

A Robust Optimization Approach to Statistical Estimation Problems

by

Apostolos G. Fertis

Submitted to the Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the


MASSACHUSETTS INSTITUTE OF TECHNOLOGY

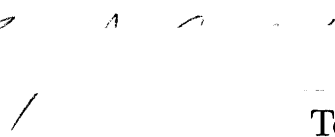
June 2009

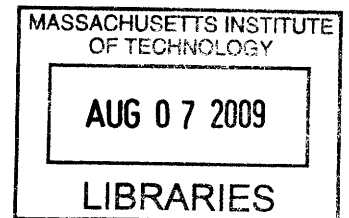
© Massachusetts Institute of Technology 2009. All rights reserved.

ARCHIVES

1 A /
Author
Department of Electrical Engineering and Computer Science
May 19, 2009

Certified by

Dimitris J. Bertsimas
Boeing Professor of Operations Research
Sloan School of Management
Thesis Supervisor

→ A A →
Accepted by

Terry P. Orlando
Chairman, Department Committee on Graduate Students



A Robust Optimization Approach to Statistical Estimation Problems

by

Apostolos G. Fertis

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

There have long been intuitive connections between robustness and regularization in statistical estimation, for example, in lasso and support vector machines. In the first part of the thesis, we formalize these connections using robust optimization. Specifically

- (a) We show that in classical regression, regularized estimators like lasso can be derived by applying robust optimization to the classical least squares problem. We discover the explicit connection between the size and the structure of the uncertainty set used in the robust estimator, with the coefficient and the kind of norm used in regularization. We compare the out-of-sample performance of the nominal and the robust estimators in computer generated and real data.
- (b) We prove that the support vector machines estimator is also a robust estimator of some nominal classification estimator (this last fact was also observed independently and simultaneously by Xu, Caramanis, and Mannor [52]). We generalize the support vector machines estimator by considering several sizes and structures for the uncertainty sets, and proving that the respective max-min optimization problems can be expressed as regularization problems.

In the second part of the thesis, we turn our attention to constructing robust maximum likelihood estimators. Specifically

- (a) We define robust estimators for the logistic regression model, taking into consideration uncertainty in the independent variables, in the response variable, and in both. We consider several structures for the uncertainty sets, and prove that, in all cases, they lead to convex optimization problems. We provide efficient algorithms to compute the estimates in all cases. We report on the out-of-sample performance of the robust, as well as the nominal estimators in both computer generated and real data sets, and conclude that the robust estimators achieve a higher success rate.
- (b) We develop a robust maximum likelihood estimator for the multivariate normal distribution by considering uncertainty sets for the data used to produce it. We develop an efficient first order gradient descent method to compute the estimate and compare the

efficiency of the robust estimate to the respective nominal one in computer generated data.

Thesis Supervisor: Dimitris J. Bertsimas
Title: Boeing Professor of Operations Research
Sloan School of Management

Acknowledgements

I would like to express my gratitude for my supervisor Professor Dimitris Bertsimas for inspiring me throughout my PhD research. He has been very helpful in identifying challenging problems, providing his insight and encouraging my endeavors, even in difficult situations. Furthermore, I would like to thank Professors Pablo Parrilo, Georgia Perakis and John Tsitsiklis, who, as members of my PhD committee, aided me with experienced comments and thoughtful observations. I would also like to thank Dr. Omid Nohadani for supporting me with his cooperation and advice, and for providing his very wise views in our long conversations. Finally, I am very grateful to my family, my father George, my mother Lori, and my sister Andriani, as well as my friends, for believing faithfully in my dreams and always providing me with invaluable support to overcome any obstacle I met.

Contents

1	Introduction	13
1.1	Robust Optimization in Statistical Estimation Problems	14
1.2	Contributions of the Thesis	15
2	Equivalence of Robust Regression and Regularized Regression	17
2.1	Introduction	17
2.2	Robust Regression	18
2.3	Support Vector Machines for Regression	23
2.4	Experimental Results	26
2.5	Conclusions	31
3	Support Vector Machines as Robust Estimators	33
3.1	Introduction	33
3.2	Robust Properties of Support Vector Machines	34
3.3	Experimental Results	39
3.4	Conclusions	40
4	Robust Logistic Regression	43
4.1	Logistic Regression	43
4.2	Robust logistic regression under independent variables uncertainty	44
4.3	Robust logistic regression under response variable uncertainty	47
4.4	Globally robust logistic regression	49
4.5	Experimental Results	55
4.5.1	Artificial Data Sets	55
4.5.2	Real Data Sets	56
4.6	Conclusions	60

