{-2}	0.974	0.728
RLARS	3.03	4.59×10^{-2}	0.844	0.876	2.85	4.33×10^{-2}	0.862	0.896
sLTS ($\alpha = 0.75$, 80 grids)	3.73	7.95×10^{-2}	0.864	0.872	4.20	8.17×10^{-2}	0.878	0.87
sLTS ($\alpha = 0.65$, 80 grids)	4.45	1.23×10^{-1}	0.85	0.886	3.61	8.95×10^{-2}	0.904	0.908
sLTS ($\alpha = 0.5$, 80 grids)	9.05	4.24×10^{-1}	0.66	0.853	8.63	3.73×10^{-1}	0.748	0.864
sparse γ-linear reg ($\gamma = 0.1$)	1.78	1.62×10^{-2}	0.994	0.731	1.82	1.62×10^{-2}	0.988	0.844
sparse γ-linear reg ($\gamma = 0.5$)	1.79	1.69×10^{-2}	0.988	0.79	1.77	1.51×10^{-2}	0.996	0.77

Table 3.5 No contamination case with $p = 100,\ 200$, $\varepsilon = 0$ and $\rho = 0.2,\ 0.5$.

	$p = 100$, $\varepsilon = 0$, $\rho = 0.2$				$p = 100$, $\varepsilon = 0$, $\rho = 0.5$			
Methods	**RMSPE**	**MSE**	**TPR**	**TNR**	**RMSPE**	**MSE**	**TPR**	**TNR**
Lasso	0.621	1.34×10^{-3}	1.0	0.987	0.621	1.12×10^{-3}	1.0	0.987
RLARS	0.551	7.15×10^{-4}	0.996	0.969	0.543	6.74×10^{-4}	0.996	0.971
sLTS ($\alpha = 0.75$, 40 grids)	0.954	4.47×10^{-3}	1.0	0.996	0.899	4.53×10^{-3}	1.0	0.993
sparse γ-linear reg ($\gamma = 0.1$)	0.564	7.27×10^{-4}	1.0	0.878	0.565	6.59×10^{-4}	1.0	0.908
sparse γ-linear reg ($\gamma = 0.5$)	0.59	1.0×10^{-3}	1.0	0.923	0.584	8.47×10^{-4}	1.0	0.94
	$p = 200$, $\varepsilon = 0$, $\rho = 0.2$				$p = 200$, $\varepsilon = 0$, $\rho = 0.5$			
Methods	**RMSPE**	**MSE**	**TPR**	**TNR**	**RMSPE**	**MSE**	**TPR**	**TNR**
Lasso	0.635	7.18×10^{-4}	1.0	0.992	0.624	6.17×10^{-4}	1.0	0.991
RLARS	0.55	3.63×10^{-4}	0.994	0.983	0.544	3.48×10^{-4}	0.996	0.985
sLTS ($\alpha = 0.75$, 40 grids)	1.01	3.76×10^{-3}	1.0	0.996	0.909	2.47×10^{-3}	1.0	0.996
sparse γ-linear reg ($\gamma = 0.1$)	0.584	4.45×10^{-4}	1.0	0.935	0.573	3.99×10^{-4}	1.0	0.938
sparse γ-linear reg ($\gamma = 0.5$)	0.621	6.55×10^{-4}	1.0	0.967	0.602	5.58×10^{-4}	1.0	0.966

3.4.6 Computational Cost

In this subsection, we consider the CPU times for Lasso, RLARS, sLTS and our method. The data were generated from the simulation model in Sect. 4.5.1. The sample size and the number of explanatory variables were set to be $n = 100$ and $p = 100, 500, 1000, 2000, 5000$, respectively. In Lasso, RLARS and sLTS, all parameters were used by default (see Sect. 3.4.3). Our method used the estimate of the RANSAC algorithm as an initial point. The number of candidates for the RANSAC algorithm was set to 1000. The parameters γ and γ_0 were set to 0.1 and 0.5, respectively. No method used parallel computing methods. Figure 3.1 shows the average CPU times over 10 runs in seconds. All results were obtained in R Version 3.3.0 with an Intel Core i7-4790K machine. sLTS shows very high computational cost. RLARS is faster, but does not give a good estimate, as seen in Sect. 3.4.5. Our proposed method is fast enough even for $p = 5000$.

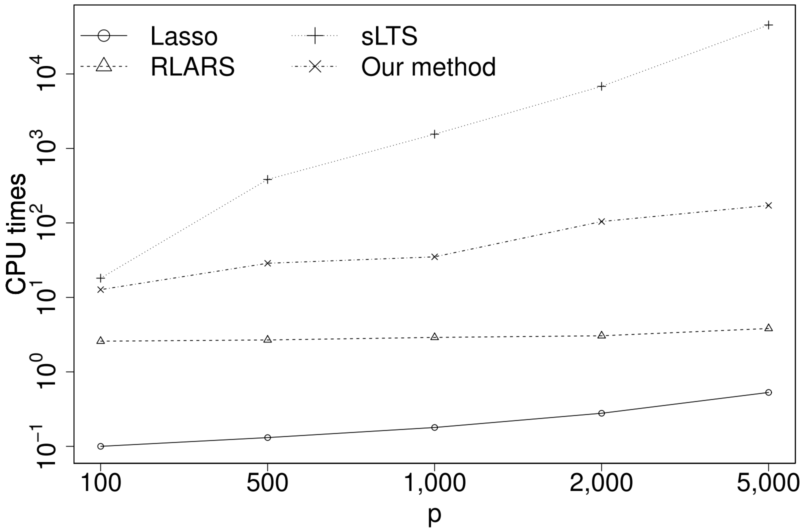

Fig. 3.1 CPU times (in seconds).

3.5 Real Data Analyses

In this section, we use two real datasets to compare our method with comparative methods in real data analysis. We show the best result of comparative methods among some parameter situations (e.g., Sect. 3.4.3).

3.5.1 NCI-60 Cancer Cell Panel

We applied our method and comparative methods to regress protein expression on gene expression data at the cancer cell panel of the National Cancer Institute. Experimental conditions were set in the same way as in Alfons, Croux, and Gelper [2] as follows. The gene expression data were obtained with an Affymetrix HG-U133A chip and the normalized GCRMAmethod, resulting in a set of $p = 22,283$ explanatory variables. The protein expressions based on 162 antibodies were acquired via reverse-phase protein lysate arrays and $\log_2$ transformed. One observation had to be removed since all values were missing in the gene expression data, reducing the number of observations to $n = 59$. Then, the KRT18 antibody was selected as the response variable because it had the largest MAD among 162 antibodies, i.e., KRT18 may include a large number of outliers. Both the protein expressions and the gene expression data can be downloaded via the web application CellMiner (http://discover.nci.nih.gov/cellminer/). As a measure of prediction performance, the root trimmed mean squared prediction error (RTMSPE) was computed via leave-one-out cross-validation given by:

$$\mathrm{RTMSPE} = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(e)^2_{[i:n]}},$$

where $e^2 = ((y_1 - x_1^T\hat{\beta}^{[-1]})^2, \dots, (y_n - x_n^T\hat{\beta}^{[-n]})^2)$ and $(e)^2_{[1:n]} \leq \cdots \leq (e)^2_{[n:n]}$ are the order statistics of e^2 and $h = \lfloor (n+1)0.75 \rfloor$. The choice of h is important because it is preferable for estimating prediction performance that trimmed squares does not include outliers. We set h in the same way as in Alfons, Croux, and Gelper [2], because the sLTS detected 13 outliers in Alfons, Croux, and Gelper [2]. In this experiment, we used the estimate of the RANSAC algorithm as an initial point instead of sLTS because sLTS required high computational cost with such high dimensional data.

Table 3.6 shows that our method outperformed other comparative methods for the RTMSPE with high dimensional data. Our method presented the smallest RTMSPE with the second smallest number of explanatory variables. RLARS presented the smallest number of explanatory variables, but a much larger RTMSPE than our method.

Table 3.6 Root trimmed mean squared prediction error (RTMSPE) for protein expressions based on the KRT18 antibody (NCI-60 cancer cell panel data), computed from leave-one-out cross-validation.

Methods	**RTMSPE**	**Selected Variables**
Lasso	1.058	52
RLARS	0.936	18
sLTS	0.721	33
Our method ($\gamma = 0.1$)	0.679	29
Our method ($\gamma = 0.5$)	0.700	30

3.5.2 Protein Homology Data

We applied our method and comparative methods to the protein sequence dataset used for KDD-Cup 2004. Experimental conditions were set in the same way as in Khan, Van Aelst, and Zamar [58] as follows. The whole dataset consists of $n = 145{,}751$ protein sequences, which has 153 blocks corresponding to native protein. Each data point in a particular block is a candidate homologous protein. There were 75 variables in the dataset: the block number (categorical) and 74 measurements of protein features. The first protein feature was used as the response variable. Then, five blocks with a total of $n = 4141$ protein sequences were selected because they contained the highest proportions of homologous proteins (and hence, the highest proportions of potential outliers). The data of each block were split into two almost equal parts to get a training sample of size $n_{tra} = 2072$ and a test sample of size $n_{test} = 2069$. The number of explanatory variables was $p = 77$, consisting of four block indicators (Variables 1–4) and 73 features. The whole protein, training and test dataset can be downloaded from http://users.ugent.be/~svaelst/software/RLARS.html. As a measure of prediction performance, the root trimmed mean squared prediction error (RTMSPE) was computed for the test sample given by:

$$\text{RTMSPE} = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(e)^2_{[i:n_{test}]}},$$

where $e^2 = ((y_1 - {x_1}^T\hat{\beta})^2, \ldots, (y_{n_{test}} - x_{n_{test}}^T\hat{\beta})^2)$ and $(e)^2_{[1:n_{test}]} \leq \cdots \leq (e)^2_{[n_{test}:n_{test}]}$ are the order statistics of e^2 and $h = \lfloor (n_{test}+1)0.99 \rfloor$, $\lfloor (n_{test}+1)0.95 \rfloor$ or $\lfloor (n_{test}+1)0.9 \rfloor$. In this experiment, we used the estimate of sLTS as an initial point.

Table 3.7 shows that our method outperformed other comparative methods for the RTMSPE. Our method presented the smallest RTMSPE with the largest number of explanatory variables. It might seem that other methods gave a smaller number of explanatory variables than necessary.

Table 3.7 Root trimmed mean squared prediction error in the protein test set.

	Trimming Fraction			
Methods	**1%**	**5%**	**10%**	**# Selected Variables**
Lasso	10.697	9.66	8.729	22
RLARS	10.473	9.435	8.527	27
sLTS	10.614	9.52	8.575	21
Our method ($\gamma = 0.1$)	10.461	9.403	8.481	44
Our method ($\gamma = 0.5$)	10.463	9.369	8.419	42

Chapter 4

Robust Generalized Linear Model with Sparsity by Stochastic Optimization

4.1 Revisiting Sparse γ-Regression

We adopt the type I of γ-divergence in this chapter. Here, we reconsider the estimation of γ-regression, which was stated in Sect. 2.1.4.

Let $f(y|x;\theta)$ be the parametric probability density function with parameter θ. The target parameter can be considered by

$$\begin{aligned}\theta^*_{\gamma,1} &= \arg\min_{\theta} D_{\gamma,1}(g(y|x), f(y|x;\theta); g(x))\\ &= \arg\min_{\theta} d_{\gamma,1}(g(y|x), f(y|x;\theta); g(x))\\ &= \arg\min_{\theta} -\frac{1}{\gamma}\log E_{g(x,y)}\left[\frac{f(y|x)^{\gamma}}{(\int f(y|x)^{1+\gamma}dy)^{\frac{\gamma}{1+\gamma}}}\right].\end{aligned}$$

Moreover, we can also consider the target parameter with a convex regularization term, given by

$$\begin{aligned}\theta^*_{\gamma_1,pen} &= \arg\min_{\theta} D_{\gamma,1}(g(y|x), f(y|x;\theta); g(x)) + \lambda P(\theta)\\ &= \arg\min_{\theta} d_{\gamma,1}(g(y|x), f(y|x;\theta); g(x)) + \lambda P(\theta)\\ &= \arg\min_{\theta} -\frac{1}{\gamma}\log E_{g(x,y)}\left[\frac{f(y|x)^{\gamma}}{(\int f(y|x)^{1+\gamma}dy)^{\frac{\gamma}{1+\gamma}}}\right] + \lambda P(\theta), \qquad (4.1)\end{aligned}$$

where $P(\theta)$ is a convex regularization term for parameter θ and λ is a tuning parameter. As an example of convex regularization term, we can consider L_1 (Lasso, [87]), elasticnet [104], the indicator function of a closed convex set [59, 24] and so on.

Let $(x_1,y_1),\ldots,(x_n,y_n)$ be the observations randomly drawn from the underlying distribution $g(x,y)$. As we have seen in Sect. 2.1.4, the γ-cross entropy, $d_{\gamma,1}(g(y|x),f(y|x;\theta);g(x))$, can be empirically estimated by

$$\bar{d}_{\gamma,1}(f(y|x;\theta)) = -\frac{1}{\gamma}\log\frac{1}{n}\sum_{i=1}^{n}\frac{f(y_i|x_i)^{\gamma}}{\left(\int f(y|x_i)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}}.$$

By virtue of (4.1), the sparse γ-estimator can be proposed by

$$\hat{\theta}_{\gamma_1,pen} = \arg\min_{\theta}\bar{d}_{\gamma,1}(f(y|x;\theta)) + \lambda P(\theta). \tag{4.2}$$

To obtain the minimizer, we solve a non-convex and non-smooth optimization problem. Iterative estimation algorithms for such a problem can not easily achieve numerical stability and efficiency.

4.1.1 MM Algorithm for Sparse γ-Regression

In Chapter 3, we proposed the iterative estimation algorithm for (3.2) by MM algorithm [52]. It has a monotone decreasing property, i.e., the objective function monotonically decreases at each iterative step, which property leads to numerical stability and efficiency. In particular, the linear regression with L_1 penalty was deeply considered.

In a similar way to in Chapter 3, the following majorization function of MM algorithm was proposed for (4.2) by using Jensen's inequality:

$$h_{MM}(\theta|\theta^{(m)}) = -\frac{1}{\gamma}\sum_{i=1}^{n}\alpha_i^{(m)}\log\left\{\frac{f(y_i|x_i;\theta)^{\gamma}}{\left(\int f(y|x_i;\theta)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}}\right\} + \lambda P(\theta), \tag{4.3}$$

where

$$\alpha_i^{(m)} = \frac{\frac{f(y_i|x_i;\theta^{(m)})^{\gamma}}{\left(\int f(y|x_i;\theta^{(m)})^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}}}{\sum_{l=1}^{n}\frac{f(y_l|x_l;\theta^{(m)})^{\gamma}}{\left(\int f(y|x_l;\theta^{(m)})^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}}}.$$

Moreover, for linear regression $y=\beta_0+x^T\beta+e$ $(e\sim N(0,\sigma^2))$ with L_1 regularization, the following majorization function and iterative estimation algorithm based on a coordinate descent method were obtained:

$$
\begin{aligned}
h_{MM,\ linear}(\theta|\theta^{(m)}) &= \frac{1}{2(1+\gamma)}\log\sigma^2 + \frac{1}{2}\sum_{i=1}^{n}\alpha_i^{(m)}\frac{(y_i-\beta_0-x_i^T\beta)^2}{\sigma^2} + \lambda||\beta||_1,\\
\beta_0^{(m+1)} &= \sum_{i=1}^{n}\alpha_i^{(m)}(y_i - x_i{}^T\beta^{(m)}),\\
\beta_j^{(m+1)} &= \frac{S\left(\sum_{i=1}^{n}\alpha_i^{(m)}(y_i-\beta_0^{(m+1)}-r_{i,-j}^{(m)})x_{ij},\ \sigma^{2(m)}\lambda\right)}{\left(\sum_{i=1}^{n}\alpha_i^{(m)}x_{ij}^2\right)} \quad (j=1,\dots,p),\\
\sigma^{2(m+1)} &= (1+\gamma)\sum_{i=1}^{n}\alpha_i^{(m)}(y_i-\beta_0^{(m+1)}-x_i^T\beta^{(m+1)})^2,
\end{aligned}
$$

where $S(t,\lambda) = \text{sign}(t)(|t|-\lambda)$ and $r_{i,-j}^{(m)} = \sum_{k\neq j} x_{ik}(\mathbb{1}_{(k<j)}\beta_k^{(m+1)} + \mathbb{1}_{(k>j)}\beta_k^{(m)})$. This iterative estimation algorithm is equal to Algorithm 1 in Sect. 3.2.2.

4.1.2 Sparse γ-Poisson Regression Case

Typical GLMs are a linear regression, logistic regression and Poisson regression: The former two regressions are easily treated with the above coordinate descent algorithm, but the Poisson regression has a problem as described in the following. Here, we consider a Poisson regression with a regularization term. Let $f(y|x;\theta)$ be the conditional density with $\theta = (\beta_0,\beta)$, given by

$$
f(y|x;\theta) = \frac{\exp(-\mu_x(\theta))}{y!}\mu_x(\theta)^y,
$$

where $\mu_x(\theta) = \mu_x(\beta_0,\beta) = \exp(\beta_0 + x^T\beta)$. By virtue of (4.3), we can obtain the majorization function for Poisson regression with a regularization term, given by

$$
\begin{aligned}
h_{MM,\ poisson}(\theta|\theta^{(m)}) &= -\sum_{i=1}^{n}\alpha_i^{(m)}\log\frac{\exp(-\mu_{x_i}(\theta))}{y_i!}\mu_{x_i}(\theta)^{y_i}\\
&+\frac{1}{1+\gamma}\sum_{i=1}^{n}\alpha_i^{(m)}\log\left\{\sum_{y=0}^{\infty}\frac{\exp(-(1+\gamma)\mu_{x_i}(\theta))}{y!^{1+\gamma}}\mu_{x_i}(\theta)^{(1+\gamma)y}\right\} + \lambda P(\theta). \qquad (4.4)
\end{aligned}
$$

The second term of the right hand side in (4.4) contains the hypergeometric series, and then we cannot obtain a closed form on the MM algorithm with respect to the parameters β_0,β although this series converges (see Sect. 4.3.3). Therefore, we cannot derive an efficient iterative estimation algorithm based on a coordinate descent method in a similar way to in Chapter 3. Other sparse optimization methods which use a linear approximation on the loss function, e.g., proximal gradient descent [70, 25, 6], can solve (4.4). However, these methods require at least sample size n times of an approximate calculation for the hypergeometric series at each iterative step in sub-problem $\arg\min_\theta h_{MM}(\theta|\theta^{(m)})$. Therefore, it requires high computation cost, especially for very large problems. We need another optimization approach to overcome such problems. In this paper, we consider minimizing the regularized expected risk (4.1) directly by a stochastic optimization approach. In what follows, we refer to the sparse γ-regression in the GLM as the sparse γ-GLM.

4.2 Stochastic Optimization Approach for Regularized Expected Risk Minimization

The regularized expected risk minimization is generally the following form:

$$\Psi^* := \min_{\theta\in\Theta} \left\{ \Psi(\theta) := E_{(x,y)}\left[l((x,y);\theta)\right] + \lambda P(\theta) \right\}, \tag{4.5}$$

where Θ is a closed convex set in $\mathbb{R}^n$, l is a loss function with a parameter θ and $\Psi(\theta)$ is bounded below over Θ by $\Psi^* > -\infty$. Stochastic optimization approach solves (4.5) sequentially. More specifically, we draw a sequence of i.i.d. paired samples $(x_1,y_1),(x_2,y_2),\ldots$ $,(x_t,y_t),\ldots$ and, at t-th time, update the parameter $\theta^{(t)}$ based on the latest paired sample (x_t,y_t) and the previous updated parameter $\theta^{(t-1)}$. Therefore, it requires low computational complexity per iteration and stochastic optimization can scale well for very large problems.

4.2.1 Stochastic Gradient Descent

The stochastic gradient descent (SGD) is one of popular stochastic optimization approaches and is widely used in machine learning community [12]. The SGD takes the form

$$\theta^{(t+1)} = \arg\min_{\theta\in\Theta} \left\langle \nabla l((x_t,y_t);\theta^{(t)}),\theta \right\rangle + \lambda P(\theta) + \frac{1}{2\eta_t}\|\theta-\theta^{(t)}\|_2^2, \tag{4.6}$$

where η_t is a step size parameter. For some important examples of $P(\theta)$, e.g., L_1 regularization, (4.6) can be solved in a closed form.

When a loss function l is convex (possibly non-differentiable) and η_t is set to be appropriate, e.g., $\eta_t = O\left(\frac{1}{\sqrt{t}}\right)$, under mild conditions, the convergence property was established for the average of the iterates, i.e., $\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^{T}\theta^{(t)}$ as follows (see, e.g., [14]):

$$E\left[\Psi(\bar{\theta}_T)\right] - \Psi^* \leq O\left(\frac{1}{\sqrt{T}}\right),$$

where the expectation is taken with respect to past paired samples $(x_t, y_t)\ldots(x_T, y_T)$. Moreover, for some variants of SGD, e.g., RDA [96], Mirror descent [23], Adagrad [22], the convergence property was established under similar assumptions.

These methods assume that a loss function is convex to establish the convergence property, but the loss function is non-convex in our problem (4.1). Then, we cannot adopt these methods directly. Recently, for non-convex loss function with convex regularization term, randomized stochastic projected gradient (RSPG) was proposed by [38]. Under mild conditions, the convergence property was established. Therefore, we consider applying the RSPG to our problem (4.1).

4.2.2 Randomized Stochastic Projected Gradient

First, we explain the RSPG, following [38]. The RSPG takes the form

$$\theta^{(t+1)} = \arg\min_{\theta\in\Theta}\left\langle \frac{1}{m_t}\sum_{i=1}^{m_t}\nabla l((x_{t,i}, y_{t,i}); \theta^{(t)}), \theta \right\rangle + \lambda P(\theta) + \frac{1}{\eta_t}V(\theta, \theta^{(t)}), \tag{4.7}$$

where m_t is the size of mini-batch at t-th time, $(x_{t,i}, y_{t,i})$ is the i-th mini-batch sample at t-th time and

$$V(a,b) = w(a) - w(b) - \langle \nabla w(b), a-b \rangle,$$

where w is continuously differentiable and α-strongly convex function satisfying $\langle a - b, \nabla w(a) - \nabla w(b)\rangle \geq \alpha\|a-b\|^2$ for $a, b \in \Theta$. When $w(\theta) = \frac{1}{2}||\theta||_2^2$, i.e., $V(\theta, \theta^{(t)}) = \frac{1}{2}||\theta - \theta^{(t)}||_2^2$, (4.7) is almost equal to (4.6).

Here, we denote two remarks on RSPG as a difference from the SGD. One is that the RSPG uses the mini-batch strategy, i.e., taking multiple samples at t-th time. The other is that the RSPG randomly selects the output $\hat{\theta}$ from $\left\{\theta^{(1)}, \ldots, \theta^{(T)}\right\}$ according to a certain probability distribution instead of taking the average of the iterates. This is because for non-convex stochastic optimization, later iterates does not always gather around local minimum and the average of the iterates cannot work in such a convex case.

Next, we show the implementation of the RSPG, given by Algorithm2. However, Algorithm 2 has a large deviation of the output because the only one final output is selected

Algorithm 2 Randomized stochastic projected gradient

```
Input: The initial point $\theta^{(1)}$, the step size $\eta_t$, the mini-batch size $m_t$, the iteration limit $T$
  and the probability mass function $P_R$ supported on $\{1,\dots,T\}$.
  Let $R$ be a random variable generated by probability mass function $P_R$.
  for $t = 1,\dots,R$ do
      $\theta^{(t+1)} = \arg\min_{\theta\in\Theta} \left\langle \frac{1}{m_t}\sum_{i=1}^{m_t} \nabla l((x_{t,i},y_{t,i});\theta^{(t)}),\theta \right\rangle + \lambda P(\theta) + \frac{1}{\eta_t} V(\theta,\theta^{(t)})$.
Output: $\theta^{(R)}$.
```

via some probability mass function P_R. Therefore, [38] also proposed the two phase RSPG (2-RSPG) which has the post-optimization phase. In the post-optimization phase, multiple outputs are selected and these are validated to determine the final output, as shown in Algorithm 3. This can be expected to achieve a better complexity result of finding an $(\varepsilon,\Lambda)-solution$, i.e., $\mathrm{Prob}\left\{C(\theta^{(R)}) \le \varepsilon\right\} \ge 1-\Lambda$, where C is some convergence criterion, for some $\varepsilon > 0$ and $\Lambda \in (0,1)$. For more detailed descriptions and proofs, we refer to the Sect.4 in [38].

Algorithm 3 Two phase randomized stochastic projected gradient

```
Input: The initial point $\theta^{(1)}$, the step size $\eta_t$, the mini-batch size $m_t$, the iteration limit $T$,
  the probability mass function $P_R$ supported on $\{1,\dots,T\}$, the number of candidates $N_{cand}$
  and the sample size $N_{post}$ for validation.
  Let $R_1, R_2, \dots, R_{N_{cand}}$ be random variables generated by probability mass function $P_R$.
  for $t = 1,\dots,\max\{R_1,R_2,\dots,R_{N_{cand}}\}$ do
      $\theta^{(t+1)} = \arg\min_{\theta\in\Theta} \left\langle \frac{1}{m_t}\sum_{i=1}^{m_t} \nabla l((x_{t,i},y_{t,i});\theta^{(t)}),\theta \right\rangle + \lambda P(\theta) + \frac{1}{\eta_t} V(\theta,\theta^{(t)})$.
   Post-optimization phase:
   $\theta^{(R_s)} = \arg\min_{s=1,\dots,N_{cand}} \frac{1}{\eta_{R_s}} \|\theta^{(R_s)} - \theta^{\left(R_s^+\right)}\|$,
  where $\theta^{\left(R_s^+\right)} = \arg\min_{\theta\in\Theta} \left\langle \frac{1}{N_{post}}\sum_{i=1}^{N_{post}} \nabla l((x_i,y_i);\theta^{(R_s)}),\theta \right\rangle + \lambda P(\theta) + \frac{1}{\eta_{R_s}} V(\theta,\theta^{(R_s)})$.
Output: $\theta^{(R_s)}$.
```

4.3 Online Robust GLM with Sparsity

In this section, we show the sparse γ-GLM with the stochastic optimization approach on three specific examples; linear regression, logistic regression and Poisson regression with L_1 regularization. In what follows, we refer to the sparse γ-GLM with the stochastic optimization approach as the online sparse γ-GLM.

In order to apply the RSPG to our methods (4.1), we prepare the monotone transformation of the γ-cross entropy for regression in (4.1) as follows

$$\underset{\theta\in\Theta}{\arg\min} E_{g(x,y)}\left[-\frac{f(y|x;\theta)^{\gamma}}{(\int f(y|x;\theta)^{1+\gamma}dy)^{\frac{\gamma}{1+\gamma}}}\right]+\lambda P(\theta), \tag{4.8}$$

and we suppose that Θ is $\mathbb{R}^n$ or closed ball with sufficiently large radius. Then, we can apply the RSPG to (4.8) and by virtue of (4.7), the update formula takes the form

$$\theta^{(t+1)}=\underset{\theta\in\Theta}{\arg\min}\left\langle -\frac{1}{m_t}\sum_{i=1}^{m_t}\nabla\frac{f(y_{t,i}|x_{t,i};\theta^{(t)})^{\gamma}}{\left(\int f(y|x_{t,i};\theta^{(t)})^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}},\theta\right\rangle+\lambda P(\theta)+\frac{1}{\eta_t}V(\theta,\theta^{(t)}). \tag{4.9}$$

More specifically, we suppose that $V(\theta,\theta^{(t)})=\frac{1}{2}||\theta-\theta^{(t)}||_2^2$ because the update formula can be obtained in closed form for some important sparse regularization terms, e.g., L_1 regularization, elasticnet. We illustrate the update algorithms based on Algorithm 2 for three specific examples. The update algorithms based on Algorithm 3 are obtained in a similar manner.

In order to implement our methods, we need to determine some tuning parameters, e.g., the step size η_t, mini-batch size m_t. In Sect. 4.4, we discuss how to determine some tuning parameters in detail.

4.3.1 Online Sparse γ-Linear Regression

Let $f(y|x;\theta)$ be the conditional density with $\theta=(\beta_0,\beta^T,\sigma^2)^T$, given by

$$f(y|x;\theta)=\phi(y;\beta_0+x^T\beta,\sigma^2),$$

where $\phi(y;\mu,\sigma^2)$ is the normal density with mean parameter μ and variance parameter σ^2. Suppose that $P(\theta)$ is the L_1 regularization $||\beta||_1$. Then, by virtue of (4.9), we can obtain the update formula given by

$$\begin{aligned}&\left(\beta_0^{(t+1)},\beta^{(t+1)},\sigma^{2^{(t+1)}}\right)\\&=\underset{\beta_0,\beta,\sigma^2}{\arg\min}\,\xi_1(\beta_0^{(t)})\beta_0+\langle\xi_2(\beta^{(t)}),\beta\rangle+\xi_3(\sigma^{2^{(t)}})\sigma^2\\&\quad+\lambda||\beta||_1+\frac{1}{2\eta_t}||\beta_0-\beta_0^{(t)}||_2^2+\frac{1}{2\eta_t}||\beta-\beta^{(t)}||_2^2+\frac{1}{2\eta_t}||\sigma^2-\sigma^{2^{(t)}}||_2^2,\end{aligned} \tag{4.10}$$

where

$$\xi_1(\beta_0^{(t)}) = -\frac{1}{m_t}\sum_{i=1}^{m_t}\left[\frac{\gamma(y_{t,i}-\beta_0^{(t)}-x_{t,i}{}^T\beta^{(t)})}{\sigma^{2(t)}}\left(\frac{1+\gamma}{2\pi\sigma^{2(t)}}\right)^{\frac{\gamma}{2(1+\gamma)}}\right.$$
$$\left.\exp\left\{-\frac{\gamma(y_{t,i}-\beta_0^{(t)}-x_{t,i}{}^T\beta^{(t)})^2}{2\sigma^{2(t)}}\right\}\right],$$
$$\xi_2(\beta^{(t)}) = -\frac{1}{m_t}\sum_{i=1}^{m_t}\left[\frac{\gamma(y_{t,i}-\beta_0^{(t)}-x_{t,i}{}^T\beta^{(t)})}{\sigma^{2(t)}}\left(\frac{1+\gamma}{2\pi\sigma^{2(t)}}\right)^{\frac{\gamma}{2(1+\gamma)}}\right.$$
$$\left.\exp\left\{-\frac{\gamma(y_{t,i}-\beta_0^{(t)}-x_{t,i}{}^T\beta^{(t)})^2}{2\sigma^{2(t)}}\right\}x_{t,i}\right],$$
$$\xi_3(\sigma^{2(t)}) = \frac{1}{m_t}\sum_{i=1}^{m_t}\left[\frac{\gamma}{2}\left(\frac{1+\gamma}{2\pi\sigma^{2(t)}}\right)^{\frac{\gamma}{2(1+\gamma)}}\left\{\frac{1}{(1+\gamma)\sigma^{2(t)}}-\frac{(y_{t,i}-\beta_0^{(t)}-x_{t,i}{}^T\beta^{(t)})^2}{\sigma^{4(t)}}\right\}\right.$$
$$\left.\exp\left\{-\frac{\gamma(y_{t,i}-\beta_0^{(t)}-x_{t,i}{}^T\beta^{(t)})^2}{2\sigma^{2(t)}}\right\}\right].$$

Consequently, we can obtain the update algorithm, as shown in Algorithm 4.

```
Algorithm 4 Online sparse $\gamma$-linear regression
Input: The initial points $\beta_0^{(1)}$, $\beta^{(1)}$, $\sigma^{2(1)}$, the step size $\eta_t$, the mini-batch size $m_t$, the
    iteration limit $T$ and the probability mass function $P_R$ supported on $\{1,\ldots,T\}$.
    Let $R$ be a random variable generated by probability mass function $P_R$.
    for $t = 1,\ldots,R$ do
        $\beta_0^{(t+1)} = \beta_0^{(t)} - \eta_t\xi_1(\beta_0^{(t)})$.
        $\beta_j^{(t+1)} = S(\beta_j^{(t)} - \eta_t\xi_{2j}(\beta^{(t)}), \eta_t\lambda)\ \ (j = 1,\ldots,p)$.
        $\sigma^{2(t+1)} = \sigma^{2(t)} - \eta_t\xi_3(\sigma^{2(t)})$.
Output: $\beta_0^{(R)}$, $\beta^{(R)}$, $\sigma^{2(R)}$.
```

Here, we briefly show the robustness of online sparse γ-linear regression. For simplicity, we consider the intercept parameter β_0. Suppose that the $(x_{t,k}, y_{t,k})$ is an outlier at t-th time. The conditional probability density $f(y_{t,k}|x_{t,k};\theta^{(t)})$ can be expected to be sufficiently small.