5	Robust Maximum Likelihood In Normally Distributed Data	61
5.1	Introduction	61
5.2	Method	61
5.3	Experiments	68
5.3.1	Worst-Case and Average Probability Density	69
5.3.2	Distance From the Nominal Estimator	70
5.3.3	Comparison of the Error Distributions	71
5.4	Conclusions	71
A	Algebraic Propositions	73
A.1	Properties of function $\mathbf{f}(\mathbf{x}, p)$	73
A.2	Propositions on matrix norms	74
B	Robust Logistic Regression Algorithms	77
B.1	Robust logistic regression under independent variables uncertainty solution algorithm	77
B.2	Robust logistic regression under response variable uncertainty solution algorithm	78
B.3	Globally robust logistic regression solution algorithm	79
C	The partial derivatives of $Z_1(\boldsymbol{\beta}, \beta_0)$ and $Z_3(\boldsymbol{\beta}, \beta_0)$	81

List of Figures

2-1	The average mean absolute error of the regression estimates according to ρ .	28
2-2	The average mean squared error of the regression estimates according to ρ .	29
3-1	The average classification error of the classification estimates according to ρ .	40
4-1	Success rate of the robust logistic regression under independent variables uncertainty estimate in the testing set.	56
4-2	Success rate of the robust logistic regression under response variable uncertainty estimate in the testing set.	57
4-3	Success rate of the globally robust logistic regression estimate in the testing set ($\Gamma = 1$).	57
4-4	Success rate of the globally robust logistic regression estimate in the testing set ($\Gamma = 2$).	58
5-1	Worst-case (left) and average (right) value of ψ , normal errors	69
5-2	Error in μ (left) and Σ (right), normal errors	71
5-3	Error in μ (left) and Σ (right), uniform errors	72

List of Tables

2.1	Sizes of real data sets for regression.	29
2.2	Mean absolute error in testing set for real data sets. * denotes the estimate with the best performance.	30
2.3	Mean squared error in testing set for real data sets. * denotes the estimate with the best performance.	31
3.1	Sizes of real data sets for classification.	41
3.2	Classification error in testing set for real data sets. * denotes the estimate with the best performance.	41
4.1	Sizes of real data sets.	59
4.2	Success rate in testing set for real data sets. * denotes the estimate with the best performance.	60

Chapter 1

Introduction

Statistical estimation has a long and distinguished history. In the context of regression, the idea of least squares has been used extensively. More generally, the established paradigm is to use the maximum likelihood principles.

Researchers soon realized that data on which these estimators are based are subject to error. The origins of the errors can be multiple (measurement, reporting, even classification (a clinical trial can be classified as success while it can be indeed a failure)). In order to deal with errors several researchers have introduced regularization methods:

- a) Regularized regression. Given a set of observations (y_i, \mathbf{x}_i) , $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^m$, $i \in \{1, 2, \dots, n\}$, Tibshirani [49] defines the Least Absolute Shrinkage and Selection Operator (lasso) estimate as the optimal solution to the optimization problem

$$\min_{\beta_0, \boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0 \mathbf{1}\|_2^2 + \rho \|\boldsymbol{\beta}\|_1, \quad (1.1)$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}, \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \left. \vphantom{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} \right\} n \text{ entries}, \quad (1.2)$$

and $\rho \geq 0$. Tibshirani [49] demonstrated using simulation results that the lasso estimate tends to have small support which yields more adaptive models and higher empirical success. Tikhonov and Arsenin proposed the ridge regression estimate, which is the

solution to

$$\min_{\beta_0, \boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0 \mathbf{1}\|_2^2 + \rho \|\boldsymbol{\beta}\|_2^2, \quad (1.3)$$

see [50].

- b) Support vector machines in classification problems, introduced by Vapnik et al. [10]. Given a set of data (y_i, \mathbf{x}_i) , $i \in \{1, 2, \dots, n\}$, $y_i \in \{-1, 1\}$, $\mathbf{x}_i \in \mathbb{R}^m$, the support vector machines estimate is the optimal solution to optimization problem

$$\begin{aligned} \min_{\boldsymbol{\beta}, \beta_0, \boldsymbol{\xi}} \quad & \|\boldsymbol{\beta}\|_2 + \rho \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & \xi_i \geq 1 - y_i(\boldsymbol{\beta}' \mathbf{x}_i + \beta_0), i \in \{1, 2, \dots, n\} \\ & \xi_i \geq 0, i \in \{1, 2, \dots, n\}, \end{aligned} \quad (1.4)$$

where $\rho \geq 0$. Support vector machines classifiers have been very successful in experiments (see Schölkopf [43]). Note that Problem (1.4) has a regularization term $\|\boldsymbol{\beta}\|_2$ in its objective.