We see from $f(y_{t,k}|x_{t,k};\theta^{(t)}) \approx 0$ and (4.10) that

$$
\begin{aligned}
&\beta_0^{(t+1)}\\
&= \underset{\beta_0}{\arg\min} -\frac{1}{m_t}\sum_{1\le i\ne k\le m_t}\left[\frac{\gamma(y_{t,i}-\beta_0^{(t)}-x_{t,i}{}^T\beta^{(t)})}{\sigma^{2(t)}}\left(\frac{1+\gamma}{2\pi\sigma^{2(t)}}\right)^{\frac{\gamma}{2(1+\gamma)}}\right.\\
&\qquad\qquad \left.\exp\left\{-\frac{\gamma(y_{t,i}-\beta_0^{(t)}-x_{t,i}{}^T\beta^{(t)})^2}{2\sigma^{2(t)}}\right\}\right]\times\beta_0\\
&\quad -\frac{1}{m_t}\frac{\gamma(y_{t,k}-\beta_0^{(t)}-x_{t,k}{}^T\beta^{(t)})}{\sigma^{2(t)}}\left(\frac{1+\gamma}{2\pi\sigma^{2(t)}}\right)^{\frac{\gamma}{2(1+\gamma)}}\exp\left\{-\frac{\gamma(y_{t,k}-\beta_0^{(t)}-x_{t,k}{}^T\beta^{(t)})^2}{2\sigma^{2(t)}}\right\}\times\beta_0\\
&\qquad\qquad +\frac{1}{2\eta_t}\|\beta_0-\beta_0^{(t)}\|_2^2\\
&= \underset{\beta_0}{\arg\min} -\frac{1}{m_t}\sum_{1\le i\ne k\le m_t}\left[\frac{\gamma(y_{t,i}-\beta_0^{(t)}-x_{t,i}{}^T\beta^{(t)})}{\sigma^{2(t)}}\left(\frac{1+\gamma}{2\pi\sigma^{2(t)}}\right)^{\frac{\gamma}{2(1+\gamma)}}\right.\\
&\qquad\qquad \left.\exp\left\{-\frac{\gamma(y_{t,i}-\beta_0^{(t)}-x_{t,i}{}^T\beta^{(t)})^2}{2\sigma^{2(t)}}\right\}\right]\times\beta_0\\
&\quad \underbrace{-\frac{1}{m_t}\frac{\gamma(1+\gamma)^{\frac{\gamma}{2(1+\gamma)}}(y_{t,k}-\beta_0^{(t)}-x_{t,k}{}^T\beta^{(t)})}{\sigma^{2(t)}}\left(2\pi\sigma^{2(t)}\right)^{\frac{\gamma^2}{2(1+\gamma)}} f(y_{t,k}|x_{t,k};\theta^{(t)})^{\gamma}\times\beta_0}_{\approx 0}\\
&\qquad +\frac{1}{2\eta_t}\|\beta_0-\beta_0^{(t)}\|_2^2.
\end{aligned}
$$

Therefore, the effect of an outlier is naturally ignored in (4.10). Similarly, we can also see the robustness for parameters β and σ^2.

4.3.2 Online Sparse γ-Logistic Regression

Let $f(y|x;\theta)$ be the conditional density with $\theta = (\beta_0, \beta^T)^T$, given by

$$
f(y|x;\beta_0,\beta) = F(\tilde{x}^T\theta)^y(1-F(\tilde{x}^T\theta))^{(1-y)},
$$

where $\tilde{x} = (1, x^T)^T$ and $F(u) = \frac{1}{1+\exp(-u)}$. Then, by virtue of (4.9), we can obtain the update formula given by

$$\begin{aligned}&\left(\beta_0^{(t+1)}, \beta^{(t+1)}\right)\\ &= \arg\min_{\beta_0,\beta} \nu_1(\beta_0^{(t)})\beta_0 + \langle \nu_2(\beta^{(t)}), \beta\rangle + \lambda||\beta||_1 + \frac{1}{2\eta_t}||\beta_0 - \beta_0^{(t)}||_2^2 + \frac{1}{2\eta_t}||\beta - \beta^{(t)}||_2^2,\end{aligned} \tag{4.11}$$

where

$$\nu_1(\beta_0^{(t)}) = -\frac{1}{m_t}\sum_{i=1}^{m_t}\left[\frac{\gamma \exp(\gamma y_{t,i}\tilde{x}_{t,i}^T\theta^{(t)})\left\{y_{t,i} - \frac{\exp((1+\gamma)\tilde{x}_{t,i}^T\theta^{(t)})}{1+\exp((1+\gamma)\tilde{x}_{t,i}^T\theta^{(t)})}\right\}}{\left\{1+\exp((1+\gamma)\tilde{x}_{t,i}^T\theta^{(t)})\right\}^{\frac{\gamma}{1+\gamma}}}\right],$$

$$\nu_2(\beta^{(t)}) = -\frac{1}{m_t}\sum_{i=1}^{m_t}\left[\frac{\gamma \exp(\gamma y_{t,i}\tilde{x}_{t,i}^T\theta^{(t)})\left\{y_{t,i} - \frac{\exp((1+\gamma)\tilde{x}_{t,i}^T\theta^{(t)})}{1+\exp((1+\gamma)\tilde{x}_{t,i}^T\theta^{(t)})}\right\}}{\left\{1+\exp((1+\gamma)\tilde{x}_{t,i}^T\theta^{(t)})\right\}^{\frac{\gamma}{1+\gamma}}} x_{t,i}\right].$$

Consequently, we can obtain the update algorithm as shown in Algorithm 5. In a similar way to online sparse γ-linear regression, we can also see the robustness for parameters β_0 and β in online sparse γ-logistic regression (4.11).

Algorithm 5 Online sparse γ-logistic regression

```
Input: The initial points $\beta_0^{(1)}$, $\beta^{(1)}$, the step size $\eta_t$, the mini-batch size $m_t$, the iteration
  limit $T$ and the probability mass function $P_R$ supported on $\{1,\ldots,T\}$.
  Let $R$ be a random variable generated by probability mass function $P_R$.
  for $t = 1,\ldots,R$ do
    $\beta_0^{(t+1)} = \beta_0^{(t)} - \eta_t \nu_1(\beta_0^{(t)})$.
    $\beta_j^{(t+1)} = S(\beta_j^{(t)} - \eta_t \nu_{2j}(\beta^{(t)}), \eta_t\lambda)$ $(j = 1,\ldots,p)$.
Output: $\beta_0^{(R)}$, $\beta^{(R)}$.
```

4.3.3 Online Sparse γ-Poisson Regression

Let $f(y|x;\theta)$ be the conditional density with $\theta = (\beta_0, \beta^T)^T$, given by

$$f(y|x;\theta) = \frac{\exp(-\mu_x(\theta))}{y!}\mu_x(\theta)^y,$$

where $\mu_x(\theta) = \mu_x(\beta_0, \beta) = \exp(\beta_0 + x^T\beta)$. Then, by virtue of (4.9), we can obtain the update formula given by

$$\begin{aligned}
&\left(\beta_0^{(t+1)}, \beta^{(t+1)}\right) \\
&= \arg\min_{\beta_0,\beta} \zeta_1(\beta_0^{(t)})\beta_0 + \langle \zeta_2(\beta^{(t)}), \beta\rangle + \lambda||\beta||_1 + \frac{1}{2\eta_t}||\beta_0 - \beta_0^{(t)}||_2^2 + \frac{1}{2\eta_t}||\beta - \beta^{(t)}||_2^2,
\end{aligned} \tag{4.12}$$

where

$$\zeta_1(\beta_0^{(t)}) = \frac{1}{m_t}\sum_{i=1}^{m_t}\left[\frac{\gamma f(y_{t,i}|x_{t,i};\theta^{(t)})^\gamma \left\{\sum_{y=0}^{\infty}(y - y_{t,i}) f(y|x_{t,i};\theta^{(t)})^{1+\gamma}\right\}}{\left\{\sum_{y=0}^{\infty} f(y|x_{t,i};\theta^{(t)})^{1+\gamma}\right\}^{\frac{1+2\gamma}{1+\gamma}}}\right],$$

$$\zeta_2(\beta^{(t)}) = \frac{1}{m_t}\sum_{i=1}^{m_t}\left[\frac{\gamma f(y_{t,i}|x_{t,i};\theta^{(t)})^\gamma \left\{\sum_{y=0}^{\infty}(y - y_{t,i}) f(y|x_{t,i};\theta^{(t)})^{1+\gamma}\right\}}{\left\{\sum_{y=0}^{\infty} f(y|x_{t,i};\theta^{(t)})^{1+\gamma}\right\}^{\frac{1+2\gamma}{1+\gamma}}} x_{t,i}\right].$$

In (4.12), two types hypergeometric series exist. Here, we prove a convergence of $\sum_{y=0}^{\infty} f(y|x_{t,i};\theta^{(t)})^{1+\gamma}$ and $\sum_{y=0}^{\infty}(y - y_{t,i}) f(y|x_{t,i};\theta^{(t)})^{1+\gamma}$. First, let us consider $\sum_{y=0}^{\infty} f(y|x_{t,i};\theta^{(t)})^{1+\gamma}$ and we denote n-th term that $S_n = f(n|x_{t,i};\theta^{(t)})^{1+\gamma}$. Then, we use the dalembert ratio test for S_n:

$$\begin{aligned}
&\lim_{n\to\infty}\left|\frac{S_{n+1}}{S_n}\right| \\
&= \lim_{n\to\infty}\left|\frac{f(n+1|x_{t,i};\theta^{(t)})^{1+\gamma}}{f(n|x_{t,i};\theta^{(t)})^{1+\gamma}}\right| \\
&= \lim_{n\to\infty}\left|\frac{\frac{\exp(-\mu_{x_{t,i}}(\beta_0^{(t)},\beta^{(t)}))}{n+1!}\mu_{x_{t,i}}(\beta_0^{(t)},\beta^{(t)})^{n+1}}{\frac{\exp(-\mu_{x_{t,i}}(\beta_0^{(t)},\beta^{(t)}))}{n!}\mu_{x_{t,i}}(\beta_0^{(t)},\beta^{(t)})^{n}}\right|^{1+\gamma} \\
&= \lim_{n\to\infty}\left|\frac{\mu_{x_{t,i}}(\beta_0^{(t)},\beta^{(t)})}{n+1}\right|^{1+\gamma}
\end{aligned}$$

If the term $\mu_{x_{t,i}}(\beta_0^{(t)},\beta^{(t)})$ is bounded,

$= 0.$

Therefore, $\sum_{y=0}^{\infty} f(y|x_{t,i};\theta^{(t)})^{1+\gamma}$ converges.

Next, let us consider $\sum_{y=0}^{\infty}(y-y_{t,i})f(y|x_{t,i};\theta^{(t)})^{1+\gamma}$ and we denote n-th term that $S_n^{'} = (n-y_{t,i})f(n|x_{t,i};\theta^{(t)})^{1+\gamma}$. Then, we use the dalembert ratio test for $S_n^{'}$:

$$
\begin{aligned}
&\lim_{n\to\infty}\left|\frac{S_{n+1}^{'}}{S_n^{'}}\right| \\
&= \lim_{n\to\infty}\left|\frac{(1+\frac{1}{n}-\frac{y_{t,i}}{n})f(n+1|x_{t,i};\theta^{(t)})^{1+\gamma}}{(1-\frac{y_{t,i}}{n})f(n|x_{t,i};\theta^{(t)})^{1+\gamma}}\right| \\
&= \lim_{n\to\infty}\left|\frac{(1+\frac{1}{n}-\frac{y_{t,i}}{n})}{(1-\frac{y_{t,i}}{n})}\right|\left|\frac{f(n+1|x_{t,i};\theta^{(t)})^{1+\gamma}}{f(n|x_{t,i};\theta^{(t)})^{1+\gamma}}\right| \\
&= 0.
\end{aligned}
$$

Therefore, $\sum_{y=0}^{\infty}(y-y_{t,i})f(y|x_{t,i};\theta^{(t)})^{1+\gamma}$ converges.

Consequently, we can obtain the update algorithm as shown in Algorithm 6. In a similar way to online sparse γ-linear regression, we can also see the robustness for parameters β_0 and β in online sparse γ-Poisson regression (4.12). Moreover, this update algorithm requires at most twice sample size $2n = 2\times\sum_{t=1}^{T} m_t$ times of an approximate calculation for the hypergeometric series in Algorithm 6. Therefore, we can achieve a significant reduction in computational complexity.

Algorithm 6 Online sparse γ-Poisson regression

```
Input: The initial points $\beta_0^{(1)}$, $\beta^{(1)}$, the step size $\eta_t$, the mini-batch size $m_t$, the iteration
  limit $T$ and the probability mass function $P_R$ supported on $\{1,\dots,T\}$.
  Let $R$ be a random variable generated by probability mass function $P_R$.
  for $t = 1,\dots,R$ do
      $\beta_0^{(t+1)} = \beta_0^{(t)} - \eta_t\zeta_1(\beta_0^{(t)})$.
      $\beta_j^{(t+1)} = S(\beta_j^{(t)} - \eta_t\zeta_{2j}(\beta^{(t)}), \eta_t\lambda)\ \ (j = 1,\dots,p)$.
Output: $\beta_0^{(R)}$, $\beta^{(R)}$.
```

4.4 Convergence Property of Online Sparse γ-GLM

In this section, we show the global convergence property of the RSPG established by [38]. Moreover, we extend it to the classical first-order necessary condition, i.e., at a local minimum, the directional derivative, if it exists, is non-negative for any direction (see, e.g., [11]).

First, we show the global convergence property of the RSPG. In order to apply to online sparse γ-GLM, we slightly modify some notations. We consider the following optimization problem (4.5) again:

$$\Psi^* := \min_{\theta\in\Theta} \underbrace{E_{(x,y)}\left[l((x,y);\theta)\right] + \lambda P(\theta)}_{:=\Psi(\theta)},$$

where $E_{(x,y)}\left[l((x,y);\theta)\right]$ is continuously differentiable and possibly non-convex. The update formula (4.7) of the RSPG is as follows:

$$\theta^{(t+1)} = \arg\min_{\theta\in\Theta} \left\langle \frac{1}{m_t}\sum_{i=1}^{m_t} \nabla l((x_{t,i},y_{t,i});\theta^{(t)}), \theta \right\rangle + \lambda P(\theta) + \frac{1}{\eta_t} V(\theta,\theta^{(t)}),$$

where

$$V(a,b) = w(a) - w(b) - \langle \nabla w(b), a-b\rangle,$$

and w is continuously differentiable and α-strongly convex function satisfying $\langle a-b, \nabla w(a) - \nabla w(b)\rangle \geq \alpha \|a-b\|^2$ for $a,b\in\Theta$. We make the following assumptions.

Assumption 1 $\nabla E_{(x,y)}\left[l((x,y);\theta)\right]$ is L-Lipschitz continuous for some $L>0$, i.e.,

$$\|\nabla E_{(x,y)}\left[l((x,y);\theta_1)\right] - \nabla E_{(x,y)}\left[l((x,y);\theta_2)\right]\| < L\|\theta_1-\theta_2\|, \text{ for any } \theta_1,\theta_2 \in \Theta. \quad (4.13)$$

Assumption 2 For any $t\geq 1$,

$$E_{(x_t,y_t)}\left[\nabla l((x_t,y_t);\theta^{(t)})\right] = \nabla E_{(x_t,y_t)}\left[l((x_t,y_t);\theta^{(t)})\right], \quad (4.14)$$

$$E_{(x_t,y_t)}\left[\left\|\nabla l((x_t,y_t);\theta^{(t)}) - \nabla E_{(x_t,y_t)}\left[l((x_t,y_t);\theta^{(t)})\right]\right\|^2\right] \leq \tau^2, \quad (4.15)$$

where $\tau>0$ is a constant.

Let us define

$$P_{X,R} = \frac{1}{\eta_R}\left(\theta^{(R)} - \theta^+\right),$$

$$\tilde{P}_{X,R} = \frac{1}{\eta_R}\left(\theta^{(R)} - \tilde{\theta}^+\right),$$

where

$$\theta^{+} = \mathop{\arg\min}_{\theta\in\Theta} \left\langle \nabla E_{(x,y)}\left[l((x,y);\theta^{(R)})\right], \theta \right\rangle + \lambda P(\theta) + \frac{1}{\eta_R} V(\theta, \theta^{(R)}), \tag{4.16}$$

$$\tilde{\theta}^{+} = \mathop{\arg\min}_{\theta\in\Theta} \left\langle \frac{1}{m_R}\sum_{i=1}^{m_R} \nabla l((x_{R,i}, y_{R,i});\theta^{(R)}), \theta \right\rangle + \lambda P(\theta) + \frac{1}{\eta_R} V(\theta, \theta^{(R)}).$$

Then, the following global convergence property was obtained.

Theorem 4.4.1. [Global Convergence Property in [38]]
Suppose that the step sizes $\{\eta_t\}$ are chosen such that $0 < \eta_t \leq \frac{\alpha}{L}$ with $\eta_t < \frac{\alpha}{L}$ for at least one t, and the probability mass function P_R is chosen such that for any $t = 1,\dots,T$,

$$P_R(t) := Prob\{R = t\} = \frac{\alpha\eta_t - L\eta_t^2}{\sum_{t=1}^{T}\left(\alpha\eta_t - L\eta_t^2\right)}. \tag{4.17}$$

Then, we have

$$E\left[||\tilde{P}_{X,R}||^2\right] \leq \frac{LD_\Psi^2 + \left(\tau^2/\alpha\right)\sum_{t=1}^{T}\left(\eta_t/m_t\right)}{\sum_{t=1}^{T}\left(\alpha\eta_t - L\eta_t^2\right)},$$

where the expectation was taken with respect to R and past samples $(x_{t,i}, y_{t,i})$ $(t = 1,\dots,T;\ i = 1,\dots,m_t)$ and $D_\Psi = \left[\frac{\Psi(\theta^{(1)})-\Psi^*}{L}\right]^{\frac{1}{2}}$.

Proof. See [38], Theorem 2. □

In particular, [38] investigated the constant step size and mini-batch size policy as follows.

Corollary 4.4.1. [Global Convergence Property with constant step size and mini-batch size in [38]]
Suppose that the step sizes and mini-batch sizes are $\eta_t = \frac{\alpha}{2L}$ and $m_t = m$ (≥ 1) for all $t = 1,\dots,T$, and the probability mass function P_R is chosen as (4.17). Then, we have

$$E\left[||\tilde{P}_{X,R}||^2\right] \leq \frac{4L^2D_\Psi^2}{\alpha^2 T} + \frac{2\tau^2}{\alpha^2 m} \quad \textit{and} \quad E\left[||P_{X,R}||^2\right] \leq \frac{8L^2D_\Psi^2}{\alpha^2 T} + \frac{6\tau^2}{\alpha^2 m}.$$

Moreover, the appropriate choice of mini-batch size m is given by

$$m = \left\lceil \min\left\{ \max\left\{ 1, \frac{\tau\sqrt{6N}}{4L\tilde{D}} \right\}, N \right\} \right\rceil,$$

where $\tilde{D} > 0$ and $N\,(= m \times T)$ is the number of total samples. Then, with the above setting, we have the following result

$$\frac{\alpha^2}{L} E\left[\|P_{X,R}\|^2\right] \leq \frac{16LD_\Psi^2}{N} + \frac{4\sqrt{6}\tau}{\sqrt{N}}\left(\frac{D_\Psi^2}{\tilde{D}} + \tilde{D}\max\left\{1, \frac{\sqrt{6}\tau}{4L\tilde{D}\sqrt{N}}\right\}\right). \tag{4.18}$$

Furthermore, when N is relatively large, the optimal choice of $\tilde{D}$ would be D_Ψ and (4.18) reduces to

$$\frac{\alpha^2}{L} E\left[\|P_{X,R}\|^2\right] \leq \frac{16LD_\Psi^2}{N} + \frac{8\sqrt{6}D_\Psi\tau}{\sqrt{N}}.$$

Proof. See [38], Corollary 4. □

Finally, we extend (4.18) to the classical first-order necessary condition as follows

Theorem 4.4.2. [The Modified Global Convergence Property]
Under the same assumptions in Theorem 4.4.1, we can expect $P_{X,R} \approx 0$ with high probability from (4.18) and Markov inequality. Then, for any direction δ and $\theta^{(R)} \in$ relint(Θ), we have

$$\Psi'(\theta^{(R)};\delta) = \lim_{k\downarrow 0}\frac{\Psi(\theta^{(R)}+k\delta)-\Psi(\theta^{(R)})}{k} \geq 0 \text{ with high probability.} \tag{4.19}$$

The proof is in Appendix C. Under the above assumptions and results, online sparse γ-GLM has the global convergence property. Therefore, we adopted the following parameter setting in online sparse γ-GLM:

$$\begin{aligned}\text{step size: } \eta_t &= \frac{1}{2L},\\ \text{mini-batch size: } m_t &= \left\lceil \min\left\{\max\left\{1, \frac{\tau\sqrt{6N}}{4L\tilde{D}}\right\}, N\right\}\right\rceil.\end{aligned}$$

More specifically, when the (approximate) minimum value of the objective function Ψ^* is known, e.g., the objective function is non-negative, we should use D_Ψ instead of $\tilde{D}$. In numerical experiment, we used the D_Ψ because we can obtain Ψ^* in advance. In real data analysis, we cannot obtain Ψ^* in advance. Then, we used the some values of $\tilde{D}$, i.e., the some values of mini-batch size m_t.

4.5 Numerical Experiment

In this section, we present the numerical results of online sparse γ-linear regression. We compared online sparse γ-linear regression based on the RSPG with online sparse γ-linear regression based on the SGD, which does not guarantee convergence for non-convex case. The RSPG has two variants, which are shown in Algorithms 2 and 3. In this experiment, we adopted the 2-RSPG for the numerical stability. In what follows, we refer to the 2-RSPG as the RSPG. As a comparative method, we implemented the SGD with the same parameter setting described in Sect. 4.2.1. All results were obtained in R version 3.3.0 with Intel Core i7-4790K machine.

4.5.1 Simulation Model

We used the simulation model given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e, \quad e \sim N(0, 0.5^2).$$

The sample size and the number of explanatory variables were set to be $N = 10000, 30000$ and $p = 1000, 2000$, respectively. The true coefficients were given by

$$\beta_1 = 1,\ \beta_2 = 2,\ \beta_4 = 4,\ \beta_7 = 7,\ \beta_{11} = 11,$$
$$\beta_j = 0 \text{ for } j \in \{0, \ldots, p\} \backslash \{1, 2, 4, 7, 11\}.$$

We arranged a broad range of regression coefficients to observe sparsity for various degrees of regression coefficients. The explanatory variables were generated from a normal distribution $N(0, \Sigma)$ with $\Sigma = (0.2^{|i-j|})_{1 \leq i,j \leq p}$. We generated 30 random samples.

Outliers were incorporated into simulations. We set the outlier ratio ($\varepsilon = 0.2$) and the outlier pattern that the outliers were generated around the middle part of the explanatory variable, where the explanatory variables were generated from $N(0, 0.5^2)$ and the error terms were generated from $N(20, 0.5^2)$.

4.5.2 Performance Measure

The empirical regularized risk and the (approximated) expected regularized risk were used to verify the fitness of regression:

$$\begin{aligned}
\text{EmpRisk} &= \frac{1}{N}\sum_{i=1}^{N} -\frac{f(y_i|x_i;\hat{\theta})^{\gamma}}{\left(\int f(y|x_i;\hat{\theta})^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}} + \lambda\|\hat{\beta}\|_1, \\
\text{ExpRisk} &= E_{g(x,y)}\left[-\frac{f(y|x;\hat{\theta})^{\gamma}}{\left(\int f(y|x;\hat{\theta})^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}}\right] + \lambda\|\hat{\beta}\|_1 \\
&\approx \frac{1}{N_{test}}\sum_{i=1}^{N_{test}} -\frac{f(y_i^*|x_i^*;\hat{\theta})^{\gamma}}{\left(\int f(y|x_i^*;\hat{\theta})^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}} + \lambda\|\hat{\beta}\|_1,
\end{aligned}$$

where $f(y|x;\hat{\theta}) = \phi(y;\hat{\beta}_0 + x^T\hat{\beta}, \hat{\sigma}^2)$ and (x_i^*, y_i^*) $(i = 1,\ldots,N_{test})$ is test samples generated from the simulation model with outlier scheme. In this experiment, we used $N_{test} = 70000$.

4.5.3 Initial Point and Tuning Parameter

In our method, we need an initial point and some tuning parameters to obtain the estimate. Therefore, we used $N_{init} = 200$ samples which were used for estimating an initial point and other parameters L in (4.13) and τ^2 in (4.15) to calculate in advance. We suggest the following ways to prepare an initial point. The estimate of other conventional robust and sparse regression methods would give a good initial point. For another choice, the estimate of the RANSAC (random sample consensus) algorithm would also give a good initial point. In this experiment, we added the noise to the estimate of the RANSAC and used it as an initial point.

For estimating L and τ^2, we followed the way to in Sect. 6 of [38]. Moreover, we used the following value of tuning parameters in this experiment. The parameter γ in the γ-divergence was set to 0.1. The parameter λ of L_1 regularization was set to $10^{-1}, 10^{-2}, 10^{-3}$.

The RSPG needed the number of candidates N_{cand} and post-samples N_{post} for post-optimization as described in Algorithm 3. Then, we used $N_{cand} = 5$ and $N_{post} = \lceil N/10 \rceil$.

4.5.4 Result

Tables 4.1-4.3 show the EmpRisk, ExpRisk and computation time in the case $\lambda = 10^{-3}, 10^{-2}$, and 10^{-1}. Except for the computation time, our method outperformed comparative methods with several sizes of sample and dimension. We verify that the SGD, which are not theoretically guaranteed to converge for non-convex loss, cannot reach the stationary point numerically. For the computation time, our method was comparable to the SGD.

Table 4.1 EmpRisk, ExpRisk and computation time for $\lambda = 10^{-3}$

	$N = 10000$, $p = 1000$			$N = 30000$, $p = 1000$		
Methods	EmpRisk	ExpRisk	Time	EmpRisk	ExpRisk	Time
RSPG	-0.629	-0.628	75.2	-0.692	-0.691	78.3
SGD with 1 mini-batch	-0.162	-0.155	95.9	-0.365	-0.362	148
SGD with 10 mini-batch	1.1×10^{-2}	1.45×10^{-2}	73.2	-5.71×10^{-2}	-5.6×10^{-2}	73.7
SGD with 30 mini-batch	4.79×10^{-2}	5.02×10^{-2}	71.4	-5.71×10^{-2}	-5.6×10^{-2}	73.7
SGD with 50 mini-batch	6.03×10^{-2}	6.21×10^{-2}	71.1	-3.98×10^{-2}	-3.88×10^{-2}	238

	$N = 10000$, $p = 2000$			$N = 30000$, $p = 2000$		
Methods	EmpRisk	ExpRisk	Time	EmpRisk	ExpRisk	Time
RSPG	-0.646	-0.646	117	-0.696	-0.696	125
SGD with 1 mini-batch	0.187	0.194	145	-3.89×10^{-2}	-3.56×10^{-2}	251
SGD with 10 mini-batch	0.428	0.431	99.2	0.357	0.359	112
SGD with 30 mini-batch	0.479	0.481	95.7	0.442	0.443	101
SGD with 50 mini-batch	0.496	0.499	166	0.469	0.47	337

Table 4.2 EmpRisk, ExpRisk and computation time for $\lambda = 10^{-2}$

	$N = 10000$, $p = 1000$			$N = 30000$, $p = 1000$		
Methods	EmpRisk	ExpRisk	Time	EmpRisk	ExpRisk	Time
RSPG	-0.633	-0.632	75.1	-0.65	-0.649	78.4
SGD with 1 mini-batch	-0.322	-0.322	96.1	-0.488	-0.487	148
SGD with 10 mini-batch	1.36	1.37	73.4	0.164	0.165	79.7
SGD with 30 mini-batch	2.61	2.61	71.6	1.34	1.34	73.9
SGD with 50 mini-batch	3.08	3.08	409	1.95	1.95	576

	$N = 10000$, $p = 2000$			$N = 30000$, $p = 2000$		
Methods	EmpRisk	ExpRisk	Time	EmpRisk	ExpRisk	Time
RSPG	-0.647	-0.646	117	-0.66	-0.66	125
SGD with 1 mini-batch	-0.131	-0.13	144	-0.436	-0.435	250
SGD with 10 mini-batch	3.23	3.23	99.1	0.875	0.875	112
SGD with 30 mini-batch	5.63	5.63	95.6	3.19	3.19	100
SGD with 50 mini-batch	6.52	6.53	503	4.38	4.38	675

Table 4.3 EmpRisk, ExpRisk and computation time for $\lambda = 10^{-1}$

	$N = 10000$, $p = 1000$			$N = 30000$, $p = 1000$		
Methods	EmpRisk	ExpRisk	Time	EmpRisk	ExpRisk	Time
RSPG	-0.633	-0.632	74.6	-0.64	-0.639	78.1
SGD with 1 mini-batch	-0.411	-0.411	95.6	-0.483	-0.482	148
SGD with 10 mini-batch	0.483	0.483	72.9	-4.56$\times 10^{-2}$	-4.5$\times 10^{-2}$	79.6
SGD with 30 mini-batch	1.53	1.53	71.1	0.563	0.563	73.7
SGD with 50 mini-batch	2.39	2.39	70.8	0.963	0.963	238
	$N = 10000$, $p = 2000$			$N = 30000$, $p = 2000$		
Methods	EmpRisk	ExpRisk	Time	EmpRisk	ExpRisk	Time
RSPG	-0.654	-0.653	116	-0.66	-0.66	130
SGD with 1 mini-batch	-0.462	-0.461	144	-0.559	-0.558	262
SGD with 10 mini-batch	0.671	0.672	98.9	-9.71$\times 10^{-2}$	-9.62$\times 10^{-2}$	116
SGD with 30 mini-batch	2.43	2.44	95.4	0.697	0.697	104
SGD with 50 mini-batch	4.02	4.02	165	1.32	1.32	340

4.6 Application to Real Data

We applied our method ‘online sparse γ-Poisson’ to real data ‘Online News Popularity’ (Fernandes, Vinagre, and Cortez [30]), which is available at https://archive.ics.uci.edu/ml/datasets/online+news+popularity. We compared our method with sparse Poisson regression which was implemented by R-package ‘glmnet’ with default parameter setting.

Online News Popularity dataset contains 39644 samples with 58 dimensional explanatory variables. We divided the dataset to 20000 training and 19644 test samples. In Online News Popularity dataset, the exposure time of each sample is different. Then, we used the log transformed feature value ‘timedelta’ as the offset term. Moreover, 2000 training samples were randomly selected. Outliers were incorporated into training samples as follows:

$$y_{outlier,i} = y_i + 100 \times t_i \quad (i = 1, \dots, 2000),$$

where i is the index of the randomly selected sample and y_i is the response variable of the i-th randomly selected sample and t_i is the offset term of the i-th randomly selected sample.

As a measure of predictive performance, the root trimmed mean squared prediction error (RTMSPE) was computed for the test samples given by

$$\text{RTMSPE} = \sqrt{\frac{1}{h}\sum_{j=1}^{h} e_{[j]}^2},$$

where $e_j^2 = \left(y_j - \left\lfloor \exp\left(\log(t_j) + \hat{\beta}_0 + x_j^T \hat{\beta}\right) \right\rfloor \right)^2$, $e_{[1]}^2 \leq \cdots \leq e_{[19644]}^2$ are the order statistics of $e_1^2, \cdots, e_{19644}^2$ and $h = \lfloor (19644+1)(1-\alpha) \rfloor$ with $\alpha = 0.05, \cdots, 0.3$.