Huber [30] considers any statistical estimator T to be a functional defined on the space of probability measures. An estimator is called robust, if functional T is continuous in a neighborhood around the true distribution of the data. Hampel defined the influence curve to quantify the robustness of statistical estimators [26]. However, they did not provide an algorithmic way to construct robust estimators.

1.1 Robust Optimization in Statistical Estimation Problems

In this thesis, we use the paradigm of Robust Optimization to design estimators that are immune to data errors. Before describing the specific contributions of the thesis, let us give a brief overview of Robust Optimization.

Robust optimization has been increasingly used in mathematical programming as an effective way to immunize solutions against data uncertainty. If the data of a problem is not equal to its nominal value, the optimal solution calculated using the contaminated data might not be optimal or even feasible using the true values of the data. Robust optimization considers uncertainty sets for the data of the problem and aims to calculate solutions that

are immune to such uncertainty. In general, consider optimization problem

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}; \mathbf{d})$$

with decision variable \mathbf{x} restricted in feasible set \mathcal{X} and data \mathbf{d} . The robust version of it, if we consider uncertainty set \mathcal{D} for the errors $\Delta \mathbf{d}$ in the data, is

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\Delta \mathbf{d} \in \mathcal{D}} f(\mathbf{x}; \mathbf{d} + \Delta \mathbf{d}).$$

As we observe, the efficient solution of the nominal problem does not guarantee the efficient solution of the respective robust one.

There are many ways to define uncertainty sets. Soyster [47] considers uncertainty sets in linear optimization problems where each column of the data belongs to a convex set. Ben-Tal and Nemirovski [4], [5], [6], follow a less conservative approach by considering uncertain linear optimization problems with ellipsoidal uncertainty sets and computing robust counterparts, which constitute conic quadratic problems. Bertsimas and Sim [8], [9], consider an uncertainty set for linear or integer optimization problems where the number of coefficients in each constraint subject to error is bounded by some parameter adjusting the level of conservatism, and prove that this problem has an equivalent linear or integer optimization formulation, respectively.

The idea of using robust optimization to define estimators that tackle the uncertainty of the statistical data has already been explored. El Ghaoui and Lebret [20] have used robust optimization to deal with errors in the regression data. They define the robust total least squares problem, where the Frobenious norm of the matrix consisting of the independent variables and the response variable of the observations is bounded by some parameter, and prove that it can be formulated as a second order cone problem.

1.2 Contributions of the Thesis

In this thesis, we apply robust optimization principles to many classical statistical estimation problems to define the respective robust estimators, that deal with errors in the statistical data used to produce them. We study the properties of the estimators, as well as their connection with the uncertainty set used to define them. We develop efficient algorithms that compute the robust estimators, and test their prediction accuracy on computer generated as well as real data.

In the first part of the thesis, we formalize the connections between robustness and regularization in statistical estimation using robust optimization. Specifically

- (a) We show that in classical regression, regularized estimators like lasso can be derived by applying robust optimization to the classical least squares problem. We discover the explicit connection between the size and the structure of the uncertainty set used in the robust estimator, with the coefficient and the kind of norm used in regularization. We compare the out-of-sample performance of the nominal and the robust estimators in computer generated and real data.
- (b) We prove that the support vector machines estimator is also a robust estimator of some nominal classification estimator (this last fact was also observed independently and simultaneously by Xu, Caramanis, and Mannor [52]). We generalize the support vector machines estimator by considering several sizes and structures for the uncertainty sets, and proving that the respective max-min optimization problems can be expressed as regularization problems.

In the second part of the thesis, we turn our attention to constructing robust maximum likelihood estimators. Specifically

- (a) We define robust estimators for the logistic regression model, taking into consideration uncertainty in the independent variables, in the response variable, and in both. We consider several structures for the uncertainty sets, and prove that, in all cases, they lead to convex optimization problems. We provide efficient algorithms to compute the estimates in all cases. We report on the out-of-sample performance of the robust, as well as the nominal estimators in both computer generated and real data sets, and conclude that the robust estimators achieve a higher success rate.
- (b) We develop a robust maximum likelihood estimator for the multivariate normal distribution by considering uncertainty sets for the data used to produce it. We develop an efficient first order gradient descent method to compute the estimate and compare the efficiency of the robust estimate to the respective nominal one in computer generated data.

The structure of the thesis is the following. In Chapter 2, the connection between robust regression and regularized regression is quantified and studied. In Chapter 3, the Support Vector Machines estimator is proved to be the robust estimator corresponding to some nominal classification estimator, and its properties are investigated. In Chapter 4, robust estimators for logistic regression are defined and calculated. In Chapter 5, a robust normal distribution estimator is defined, and an efficient algorithm to calculate it is developed.