In our method, we need an initial point and some tuning parameters to obtain the estimate. Therefore, we used $N_{init} = 200$ samples which were used for estimating an initial point and other parameters L in (4.13) and τ^2 in (4.15) to calculate in advance. In this experiment, we used the estimate of the RANSAC. For estimating L, we followed the way to in [38], page 298-299. Moreover, we used the following value of tuning parameters in this experiment. The parameter γ in the γ-divergence was set to $0.1, 0.5, 1.0$. The parameter λ of L_1 regularization was selected by the robust cross-validation proposed by Kawashima and Fujisawa [57]. The robust cross-validation was given by:

$$\mathrm{RoCV}(\lambda) = -\frac{1}{n} \sum_{i=1}^{n} \frac{f(y_i|x_i; \hat{\theta}^{[-i]})^{\gamma_0}}{\left(\int f(y|x_i; \hat{\theta}^{[-i]})^{1+\gamma_0} dy\right)^{\frac{\gamma_0}{1+\gamma_0}}},$$

where $\hat{\theta}^{[-i]}$ is the estimated parameter deleting the i-th observation and γ_0 is an appropriate tuning parameter. In this experiment, γ_0 was set to 1.0. The mini-batch size was set to $100, 200, 500$. The RSPG needed the number of candidates and post-samples N_{cand} and N_{post} for post-optimization as described in Algorithm 3. We used $N_{cand} = 5$ and $N_{post} = \lceil N/10 \rceil$. We showed the best result of our method and comparative method in Table 4.4. All results were obtained in R version 3.3.0 with Intel Core i7-4790K machine. Table 4.4 shows that our method performed better than sparse Poisson regression.

Table 4.4 Root trimmed mean squared prediction error in test samples

	trimming fraction $100\alpha\%$					
Methods	5%	10%	15%	20%	25%	30%
Our method	2419.3	1760.2	1423.7	1215.7	1064	948.9
Sparse Poisson Regression	2457.2	2118.1	1902.5	1722.9	1562.5	1414.1

Chapter 5

Robust Regression via γ-divergence against Heterogeneous Contamination

5.1 Revisiting the Estimation of the γ-Regression

The γ-divergence for the i.i.d. problem was first proposed by Fujisawa and Eguchi [35]. It measures the difference between two probability density functions. As stated earlier, the γ-divergence for regression was first proposed by Fujisawa and Eguchi [35] and we refer to it as the type I. Then, we proposed the other γ-divergence for regression and refer to it as the type II. In this section, we briefly review both types of γ-divergence for regression and present the corresponding parameter estimation.

Theoretical properties of the γ-divergence for the i.i.d. problemwere deeply investigated by Fujisawa and Eguchi [35]. Theoretical properties of the γ-divergence for regression were studied by Fujisawa and Eguchi [35], Kanamori and Fujisawa [54], but not well under heterogeneous contamination, which is special in the regression problem and does not appear in the i.i.d. problem. Hung, Jou, and Huang [51] pointed out that a logistic regression model with mislabeled data can be regarded as a logistic regression model with heterogeneous contamination and then applied the type I to a usual logistic regression model, which enables us to estimate the parameter of the logistic regression model without modelling mislabeled scheme even if mislabeled data exist. They also investigated theoretical properties on robustness, but they assumed that γ is sufficiently large. In Sect. 5.2, we will see that the type I is superior to type II under heterogeneous contamination in the sense of the strong robustness without assuming that γ is sufficiently large. Here we mention that the density power divergence [4] is another candidate of divergence which gives robustness, but it does not have the strong robustness [35, 51].

5.1.1 Estimation for γ-Regression

Let $f(y|x;\theta)$ be a conditional probability density function of y given x with parameter θ. Both types of γ-cross entropy for regression are given by

$$d_{\gamma,1}(g(y|x), f(y|x;\theta); g(x)) = -\frac{1}{\gamma}\log E_{g(x,y)}\left[\frac{f(y|x;\theta)^{\gamma}}{(\int f(y|x;\theta)^{1+\gamma}dy)^{\frac{\gamma}{1+\gamma}}}\right],$$
$$d_{\gamma,2}(g(y|x), f(y|x;\theta); g(x)) = -\frac{1}{\gamma}\log E_{g(x,y)}\left[f(y|x;\theta)^{\gamma}\right] + \frac{1}{1+\gamma}\log E_{g(x)}\left[f(y|x;\theta)^{1+\gamma}dy\right].$$

The target parameter can be defined as the minimizer by

$$\theta^*_{\gamma,j} = \arg\min_{\theta} d_{\gamma,j}(g(y|x), f(y|x;\theta); g(x)) \quad \text{for } j = 1, 2.$$

Suppose that $f(y|x;\theta^*)$ is the target conditional probability density function. The latent bias is expressed as $\theta^*_{\gamma,j} - \theta^*$. This is zero when the underlying model belongs to a parametric model, in other words, $g(y|x) = f(y|x;\theta^*)$, but is not always zero when the underlying model is contaminated by outliers. This issue will be discussed in Sect. 5.2.

5.1.2 Parameter Estimation for Location-Scale Family

Here we show that both types of γ-divergence give the same parameter estimation when the parametric conditional probability density function $f(y|x;\theta)$ belongs to a location-scale family in which the scale does not depend on the explanatory variables, given by

$$f(y|x;\theta) = \frac{1}{\sigma}s\left(\frac{y - q(x;\zeta)}{\sigma}\right), \tag{5.1}$$

where $s(y)$ is a probability density function, σ is a scale parameter and $q(x;\zeta)$ is a location function with a regression parameter ζ, e.g., $q(x;\zeta) = x^T\zeta$. Then, we can obtain

$$\begin{aligned}\int f(y|x;\theta)^{1+\gamma}dy &= \int \frac{1}{\sigma^{1+\gamma}}s\left(\frac{y - q(x;\zeta)}{\sigma}\right)^{1+\gamma}dy \\ &= \sigma^{-\gamma}\int s(z)^{1+\gamma}dz. \end{aligned}\tag{5.2}$$

This does not depend on the explanatory variables x. Using this property, we can show that both types of γ-cross entropy are the same as follows:

$$\begin{aligned}
&d_{\gamma,1}(g(y|x), f(y|x;\theta); g(x)) \\
&= -\frac{1}{\gamma}\log \int \left\{ \int g(y|x) f(y|x;\theta)^{\gamma} dy \Big/ \left(\int f(y|x;\theta)^{1+\gamma} dy \right)^{\frac{\gamma}{1+\gamma}} \right\} g(x) dx \\
&= -\frac{1}{\gamma}\log \left\{ \int \int g(x,y) f(y|x;\theta)^{\gamma} dx dy \Big/ \left(\int f(y|x;\theta)^{1+\gamma} dy \right)^{\frac{\gamma}{1+\gamma}} \right\} \\
&= -\frac{1}{\gamma}\log \int \int g(x,y) f(y|x;\theta)^{\gamma} dx dy + \frac{1}{1+\gamma}\log \int f(y|x;\theta)^{1+\gamma} dy \\
&= -\frac{1}{\gamma}\log \int \int g(x,y) f(y|x;\theta)^{\gamma} dx dy + \frac{1}{1+\gamma}\log \int f(y|x;\theta)^{1+\gamma} dy \int g(x) dx \\
&= d_{\gamma,2}(g(y|x), f(y|x;\theta); g(x)).
\end{aligned}$$

The second equality holds from (5.2). As a result, both types of γ-divergence give the same parameter estimation, because the estimator is defined by the empirical estimation of the cross entropy. However, it should be noted that both types of γ-divergence are not the same, because $d_{\gamma,1}(g(y|x), g(y|x); g(x)) \neq d_{\gamma,2}(g(y|x), g(y|x); g(x))$.

5.2 Robust Properties

In this section, we show a distinct difference between two types of γ-divergence.

5.2.1 Contamination Model and Basic Condition

Let $\delta(y|x)$ be the contamination conditional probability density function related to outliers. Let $\varepsilon(x)$ and ε denote the outlier ratios which depends on x and does not, respectively. Suppose that the underlying conditional probability density functions under heterogeneous and homogeneous contaminations are given by

$$\begin{aligned}
g(y|x) &= (1-\varepsilon(x)) f(y|x;\theta^*) + \varepsilon(x)\delta(y|x), \\
g(y|x) &= (1-\varepsilon) f(y|x;\theta^*) + \varepsilon\delta(y|x).
\end{aligned}$$

Let

$$\nu_{f,\gamma}(x) = \left\{ \int \delta(y|x) f(y|x)^{\gamma} dy \right\}^{\frac{1}{\gamma}}, \qquad \nu_{f,\gamma} = \left\{ \int \nu_{f,\gamma}(x)^{\gamma} g(x) dx \right\}^{\frac{1}{\gamma}}.$$

Here we assume that

$$\nu_{f_{\theta^*},\gamma} \approx 0.$$

This is an extended assumption used for the i.i.d. problem[35] to the regression problem. This assumption implies that $\nu_{f_{\theta^*},\gamma}(x) \approx 0$ for any x (a.e.) and illustrates that the contamination conditional probability density function $\delta(y|x)$ lies on the tail of the target conditional probability density function $f(y|x;\theta^*)$. For example, if $\delta(y|x)$ is the Dirac delta function at the outlier $y_\dagger(x)$ given x, then we have $\nu_{f_{\theta^*},\gamma}(x) = f(y_\dagger(x)|x;\theta^*) \approx 0$, which is reasonable because $y_\dagger(x)$ is an outlier.

Here we also consider the condition $\nu_{f_\theta,\gamma} \approx 0$, which is used later. This will be true in the neighbourhood of $\theta = \theta^*$. In addition, even when θ is not close to θ^*, if $\delta(y|x)$ lies on the tail of $f(y|x;\theta)$, we can see $\nu_{f_\theta,\gamma} \approx 0$.

To make the discussion easier, we prepare the monotone transformation of both types of γ-cross entropy for regression by

$$\begin{aligned}
\tilde{d}_{\gamma,1}(g(y|x), f(y|x;\theta); g(x)) &= -\exp\left\{-\gamma d_{\gamma,1}(g(y|x), f(y|x;\theta); g(x))\right\} \\
&= -\int\int \frac{f(y|x;\theta)^\gamma}{\left(\int f(y|x;\theta)^{1+\gamma} dy\right)^{\frac{\gamma}{1+\gamma}}} g(y|x) g(x) dx dy, \\
\tilde{d}_{\gamma,2}(g(y|x), f(y|x;\theta); g(x)) &= -\exp\left\{-\gamma d_{\gamma,2}(g(y|x), f(y|x;\theta); g(x))\right\} \\
&= -\frac{\int \left(\int g(y|x) f(y|x;\theta)^\gamma dy\right) g(x) dx}{\left\{\int \left(\int f(y|x;\theta)^{1+\gamma} dy\right) g(x) dx\right\}^{\frac{\gamma}{1+\gamma}}}.
\end{aligned}$$

5.2.2 Robustness of Type I

We see

$$\begin{aligned}
&\tilde{d}_{\gamma,1}(g(y|x), f(y|x;\theta); g(x)) \\
&= -\int \frac{\int g(y|x) f(y|x;\theta)^\gamma dy}{\left(\int f(y|x;\theta)^{1+\gamma} dy\right)^{\frac{\gamma}{1+\gamma}}} g(x) dx \\
&= -\int \frac{\int \{(1-\varepsilon(x)) f(y|x;\theta^*) + \varepsilon(x)\delta(y|x)\} f(y|x;\theta)^\gamma dy}{\left(\int f(y|x;\theta)^{1+\gamma} dy\right)^{\frac{\gamma}{1+\gamma}}} g(x) dx \\
&= -\int \frac{\int f(y|x;\theta^*) f(y|x;\theta)^\gamma dy}{\left(\int f(y|x;\theta)^{1+\gamma} dy\right)^{\frac{\gamma}{1+\gamma}}} (1-\varepsilon(x)) g(x) dx - \int \frac{\int \delta(y|x;\theta) f(y|x;\theta)^\gamma dy}{\left(\int f(y|x;\theta)^{1+\gamma} dy\right)^{\frac{\gamma}{1+\gamma}}} \varepsilon(x) g(x) dx \\
&= -\tilde{d}_{\gamma,1}(f(y|x;\theta^*), f(y|x;\theta)); \tilde{g}(x)) - \int \frac{\nu_{f_\theta,\gamma}(x)^\gamma}{\left(\int f(y|x;\theta)^{1+\gamma} dy\right)^{\frac{\gamma}{1+\gamma}}} \varepsilon(x) g(x) dx,
\end{aligned}$$

where $\tilde{g}(x) = (1-\varepsilon(x))g(x)$. From this relation, we can easily show the following theorem.

Theorem 5.2.1. *Under the condition $\nu_{f_\theta,\gamma} \approx 0$ and $\int f(y|x;\theta)^{1+\gamma}dy > 0$, we have*

$$\tilde{d}_{\gamma,1}(g(y|x), f(y|x;\theta); g(x)) \approx \tilde{d}_{\gamma,1}(f(y|x;\theta^*), f(y|x;\theta); \tilde{g}(x)).$$

Using this theorem, we can expect that the latent bias $\theta^*_{\gamma,1} - \theta^*$ is close to zero, because

$$\arg\min_\theta \tilde{d}_{\gamma,1}(g(y|x), f(y|x;\theta); g(x)) = \arg\min_\theta d_{\gamma,1}(g(y|x), f(y|x;\theta); g(x)) \; = \; \theta^*_{\gamma,1}$$
$$\arg\min_\theta \tilde{d}_{\gamma,1}(f(y|x;\theta^*), f(y|x;\theta); \tilde{g}(x)) = \arg\min_\theta d_{\gamma,1}(f(y|x;\theta^*), f(y|x;\theta); \tilde{g}(x)) \; = \; \theta^*.$$

The last equality holds even when $g(x)$ is replaced by $\tilde{g}(x) = (1-\varepsilon(x))g(x)$.

In addition, we can have the modified Pythagorean relation approximately.

Theorem 5.2.2. *Under the condition $\nu_{f_\theta,\gamma} \approx 0$ and $\int f(y|x;\theta)^{1+\gamma}dy > 0$, the modified Pythagorean relation among $g(y|x)$, $f(y|x;\theta^*)$, $f(y|x;\theta)$ approximately holds:*

$$D_{\gamma,1}(g(y|x), f(y|x;\theta); g(x)) \approx D_{\gamma,1}(g(y|x), f(y|x;\theta^*); g(x)) + D_{\gamma,1}(f(y|x;\theta^*), f(y|x;\theta); \tilde{g}(x)).$$

The modified Pythagorean relation implies that the minimizer of $D_{\gamma,1}(g(y|x), f(y|x;\theta); g(x))$ is almost the same as the minimizer of $D_{\gamma,1}(f(y|x;\theta^*), f(y|x;\theta); \tilde{g}(x))$, which is θ^*. This also implies the strong robustness.

In the theorems, we assume $\nu_{f_\theta,\gamma} \approx 0$ and $\int f(y|x;\theta)^{1+\gamma}dy > 0$. The former condition was already discussed in Sect. 5.2.1. Here we investigate the latter condition. When the parametric conditional probability density function belongs to a location-scale family (5.1), this condition will be expected to hold, because

$$\int f(y|x;\theta)^{1+\gamma}dy = \int \frac{1}{\sigma^{1+\gamma}} s\left(\frac{y-q(x;\zeta)}{\sigma}\right)^{1+\gamma} dy = \frac{1}{\sigma^\gamma}\int s(z)^{1+\gamma}dz.$$

We can also verify that this condition holds for a logistic regression model, a Poisson regression model, and so on.

Finally we mention the homogeneous contamination. The modified Pythagorean relation in Theorem 5.2.2 is changed to the usual Pythagorean relation, because we can easily see $D_{\gamma,1}(f(y|x;\theta^*), f(y|x;\theta); \tilde{g}(x)) = D_{\gamma,1}(f(y|x;\theta^*), f(y|x;\theta); g(x))$ under homogeneous contamination.

5.2.3 Robustness of Type II

First, we illustrate that the strong robustness does not hold in general under heterogeneous contamination, unlike for type I. We see

$$\begin{aligned}
&\tilde{d}_{\gamma,2}(g(y|x), f(y|x;\theta); g(x))\\
&= -\frac{\int \left(\int g(y|x) f(y|x;\theta)^{\gamma} dy\right) g(x) dx}{\left\{\int \left(\int f(y|x;\theta)^{1+\gamma} dy\right) g(x) dx\right\}^{\frac{\gamma}{1+\gamma}}}\\
&= -\frac{\int \left(\int (1-\varepsilon(x)) f(y|x;\theta^*) f(y|x;\theta)^{\gamma} dy + \int \varepsilon(x)\delta(y|x) f(y|x;\theta)^{\gamma} dy\right) g(x) dx}{\left\{\int \left(\int f(y|x;\theta)^{1+\gamma} dy\right) g(x) dx\right\}^{\frac{\gamma}{1+\gamma}}}\\
&= -\frac{\int \left(\int (1-\varepsilon(x)) f(y|x;\theta^*) f(y|x;\theta)^{\gamma} dy + \int \varepsilon(x) \nu_{f_\theta,\gamma}(x)\right) g(x) dx}{\left\{\int \left(\int f(y|x;\theta)^{1+\gamma} dy\right) g(x) dx\right\}^{\frac{\gamma}{1+\gamma}}}\\
&\approx -\frac{\int \int f(y|x;\theta^*) f(y|x;\theta)^{\gamma} dy (1-\varepsilon(x)) g(x) dx}{\left\{\int \left(\int f(y|x;\theta)^{1+\gamma} dy\right) g(x) dx\right\}^{\frac{\gamma}{1+\gamma}}}.
\end{aligned}$$

The last approximation holds from $\nu_{f_\theta,\gamma}(x) \approx 0$. This can not be expressed using $d_\gamma(f(y|x;\theta^*), f(y|x;\theta); h(x))$ with an appropriate base measure $h(x)$, unlike for type I, because the base measure of the numerator on the explanatory variables is different from that of the denominator. As will be shown in numerical experiments, the type II presents a significant bias under heterogeneous contamination. However, as already mentioned in Sect. 5.1.2, when the parametric conditional probability density function belongs to a location-scale family (5.1), the cross entropy for type II is the same as that for type I and then the type II can have the strong robustness. In addition, under homogeneous contamination, we have $\tilde{d}_{\gamma,2}(g(y|x), f(y|x;\theta); g(x)) \approx (1-\varepsilon)\tilde{d}_{\gamma,2}(f(y|x;\theta^*), f(y|x;\theta); g(x))$ and then we expect that the latent bias $\theta^*_{\gamma,2} - \theta^*$ is sufficiently small.

5.3 Numerical Experiment

In this section, using a simulation model, we compare the type I with the type II.

As shown in Sect. 5.2, the distinct difference occurs under heterogeneous contamination when the parametric conditional probability density function $f(y|x;\theta)$ does not belong to a location-scale family. Therefore, we used the logistic regression model as the simulation model, given by

$$Pr(y=1|x) = \pi(x;\beta),\ Pr(y=0|x) = 1-\pi(x;\beta),$$

where $\pi(x;\beta) = \{1+\exp(-\beta_0 - x_1\beta_1 - \cdots - x_p\beta_p)\}^{-1}$. The sample size and the number of explanatory variables were set to be $n = 2000$ and $p = 5$, respectively. The true coefficients were given by

$$\beta_0 = 0,\ \beta_1 = 1,\ \beta_2 = -1,\ \beta_3 = 1,\ \beta_4 = -1,\ \beta_5 = 0.$$

The explanatory variables were generated from a normal distribution $N(0,\Sigma)$ with $\Sigma = (0.2^{|i-j|})_{1\leq i,j\leq p}$. We generated 100 random samples.

Outliers were incorporated into simulations. We investigated four outlier ratios ($\varepsilon = 0.1,\ 0.2,\ 0.3$ and 0.4) and the following outlier pattern: The outliers were generated around the edge part of the explanatory variables, where the explanatory variables were generated from $N(\boldsymbol{\mu}_{\text{out}}, 0.5^2\mathbf{I})$ where $\boldsymbol{\mu}_{\text{out}} = (20,0,20,0,0)$ and the response variable y is set to 0.

In order to verify the fitness of the regression coefficient, we used the mean squared error (MSE) as the performance measure, given by

$$\text{MSE} = \frac{1}{p+1}\sum_{j=0}^{p}(\hat{\beta}_j - \beta_j^*)^2,$$

where β_j^*'s are the true coefficients. The tuning parameter γ in the γ-divergence was set to 0.5 and 1.0.

Table 5.1 shows the MSE in the case $\varepsilon = 0.1,\ 0.2,\ 0.3$ and 0.4. The type I presented smaller MSEs than the type II. The difference between two types was larger as the outlier ratio ε was larger.

Table 5.1 MSE under heterogeneous contamination

Methods	$\gamma = 0.5$	$\gamma = 1.0$
	$\varepsilon = 0.1$	
Type I	0.00620	0.00712
Type II	0.00810	0.0276
	$\varepsilon = 0.2$	
Type I	0.0136	0.0149
Type II	0.0215	0.110
	$\varepsilon = 0.3$	
Type I	0.0262	0.0282
Type II	0.0472	0.282
	$\varepsilon = 0.4$	
Type I	0.0514	0.0547
Type II	0.0998	0.648

Chapter 6

Conclusion

This thesis mainly focuses on robust regression methods based on the γ-divergence and incorporates some sparse regularization techniques into them and investigates theoretical robust properties on γ-divergence.

In Chapter 3, we proposed the robust linear regression method with sparsity based on the γ-divergence. We showed desirable robust properties under both homogeneous and heterogeneous contamination. In particular, we presented the Pythagorean relation for the regression case, although it was not shown in Kanamori and Fujisawa [54]. In most of the robust and sparse regression methods, it is difficult to obtain the efficient estimation algorithm, because the objective function is non-convex and non-differentiable. Nonetheless, we succeeded to propose the efficient estimation algorithm, which has a monotone decreasing property of the objective function by using the MM-algorithm. The numerical experiments and real data analyses suggested that our method was superior to comparative robust and sparse linear regression methods in terms of both accuracy and computational costs. However, in numerical experiments, a few results of performance measure "TNR" were a little less than the best results. Therefore, if more sparsity of coefficients is needed, other sparse penalties, e.g., the Smoothly Clipped Absolute Deviations (SCAD) [28] and the Minimax Concave Penalty (MCP) [100], can also be useful.

In Chapter 4, we proposed the online robust regression methods in GLM based on the γ-divergence. We applied a stochastic optimization approach in order to reduce the computational complexity and overcome the computational problem on the hypergeometric series in Poisson regression. We adopted the RSPG, which guaranteed the global convergence property for non-convex stochastic optimization problem, as a stochastic optimization approach. We proved that the global convergence property can be extended to the classical first-order necessary condition. In this paper, linear/logistic/Poisson regression problems with L_1 regularization were illustrated in detail. As a result, not only Poisson case but also linear/logistic case can scale well for very large problems by virtue of the stochastic optimization approach. To the best of our knowledge, there is no efficient method for the robust and sparse Poisson regression, but we have succeeded to propose an efficient estimation procedure with online strategy. The numerical experiments and real data analysis suggested that our methods had good performances in terms of both accuracy and computational cost. However, there are still some problems in Poisson regression problem, e.g., overdispersion [21], zero inflated Poisson [62]. Therefore, it can be useful to extend the Poisson regression to the negative binomial regression and the zero inflated Poisson regression for future work. Moreover, the accelerated RSPG was proposed in [37], and then we can adopt it as a stochastic optimization approach in order to achieve faster convergence than the RSPG.

In Chapter 5, we investigated both types of γ-divergence for regression in terms of the parameters estimation and robust properties. We pointed out that the parameter estimation of both types of γ-divergence is the same under the assumption that the parametric conditional probability density function belongs to a location-scale family. Moreover, we elucidated a distinct difference between both types of γ-divergence form the view point of the latent bias and the Pythagorean relation. The numerical experiments were illustrated to verify the difference of the theoretical robust property between the type I and type II under heterogeneous contamination.

References

[1] C. C. Aggarwal and P. S. Yu. "Outlier Detection for High Dimensional Data". In: *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*. SIGMOD '01. Santa Barbara, California, USA: ACM, 2001, pp. 37–46. ISBN: 1-58113-332-4. DOI: 10.1145/375663.375668. URL: http://doi.acm.org/10.1145/375663.375668.

[2] A. Alfons, C. Croux, and S. Gelper. "Sparse least trimmed squares regression for analyzing high-dimensional large data sets". In: *The Annals of Applied Statistics* 7.1 (2013), pp. 226–248.

[3] M. Avella-Medina. "Influence functions for penalized M-estimators". In: *Bernoulli* 23.4B (Nov. 2017), pp. 3178–3196. DOI: 10.3150/16-BEJ841. URL: https://doi.org/10.3150/16-BEJ841.

[4] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. "Robust and efficient estimation by minimising a density power divergence". In: *Biometrika* 85.3 (1998), pp. 549–559.

[5] A. Basu, H. Shioya, and C. Park. *Statistical inference: the minimum distance approach*. CRC Press, 2011.

[6] A. Beck and M. Teboulle. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". In: *SIAM J. Img. Sci.* 2.1 (Mar. 2009), pp. 183–202. ISSN: 1936-4954. DOI: 10.1137/080716542. URL: http://dx.doi.org/10.1137/080716542.

[7] A. M. Bianco and V. J. Yohai. "Robust Estimation in the Logistic Regression Model". In: *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday*. Ed. by Helmut Rieder. New York, NY: Springer New York, 1996, pp. 17–34. ISBN: 978-1-4612-2380-1. DOI: 10.1007/978-1-4612-2380-1_2. URL: https://doi.org/10.1007/978-1-4612-2380-1_2.

[8] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector". In: *Ann. Statist.* 37.4 (Aug. 2009), pp. 1705–1732. DOI: 10.1214/08-AOS620. URL: https://doi.org/10.1214/08-AOS620.

[9] P. Bloomfield and W. Steiger. "Least Absolute Deviations Curve-Fitting". In: *Siam Journal on Scientific and Statistical Computing* 1 (June 1980).

[10] J. Bootkrajang and A. Kabán. "Classification of mislabelled microarrays using robust sparse logistic regression". In: *Bioinformatics* 29.7 (2013), pp. 870–877. DOI: 10.1093/bioinformatics/btt078. URL: +%20http://dx.doi.org/10.1093/bioinformatics/btt078.

[11] J. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

[12] L. Bottou. “Large-scale machine learning with stochastic gradient descent”. In: *Proceedings of COMPSTAT’2010*. Springer, 2010, pp. 177–186.

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers”. In: *Found. Trends Mach. Learn.* 3.1 (Jan. 2011), pp. 1–122. ISSN: 1935-8237. DOI: 10.1561/2200000016. URL: http://dx.doi.org/10.1561/2200000016.

[14] S. Bubeck. “Convex Optimization: Algorithms and Complexity”. In: *Found. Trends Mach. Learn.* 8.3-4 (Nov. 2015), pp. 231–357. ISSN: 1935-8237. DOI: 10.1561/2200000050. URL: http://dx.doi.org/10.1561/2200000050.

[15] P. Bühlmann and S. A. Van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. 1st. Springer Publishing Company, Incorporated, 2011. ISBN: 3642201911, 9783642201912.

[16] F. Bunea, A. B. Tsybakov, and M. Wegkamp. “Sparsity oracle inequalities for the Lasso”. In: *Electron. J. Statist.* 1 (2007), pp. 169–194. DOI: 10.1214/07-EJS008. URL: https://doi.org/10.1214/07-EJS008.

[17] E. Candes and T. Tao. “The Dantzig selector: Statistical estimation when p is much larger than n”. In: *Ann. Statist.* 35.6 (Dec. 2007), pp. 2313–2351. DOI: 10.1214/009053606000001523. URL: https://doi.org/10.1214/009053606000001523.

[18] R. J. Carroll and S. Pederson. “On Robustness in the Logistic Regression Model”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 55.3 (1993), pp. 693–706. ISSN: 00359246. URL: http://www.jstor.org/stable/2345881.

[19] E. C. Chi and D. W. Scott. “Robust Parametric Classification and Variable Selection by a Minimum Distance Criterion”. In: *Journal of Computational and Graphical Statistics* 23.1 (2014), pp. 111–128. DOI: 10.1080/10618600.2012.737296.

[20] C. Croux, C. Flandre, and G. Haesbroeck. “The breakdown behavior of the maximum likelihood estimator in the logistic regression model”. In: *Statistics & Probability Letters* 60.4 (2002), pp. 377–386. ISSN: 0167-7152. DOI: https://doi.org/10.1016/S0167-7152(02)00292-4. URL: http://www.sciencedirect.com/science/article/pii/S0167715202002924.

[21] C. Dean and J. F. Lawless. “Tests for Detecting Overdispersion in Poisson Regression Models”. In: *Journal of the American Statistical Association* 84.406 (1989), pp. 467–472. DOI: 10.1080/01621459.1989.10478792. eprint: http://www.tandfonline.com/doi/pdf/10.1080/01621459.1989.10478792. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478792.

[22] J. C. Duchi, E. Hazan, and Y. Singer. “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization.” In: *Journal of Machine Learning Research* 12 (2011), pp. 2121–2159. URL: http://dblp.uni-trier.de/db/journals/jmlr/jmlr12.html#DuchiHS11.

[23] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. “Composite Objective Mirror Descent”. In: (2010), pp. 14–26. URL: http://colt2010.haifa.il.ibm.com/papers/COLT2010proceedings.pdf#page=22.

[24] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. “Efficient Projections Onto the L1-ball for Learning in High Dimensions”. In: ICML ’08 (2008), pp. 272–279. DOI: 10.1145/1390156.1390191. URL: http://doi.acm.org/10.1145/1390156.1390191.

[25] J. Duchi and Y. Singer. “Efficient Online and Batch Learning Using Forward Backward Splitting”. In: *J. Mach. Learn. Res.* 10 (Dec. 2009), pp. 2899–2934. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1577069.1755882.

[26] B. Efron, T. J. Hastie, I. Johnstone, and R. Tibshirani. “Least angle regression”. In: *The Annals of Statistics* 32.2 (2004), pp. 407–499.

[27] D. Endler. “Intrusion detection. Applying machine learning to Solaris audit data”. In: *Proceedings 14th Annual Computer Security Applications Conference (Cat. No.98EX217)*. Dec. 1998, pp. 268–279. DOI: 10.1109/CSAC.1998.738647.

[28] J. Fan and R. Li. “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties”. In: *Journal of the American Statistical Association* 96.456 (2001), pp. 1348–1360. DOI: 10.1198/016214501753382273.

[29] T. Fawcett and F. Provost. “Activity Monitoring: Noticing Interesting Changes in Behavior”. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’99. San Diego, California, USA: ACM, 1999, pp. 53–62. ISBN: 1-58113-143-7. DOI: 10.1145/312129.312195. URL: http://doi.acm.org/10.1145/312129.312195.

[30] K. Fernandes, P. Vinagre, and P. Cortez. “A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News”. In: *Progress in Artificial Intelligence*. Ed. by Francisco Pereira, Penousal Machado, Ernesto Costa, and Amílcar Cardoso. Cham: Springer International Publishing, 2015, pp. 535–546. ISBN: 978-3-319-23485-4.

[31] C. Fraley and T. Hesterberg. “Least Angle Regression and LASSO for Large Datasets”. In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 1.4 (2009), pp. 251–259. DOI: 10.1002/sam.10021. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.10021. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.10021.

[32] J. H. Friedman, T. J. Hastie, H. Höfling, and R. Tibshirani. “Pathwise coordinate optimization”. In: *The Annals of Applied Statistics* 1.2 (2007), pp. 302–332.

[33] J. H. Friedman, T. J. Hastie, and R. Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” In: *Journal of statistical software* 33 1 (2010), pp. 1–22.

[34] H. Fujisawa. *Robust Statistics (in Japanese)*. Kindaikagakusha, 2017.

[35] H. Fujisawa and S. Eguchi. “Robust Parameter Estimation with a Small Bias Against Heavy Contamination”. In: *Journal of Multivariate Analysis* 99.9 (2008), pp. 2053–2081.

[36] D. Gervini and V. J. Yohai. “A class of robust and fully efficient regression estimators”. In: *Ann. Statist.* 30.2 (Apr. 2002), pp. 583–616. DOI: 10.1214/aos/1021379866. URL: https://doi.org/10.1214/aos/1021379866.

[37] S. Ghadimi and G. Lan. “Accelerated gradient methods for nonconvex nonlinear and stochastic programming”. In: *Mathematical Programming* 156.1 (Mar. 2016), pp. 59–99. ISSN: 1436-4646. DOI: 10.1007/s10107-015-0871-8. URL: https://doi.org/10.1007/s10107-015-0871-8.

[38] S. Ghadimi, G. Lan, and H. Zhang. “Mini-batch Stochastic Approximation Methods for Nonconvex Stochastic Composite Optimization”. In: *Math. Program.* 155.1-2 (Jan. 2016), pp. 267–305. ISSN: 0025-5610. DOI: 10.1007/s10107-014-0846-1. URL: http://dx.doi.org/10.1007/s10107-014-0846-1.

[39] A. Ghosh and A. Basu. “Robust estimation in generalized linear models: the density power divergence approach”. In: *TEST* 25.2 (June 2016), pp. 269–290. ISSN: 1863-8260. DOI: 10.1007/s11749-015-0445-3. URL: https://doi.org/10.1007/s11749-015-0445-3.

[40] A. Ghosh and S. Majumdar. “Ultrahigh-dimensional Robust and Efficient Sparse Regression using Non-Concave Penalized Density Power Divergence”. In: *arXiv preprint arXiv:1802.04906* (2018).

[41] E. Greenshtein and Y. Ritov. “Persistence in high-dimensional linear predictor selection and the virtue of overparametrization”. In: *Bernoulli* 10.6 (Dec. 2004), pp. 971–988. DOI: 10.3150/bj/1106314846. URL: https://doi.org/10.3150/bj/1106314846.

[42] F. E. Grubbs. “Procedures for Detecting Outlying Observations in Samples”. In: *Technometrics* 11.1 (1969), pp. 1–21. DOI: 10.1080/00401706.1969.10490657.

[43] F. R. Hampel. “The Influence Curve and Its Role in Robust Estimation”. In: *Journal of the American Statistical Association* 69.346 (1974), pp. 383–393. ISSN: 01621459. URL: http://www.jstor.org/stable/2285666.

[44] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics. The Approach Based on Influence Functions*. Wiley, 1986.

[45] T. J. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2010.

[46] T. J. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015. ISBN: 1498712169, 9781498712163.

[47] L. L. Ho, C. J. Macey, and R. Hiller. “A Distributed and Reliable Platform for Adaptive Anomaly Detection in IP Networks”. In: *Active Technologies for Network and Service Management*. Ed. by Rolf Stadler and Burkhard Stiller. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 33–46. ISBN: 978-3-540-48100-3.

[48] A. E. Hoerl and R. W. Kennard. “Ridge regression: Biased estimation for nonorthogonal problems”. In: *Technometrics* 12.1 (1970), pp. 55–67.

[49] H. Hotelling. “The Generalization of Student’s Ratio”. In: *Ann. Math. Statist.* 2.3 (Aug. 1931), pp. 360–378. DOI: 10.1214/aoms/1177732979. URL: https://doi.org/10.1214/aoms/1177732979.

[50] P. J. Huber. “Robust Estimation of a Location Parameter”. In: *Ann. Math. Statist.* 35.1 (Mar. 1964), pp. 73–101. DOI: 10.1214/aoms/1177703732. URL: https://doi.org/10.1214/aoms/1177703732.

[51] H. Hung, Z. Jou, and S. Huang. “Robust mislabel logistic regression without modeling mislabel probabilities”. In: *Biometrics* (2017), n/a–n/a. ISSN: 1541-0420. DOI: 10.1111/biom.12726. URL: http://dx.doi.org/10.1111/biom.12726.

[52] D. R. Hunter and K. Lange. “A tutorial on MM algorithms”. In: *The American Statistician* 58.1 (2004), pp. 30–37.

[53] M. C. Jones, N. L. Hjort, I. R. Harris, and A. Basu. "A Comparison of related density-based minimum divergence estimators". In: *Biometrika* 88.3 (2001), pp. 865–873.

[54] T. Kanamori and H. Fujisawa. "Robust estimation under heavy contamination using unnormalized models". In: *Biometrika* 102.3 (2015), pp. 559–572.

[55] T. Kawashima and H. Fujisawa. "On Difference Between Two Types of γ-divergence for Regression". In: *ArXiv e-prints* (May 2018). arXiv: 1805.06144 [math.ST].

[56] T. Kawashima and H. Fujisawa. "Robust and Sparse Regression in GLM by Stochastic Optimization". In: *ArXiv e-prints* (Feb. 2018). arXiv: 1802.03127 [stat.ML].

[57] T Kawashima and H Fujisawa. "Robust and Sparse Regression via γ-Divergence". In: *Entropy* 19.608 (2017). ISSN: 1099-4300. DOI: 10.3390/e19110608. URL: http://www.mdpi.com/1099-4300/19/11/608.

[58] J. A. Khan, S. Van Aelst, and R. H. Zamar. "Robust linear model selection based on least angle regression". In: *Journal of the American Statistical Association* 102.480 (2007), pp. 1289–1299.

[59] J. Kivinen and M. K. Warmuth. "Exponentiated Gradient Versus Gradient Descent for Linear Predictors". In: *Information and Computation* 132 (1995).

[60] K. Knight and W. Fu. "Asymptotics for lasso-type estimators". In: *Ann. Statist.* 28.5 (Oct. 2000), pp. 1356–1378. DOI: 10.1214/aos/1015957397. URL: https://doi.org/10.1214/aos/1015957397.

[61] H. R. Künsch, L. A. Stefanski, and R. J. Carroll. "Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, with Applications to Generalized Linear Models". In: *Journal of the American Statistical Association* 84.406 (1989), pp. 460–466. DOI: 10.1080/01621459.1989.10478791.

[62] D. Lambert. "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing". In: *Technometrics* 34.1 (1992), pp. 1–14. ISSN: 00401706. URL: http://www.jstor.org/stable/1269547.

[63] J. Laurikkala, M. Juhola, and E. Kentala. "Informal Identification of Outliers in Medical Data". In: *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology* (July 2000).

[64] A. C. Lozano, N. Meinshausen, and E. Yang. "Minimum Distance Lasso for robust high-dimensional regression". In: *Electron. J. Statist.* 10.1 (2016), pp. 1296–1340. DOI: 10.1214/16-EJS1136. URL: https://doi.org/10.1214/16-EJS1136.

[65] J. Mairal. "Optimization with First-Order Surrogate Functions". In: *ICML 2013 - International Conference on Machine Learning*. Vol. 28. JMLR Proceedings. Atlanta, United States, June 2013, pp. 783–791. URL: https://hal.inria.fr/hal-00822229.

[66] R.A. Maronna, D.R. Martin, and V.J. Yohai. *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley, 2006.

[67] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989. ISBN: 9780412317606. URL: http://books.google.com/books?id=h9kFH2%5C_FfBkC.

[68] N. Meinshausen and P. Bühlmann. “High-dimensional graphs and variable selection with the Lasso”. In: *Ann. Statist.* 34.3 (June 2006), pp. 1436–1462. DOI: 10.1214/009053606000000281. URL: https://doi.org/10.1214/009053606000000281.

[69] N. Meinshausen and B. Yu. “Lasso-type recovery of sparse representations for high-dimensional data”. In: *Ann. Statist.* 37.1 (Feb. 2009), pp. 246–270. DOI: 10.1214/07-AOS582. URL: https://doi.org/10.1214/07-AOS582.

[70] Y. Nesterov. *Gradient methods for minimizing composite objective function.* CORE Discussion Papers 2007076. Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007. URL: https://EconPapers.repec.org/RePEc:cor:louvco:2007076.

[71] H. D. Nguyen. “An introduction to Majorization-Minimization algorithms for machine learning and statistical estimation”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7.2 (), e1198. DOI: 10.1002/widm.1198. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1198. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1198.

[72] N. Parikh and S. Boyd. “Proximal Algorithms”. In: *Found. Trends Optim.* 1.3 (Jan. 2014), pp. 127–239. ISSN: 2167-3888. DOI: 10.1561/2400000003. URL: http://dx.doi.org/10.1561/2400000003.

[73] D. Pregibon. “Logistic Regression Diagnostics”. In: *The Annals of Statistics* 9.4 (1981), pp. 705–724. ISSN: 00905364. URL: http://www.jstor.org/stable/2240841.

[74] R. T. Rockafellar. *Convex analysis.* Princeton Mathematical Series. Princeton, N. J.: Princeton University Press, 1970.

[75] P. J. Rousseeuw. “Least Median of Squares Regression”. In: *Journal of the American Statistical Association* 79.388 (1984), pp. 871–880. DOI: 10.1080/01621459.1984.10477105.

[76] P. J. Rousseeuw and C. Croux. “Alternatives to the Median Absolute Deviation”. In: *Journal of the American Statistical Association* 88.424 (1993), pp. 1273–1283. DOI: 10.1080/01621459.1993.10476408.

[77] P. J. Rousseeuw and K. Driessen. “Computing LTS Regression for Large Data Sets”. In: *Data Min. Knowl. Discov.* 12.1 (Jan. 2006), pp. 29–45. ISSN: 1384-5810. DOI: 10.1007/s10618-005-0024-4. URL: http://dx.doi.org/10.1007/s10618-005-0024-4.

[78] P. J. Rousseeuw and M. Hubert. “Robust statistics for outlier detection”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1 (2011), pp. 73–79. DOI: 10.1002/widm.2.

[79] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection.* Vol. 589. John wiley & sons, 2005.

[80] P. J. Rousseeuw and V. J. Yohai. “Robust Regression by Means of S-Estimators”. In: *Robust and Nonlinear Time Series Analysis.* Ed. by Jürgen Franke, Wolfgang Härdle, and Douglas Martin. New York, NY: Springer US, 1984, pp. 256–272. ISBN: 978-1-4615-7821-5.

[81] E. D. Schifano, R. L. Strawderman, and M. T. Wells. “Majorization-Minimization algorithms for nonsmoothly penalized objective functions”. In: *Electron. J. Statist.* 4 (2010), pp. 1258–1299. DOI: 10.1214/10-EJS582. URL: https://doi.org/10.1214/10-EJS582.

[82] D. W. Scott. “Parametric Statistical Modeling by Minimum Integrated Square Error”. In: *Technometrics* 43.3 (2001), pp. 274–285. ISSN: 00401706. URL: http://www.jstor.org/stable/1271214.

[83] Y. She and A. B. Owen. “Outlier Detection Using Nonconvex Penalized Regression”. In: *Journal of the American Statistical Association* 106.494 (2011), pp. 626–639. DOI: 10.1198/jasa.2011.tm10390. eprint: https://doi.org/10.1198/jasa.2011.tm10390. URL: https://doi.org/10.1198/jasa.2011.tm10390.

[84] W. A. Shewhart. “Economic Control of Quality of Manufactured Product”. In: *Bell System Technical Journal* 9 (Apr. 1930). DOI: 10.1002/j.1538-7305.1930.tb00373.x.

[85] A. J. Stromberg, O. Hossjer, and D. M. Hawkins. “The Least Trimmed Differences Regression Estimator and Alternatives”. In: *Journal of the American Statistical Association* 95.451 (2000), pp. 853–864. ISSN: 01621459. URL: http://www.jstor.org/stable/2669469.

[86] J. Tibshirani and C. D. Manning. “Robust Logistic Regression using Shift Parameters”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*. 2014, pp. 124–129. URL: http://aclweb.org/anthology/P/P14/P14-2021.pdf.

[87] R. Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B* (1996), pp. 267–288.

[88] R. Tibshirani. “THE LASSO METHOD FOR VARIABLE SELECTION IN THE COX MODEL”. In: *Statistics in Medicine* 16.4 (1997), pp. 385–395.

[89] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. “Sparsity and smoothness via the fused lasso”. In: *Journal of the Royal Statistical Society: Series B* 67.1 (2005), pp. 91–108.

[90] S. A. Van de Geer and P. Bühlmann. “On the conditions used to prove oracle results for the Lasso”. In: *Electron. J. Statist.* 3 (2009), pp. 1360–1392. DOI: 10.1214/09-EJS506. URL: https://doi.org/10.1214/09-EJS506.

[91] H. Wang and C. Leng. “Unified LASSO Estimation by Least Squares Approximation”. In: *Journal of the American Statistical Association* 102.479 (2007), pp. 1039–1048. DOI: 10.1198/016214507000000509.

[92] H. Wang, G. Li, and G. Jiang. “Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso”. In: *Journal of Business & Economic Statistics* 25 (2007), pp. 347–355. URL: https://EconPapers.repec.org/RePEc:bes:jnlbes:v:25:y:2007:p:347-355.

[93] M. P. Windham. “Robustifying model fitting”. In: *Journal of the Royal Statistical Society: Series B* 57.3 (1995), pp. 599–609.

[94] S. J. Wright. “Coordinate Descent Algorithms”. In: *Math. Program.* 151.1 (June 2015), pp. 3–34. ISSN: 0025-5610. DOI: 10.1007/s10107-015-0892-3. URL: http://dx.doi.org/10.1007/s10107-015-0892-3.

[95] T. T. Wu and K. Lange. “Coordinate descent algorithms for lasso penalized regression”. In: *Ann. Appl. Stat.* 2.1 (Mar. 2008), pp. 224–244. DOI: 10.1214/07-AOAS147. URL: https://doi.org/10.1214/07-AOAS147.

[96] L. Xiao. “Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization”. In: *Journal of Machine Learning Research* 11 (2010), pp. 2543–2596.

[97] N. Ye and Q. Chen. “An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems”. In: *Quality and Reliability Engineering International* 17.2 (2001), pp. 105–112. DOI: 10.1002/qre.392.

[98] M. Yuan and Y. Lin. “Model selection and estimation in regression with grouped variables”. In: *Journal of the Royal Statistical Society: Series B* 68.1 (2006), pp. 49–67.

[99] Y. Zang, Q. Zhao, Q. Zhang, Y. Li, S. Zhang, and S. Ma. “Inferring gene regulatory relationships with a high-dimensional robust approach”. In: *Genetic Epidemiology* 41.5 (2017), pp. 437–454. DOI: 10.1002/gepi.22047. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.22047. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22047.

[100] C.H. Zhang. “Nearly unbiased variable selection under minimax concave penalty”. In: *The Annals of Statistics* 38.2 (Apr. 2010), pp. 894–942. DOI: 10.1214/09-AOS729. URL: https://doi.org/10.1214/09-AOS729.

[101] T. Zhang. “Analysis of multi-stage convex relaxation for sparse regularization”. In: *Journal of Machine Learning Research* 11.Mar (2010), pp. 1081–1107.

[102] P. Zhao and B. Yu. “On Model Selection Consistency of Lasso”. In: *J. Mach. Learn. Res.* 7 (Dec. 2006), pp. 2541–2563. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1248547.1248637.

[103] H. Zou. “The Adaptive Lasso and Its Oracle Properties”. In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1418–1429. DOI: 10.1198/016214506000000735. eprint: https://doi.org/10.1198/016214506000000735. URL: https://doi.org/10.1198/016214506000000735.

[104] H. Zou and T. J. Hastie. “Regularization and variable selection via the Elastic Net”. In: *Journal of the Royal Statistical Society, Series B* 67 (2005), pp. 301–320.

Appendix A

Proof of Theorem 2.1.1

Here, we show the proof in the case of type II for simplicity. We can prove Properties (i), (ii) and (iii) even in the case of type I in a similar manner.

Proof of Theorem 2.1.1. For two non-negative functions $r(x,y)$ and $u(x,y)$ and probability density function $g(x)$, it follows from Hölder's inequality that:

$$\int r(x,y)u(x,y)g(x)dxdy \leq \left(\int r(x,y)^{\alpha}g(x)dxdy\right)^{\frac{1}{\alpha}}\left(\int u(x,y)^{\beta}g(x)dxdy\right)^{\frac{1}{\beta}},$$

where α and β are positive constants and $\frac{1}{\alpha}+\frac{1}{\beta}=1$. The equality holds if and only if $r(x,y)^{\alpha}=\tau u(x,y)^{\beta}$ for a positive constant τ. Let $r(x,y)=g(y|x)$, $u(x,y)=f(y|x)^{\gamma}$, $\alpha = 1+\gamma$ and $\beta=\frac{1+\gamma}{\gamma}$. Then, it holds that:

$$\begin{aligned}&\int\left(\int g(y|x)f(y|x)^{\gamma}dy\right)dg(x)\\ &\leq\left\{\int\left(\int g(y|x)^{1+\gamma}dy\right)dg(x)\right\}^{\frac{1}{1+\gamma}}\left\{\int\left(\int f(y|x)^{1+\gamma}dy\right)dg(x)\right\}^{\frac{\gamma}{1+\gamma}}.\end{aligned}$$

The equality holds if and only if $g(y|x)^{1+\gamma}=\tau(f(y|x)^{\gamma})^{\frac{1+\gamma}{\gamma}}$, i.e., $g(y|x)=f(y|x)$ because $g(y|x)$ and $f(y|x)$ are conditional probability density functions. Properties (i) and (ii) follow from this inequality, the equality condition and the definition of $D_{\gamma}(g(y|x),f(y|x);g(x))$.

Let us prove Property (iii). Suppose that γ is sufficiently small. Then, it holds that $f^{\gamma} = 1 + \gamma \log f + O(\gamma^2)$. The γ-divergence for regression is expressed by:

$$
\begin{aligned}
& D_{\gamma,2}(g(y|x), f(y|x); g(x)) \\
& = \frac{1}{\gamma(1+\gamma)} \log \int \left\{ \int g(y|x)(1+\gamma \log g(y|x)+O(\gamma^2))dy \right\} g(x)dx \\
& - \frac{1}{\gamma} \log \int \left\{ \int g(y|x)(1+\gamma \log f(y|x)+O(\gamma^2))dy \right\} g(x)dx \\
& + \frac{1}{1+\gamma} \log \int \left\{ \int f(y|x)(1+\gamma \log f(y|x)+O(\gamma^2))dy \right\} g(x)dx \\
& = \frac{1}{\gamma(1+\gamma)} \log \left\{ 1+\gamma \int \left(\int g(y|x) \log g(y|x)dy \right) g(x)dx + O(\gamma^2) \right\} \\
& - \frac{1}{\gamma} \log \left\{ 1+\gamma \int \left(\int g(y|x) \log f(y|x)dy \right) g(x)dx + O(\gamma^2) \right\} \\
& \frac{1}{1+\gamma} \log \left\{ 1+\gamma \int \left(\int f(y|x) \log f(y|x)dy \right) g(x)dx + O(\gamma^2) \right\} \\
& = \frac{1}{(1+\gamma)} \int \left(\int g(y|x) \log g(y|x)dy \right) g(x)dx \\
& - \int \left(\int g(y|x) \log f(y|x)dy \right) g(x)dx + O(\gamma) \\
& = \int D_{KL}(g(y|x), f(y|x))g(x)dx + O(\gamma).
\end{aligned}
$$

□

Appendix B

Some Proofs in Chapter 3

Proof of Theorem 3.3.1. We see that:

$$\begin{aligned}
&\int \left(\int g(y|x) f(y|x;\theta)^{\gamma} dy \right) g(x) dx \\
&= \int \left(\int \{(1-\varepsilon) f(y|x;\theta^{*}) + \varepsilon \delta(y|x)\} f(y|x;\theta)^{\gamma} dy \right) g(x) dx \\
&= (1-\varepsilon) \left\{ \int \left(\int f(y|x;\theta^{*}) f(y|x;\theta)^{\gamma} dy \right) g(x) dx \right\} \\
&\quad + \varepsilon \left\{ \int \left(\int \delta(y|x) f(y|x;\theta)^{\gamma} dy \right) g(x) dx \right\}.
\end{aligned}$$

It follows from the assumption $\varepsilon < \frac{1}{2}$ that:

$$\begin{aligned}
&\left\{ \varepsilon \int \left(\int \delta(y|x) f(y|x;\theta)^{\gamma} dy \right) g(x) dx \right\}^{\frac{1}{\gamma}} \\
&< \left\{ \frac{1}{2} \int \left(\int \delta(y|x) f(y|x;\theta)^{\gamma} dy \right) g(x) dx \right\}^{\frac{1}{\gamma}} \\
&< \left\{ \int \left(\int \delta(y|x) f(y|x;\theta)^{\gamma} dy \right) g(x) dx \right\}^{\frac{1}{\gamma}} = \nu_{f_{\theta},\gamma}.
\end{aligned}$$

Hence,

$$\begin{aligned}&\int\left(\int g(y|x)f(y|x;\theta)^{\gamma}dy\right)g(x)dx=\\&(1-\varepsilon)\left\{\int\left(\int f(y|x;\theta^{*})f(y|x;\theta)^{\gamma}dy\right)g(x)dx\right\}\\&\qquad+O\left(\nu_{f_\theta,\gamma}^{\gamma}\right).\end{aligned}$$

Therefore, it holds that:

$$\begin{aligned}&d_{\gamma,2}(g(y|x),f(y|x;\theta);g(x))\\&=-\frac{1}{\gamma}\log\int\left(\int g(y|x)f(y|x;\theta)^{\gamma}dy\right)g(x)dx\\&\quad+\frac{1}{1+\gamma}\log\int\left(\int f(y|x;\theta)^{1+\gamma}dy\right)g(x)dx\\&=-\frac{1}{\gamma}\log\int\left(\int f(y|x;\theta^{*})f(y|x;\theta)^{\gamma}dy\right)g(x)dx\\&\quad+\frac{1}{1+\gamma}\log\int\left(\int f(y|x;\theta)^{1+\gamma}dy\right)g(x)dx\\&\quad-\frac{1}{\gamma}\log(1-\varepsilon)+O\left(\nu_{f_\theta,\gamma}^{\gamma}\right)\\&=d_{\gamma,2}(f(y|x;\theta^{*}),f(y|x;\theta);g(x))\\&\quad-\frac{1}{\gamma}\log(1-\varepsilon)+O\left(\nu_{f_\theta,\gamma}^{\gamma}\right).\end{aligned}$$

Then, it follows that:

$$\begin{aligned}&D_{\gamma,2}(g(y|x),f(y|x;\theta);g(x))-D_{\gamma,2}(g(y|x),f(y|x;\theta^{*});g(x))\\&\quad-D_{\gamma,2}(f(y|x;\theta^{*}),f(y|x;\theta);g(x))\\&=\left\{-d_{\gamma,2}(g(y|x),g(y|x);g(x))+d_{\gamma,2}(g(y|x),f(y|x;\theta);g(x))\right\}\\&\quad-\left\{-d_{\gamma,2}(g(y|x),g(y|x);g(x))+d_{\gamma,2}(g(y|x),f(y|x;\theta^{*});g(x))\right\}\\&\quad-\left\{-d_{\gamma,2}(f(y|x;\theta^{*}),f(y|x;\theta^{*});g(x))+d_{\gamma,2}(f(y|x;\theta^{*}),f(y|x;\theta);g(x))\right\}\\&=d_{\gamma,2}(g(y|x),f(y|x;\theta);g(x))-d_{\gamma,2}(f(y|x;\theta^{*}),f(y|x;\theta);g(x))\\&-d_{\gamma,2}(g(y|x),f(y|x;\theta^{*});g(x))+d_{\gamma,2}(f(y|x;\theta^{*}),f(y|x;\theta^{*});g(x))\\&=O(\nu^{\gamma}).\end{aligned}$$

□

Proof of Theorem 3.3.2. We see that:

$$\begin{aligned}
&\int \left(\int g(y|x) f(y|x;\theta)^{\gamma} dy \right) g(x) dx \\
&= \left\{ \int \left(\int f(y|x;\theta^{*}) f(y|x;\theta)^{\gamma} dy \right) (1-\varepsilon(x)) g(x) dx \right. \\
&\qquad \left. + \int \left(\int \delta(y|x) f(y|x;\theta)^{\gamma} dy \right) \varepsilon(x) g(x) dx \right\}.
\end{aligned}$$

It follows from the assumption $\varepsilon(x) < \frac{1}{2}$ that:

$$\begin{aligned}
&\left\{ \int \left(\int \delta(y|x) f(y|x;\theta)^{\gamma} dy \right) \varepsilon(x) g(x) dx \right\}^{\frac{1}{\gamma}} \\
&< \left\{ \int \left(\int \delta(y|x) f(y|x;\theta)^{\gamma} dy \right) \frac{g(x)}{2} dx \right\}^{\frac{1}{\gamma}} \\
&< \left\{ \int \left(\int \delta(y|x) f(y|x;\theta)^{\gamma} dy \right) g(x) dx \right\}^{\frac{1}{\gamma}} = \nu_{f_{\theta},\gamma}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\int \left(\int g(y|x) f(y|x;\theta)^{\gamma} dy \right) g(x) dx \\
&= \left\{ \int \left(\int f(y|x;\theta^{*}) f(y|x;\theta)^{\gamma} dy \right) (1-\varepsilon(x)) g(x) dx \right\} \\
&\qquad + O(\nu_{f_{\theta},\gamma}^{\gamma}).
\end{aligned}$$

Therefore, it holds that:

$$
\begin{aligned}
&d_{\gamma,2}(g(y|x),f(y|x;\theta);g(x))\\
&=-\frac{1}{\gamma}\log\int\left(\int g(y|x)f(y|x;\theta)^{\gamma}dy\right)g(x)dx\\
&\quad+\frac{1}{1+\gamma}\log\int\left(\int f(y|x;\theta)^{1+\gamma}dy\right)g(x)dx\\
&=-\frac{1}{\gamma}\log\left\{\int\left(\int f(y|x;\theta^*)f(y|x;\theta)^{\gamma}dy\right)(1-\varepsilon(x))g(x)dx\right\}\\
&\qquad+O(\nu_{f_\theta,\gamma}^{\gamma})+\frac{1}{1+\gamma}\log\int\left(\int f(y|x;\theta)^{1+\gamma}dy\right)g(x)dx\\
&=d_{\gamma,2}(f(y|x;\theta^*),f(y|x;\theta);(1-\varepsilon(x))g(x))+O(\nu_{f_\theta,\gamma}^{\gamma})\\
&\quad-\frac{1}{1+\gamma}\log\int\left(\int f(y|x;\theta)^{1+\gamma}dy\right)(1-\varepsilon(x))g(x)dx\\
&\qquad+\frac{1}{1+\gamma}\log\int\left(\int f(y|x;\theta)^{1+\gamma}dy\right)g(x)dx\\
&=d_{\gamma,2}(f(y|x;\theta^*),f(y|x;\theta);(1-\varepsilon(x))g(x))\\
&\quad+O(\nu_{f_\theta,\gamma}^{\gamma})-\frac{1}{1+\gamma}\log\left\{1-\int\varepsilon(x)g(x)dx\right\}.
\end{aligned}
$$

Then, it follows that:

$$
\begin{aligned}
&D_{\gamma,2}(g(y|x),f(y|x;\theta);g(x))\\
&-D_{\gamma,2}(g(y|x),f(y|x;\theta^*);g(x))\\
&-D_{\gamma},2(f(y|x;\theta^*),f(y|x;\theta);(1-\varepsilon(x))g(x))\\
&=\{-d_{\gamma,2}(g(y|x),g(y|x);g(x))+d_{\gamma,2}(g(y|x),f(y|x;\theta);g(x))\}\\
&-\{-d_{\gamma,2}(g(y|x),g(y|x);g(x))+d_{\gamma,2}(g(y|x),f(y|x;\theta^*);g(x))\}\\
&-\{-d_{\gamma,2}(f(y|x;\theta^*),f(y|x;\theta^*);(1-\varepsilon(x))g(x))\\
&\qquad\qquad+d_{\gamma,2}(f(y|x;\theta^*),f(y|x;\theta);(1-\varepsilon(x))g(x))\}\\
&=d_{\gamma,2}(g(y|x),f(y|x;\theta);g(x))\\
&\quad-d_{\gamma,2}(f(y|x;\theta^*),f(y|x;\theta);(1-\varepsilon(x))g(x))\\
&\quad-d_{\gamma,2}(g(y|x),f(y|x;\theta^*);g(x))\\
&\quad+d_{\gamma,2}(f(y|x;\theta^*),f(y|x;\theta^*);(1-\varepsilon(x))g(x))\\
&=O(\nu^{\gamma}).
\end{aligned}
$$

□

Appendix C

Proof of Theorem 4.4.2

Proof of Theorem 4.4.2.

$$\begin{aligned}
&\lim_{k\downarrow 0}\frac{\Psi(\theta^{(R)}+k\delta)-\Psi(\theta^{(R)})}{k}\\
&=\lim_{k\downarrow 0}\frac{E_{(x,y)}\left[l((x,y);\theta^{(R)}+k\delta)\right]-E_{(x,y)}\left[l((x,y);\theta^{(R)})\right]+\lambda P(\theta^{(R)}+k\delta)-\lambda P(\theta^{(R)})}{k}\\
&=\lim_{k\downarrow 0}\frac{E_{(x,y)}\left[l((x,y);\theta^{(R)}+k\delta)\right]-E_{(x,y)}\left[l((x,y);\theta^{(R)})\right]}{k}\\
&\quad+\lim_{k\downarrow 0}\frac{\lambda P(\theta^{(R)}+k\delta)-\lambda P(\theta^{(R)})}{k}. \qquad \text{(C.1)}
\end{aligned}$$

The directional derivative of the differentiable function always exist and is represented by the dot product with the gradient of the differentiable function and the direction given by

$$\begin{aligned}
&\lim_{k\downarrow 0}\frac{E_{(x,y)}\left[l((x,y);\theta^{(R)}+k\delta)\right]-E_{(x,y)}\left[l((x,y);\theta^{(R)})\right]}{k}\\
&=\left\langle \nabla E_{(x,y)}\left[l((x,y);\theta^{(R)})\right],\delta\right\rangle. \qquad \text{(C.2)}
\end{aligned}$$

Moreover, the directional derivative of the (proper) convex function exists at the relative interior point of the domain and is greater than the dot product with the subgradient of the convex function and direction [74] given by

$$\begin{aligned}
\lim_{k\downarrow 0}\frac{\lambda P(\theta^{(R)}+k\delta)-\lambda P(\theta^{(R)})}{k} &= \sup_{g\in\partial P(\theta^{(R)})}\lambda\langle g,\delta\rangle\\
&\geq \lambda\langle g,\delta\rangle\ \ for\ any\ g\in\partial P(\theta^{(R)}). \qquad \text{(C.3)}
\end{aligned}$$

Then, by the optimality condition of (4.16), we have the following equation

$$\begin{aligned}
&0 \in \nabla E_{(x,y)}\left[l((x,y);\theta^{(R)})\right] + \lambda \partial P(\theta^{+}) + \frac{1}{\eta_R}\left\{\nabla w\left(\theta^{+}\right) - \nabla w\left(\theta^{(R)}\right)\right\} \\
&\frac{1}{\eta_R}\left\{\nabla w\left(\theta^{(R)}\right) - \nabla w\left(\theta^{+}\right)\right\} \in \nabla E_{(x,y)}\left[l((x,y);\theta^{(R)})\right] + \lambda \partial P(\theta^{+}). \qquad \text{(C.4)}
\end{aligned}$$

Therefore, we can obtain (4.19) from $P_{X,R} \approx 0$, (C.1), (C.2), (C.3) and (C.4) as follows;

$$\begin{aligned}
&\lim_{k \downarrow 0} \frac{E_{(x,y)}\left[l((x,y);\theta^{(R)} + k\delta)\right] - E_{(x,y)}\left[l((x,y);\theta^{(R)})\right]}{k} \\
&\quad + \lim_{k \downarrow 0} \frac{\lambda P(\theta^{(R)} + k\delta) - \lambda P(\theta^{(R)})}{k} \\
&\geq \left\langle \nabla E_{(x,y)}\left[l((x,y);\theta^{(R)})\right], \delta \right\rangle + \lambda \left\langle g, \delta \right\rangle \quad for\ any\ g \in \partial P(\theta^{(R)}) \\
&= \left\langle \nabla E_{(x,y)}\left[l((x,y);\theta^{(R)})\right] + \lambda g, \delta \right\rangle \quad for\ any\ g \in \partial P(\theta^{(R)}) \\
&\ni 0.
\end{aligned}$$

□