Chapter 2

Equivalence of Robust Regression and Regularized Regression

2.1 Introduction

A way to improve the performance of the regression estimate is to impose a regularization term in the objective function of the optimization problem which defines it. Given a set of observations (y_i, \mathbf{x}_i) , $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^m$, $i \in \{1, 2, \dots, n\}$, Tibshirani [49] defines the Least Absolute Shrinkage and Selection Operator (lasso) estimate as the optimal solution to the optimization problem

$$\min_{\beta_0, \boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0 \mathbf{1}\|_2^2 + \rho \|\boldsymbol{\beta}\|_1, \quad (2.1)$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}, \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \left. \vphantom{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} \right\} n \text{ entries}, \quad (2.2)$$

and $\rho \geq 0$.

Tibshirani [49] demonstrated using simulation results that the lasso estimate tends to have small support which yields more adaptive models and higher empirical success. Candès and Plan [13] proved that if the coefficient vector $\boldsymbol{\beta}$ and the data matrix \mathbf{X} follow certain probability distributions then, lasso nearly selects the best subset of variables with non-zero coefficients.

The connection between Robust Optimization and Regularization has been explored in the past. El Ghaoui and Le Bret prove that the minimization of the worst-case least squares error can be formulated as a Tikhonov regularization procedure [20]. Golub et al proved that Tikhonov's regularization method can be expressed as a total least squares formulation, where both the coefficient matrix and the right-hand side are known to reside in some sets.

In this chapter, we prove that regularization and robust optimization are essentially equivalent, that is the application of the robust optimization paradigm in statistical estimation leads to regularized solutions. We investigate the nature of the regularized solutions as the uncertainty sets in robust optimization vary. We present empirical evidence that demonstrates that the application of robust optimization in statistics, which is equivalent to regularization, has an improved out-of-sample performance in both artificial and real data. We further investigate the effectiveness of different uncertainty sets and their corresponding regularizations. In summary, the key contribution of this section is that the strong empirical performance of regularized solutions, which we also observe in this paper, can be explained by the fact that the process of regularization immunizes the estimation from data uncertainty.

The structure of the chapter is as follows. In Section 2.2, we prove that the robust regression estimate for uncertainty sets of various kinds of norms can be expressed as a regularized regression problem, and we calculate the relation between the norms used to define the uncertainty sets and the norms used in regularization, as well as their coefficients. In Section 2.3, we prove that the optimization problem used to define the support vector machines regression estimate is the robust counterpart of the ϵ -insensitive regression estimate problem. In Section 2.4, we report on the improved out-of-sample performance of the robust and regularized estimates in comparison to the classical ones in the experiments we carried out on artificial and real data.

2.2 Robust Regression

Given a set of data (y_i, \mathbf{x}_i) , $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^m$, $i \in \{1, 2, \dots, n\}$, we consider the Robust L_p Regression optimization problem

$$\min_{\beta, \beta_0} \max_{\Delta \mathbf{X} \in \mathcal{N}} \|\mathbf{y} - (\mathbf{X} + \Delta \mathbf{X})\beta - \beta_0 \mathbf{1}\|_p, \quad (2.3)$$

where \mathbf{y} , \mathbf{X} , and $\mathbf{1}$ are defined in Eq. (2.2) and \mathcal{N} is the uncertainty set for $\Delta \mathbf{X}$.

The uncertainty sets for $\Delta \mathbf{X}$ are going to be defined by bounding various kinds of norms of matrix $\Delta \mathbf{X}$. There exist several matrix norms (see Golub and Van Loan [24]). For

example, norm $\|\bullet\|_{q,p}$ for an $n \times m$ matrix \mathbf{A} is defined by

$$\|\mathbf{A}\|_{q,p} \equiv \sup_{\mathbf{x} \in \mathbb{R}^m, \mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_q}, \quad q, p \geq 1,$$

see Golub and Van Loan [24], p. 56.

Note that

$$\|\mathbf{A}\|_{q,p} = \max_{\|\mathbf{x}\|_q=1} \|\mathbf{A}\mathbf{x}\|_p,$$

and that for some $\mathbf{x}^* \in \mathbb{R}^m$ with $\|\mathbf{x}^*\|_q = 1$, we have that

$$\|\mathbf{A}\|_{q,p} = \|\mathbf{A}\mathbf{x}^*\|_p,$$

see Golub and Van Loan [24], p. 56.

Moreover, we define the p -Frobenius norm which is also going to be used in defining uncertainty sets for $\Delta \mathbf{X}$.

Definition 1. *The p -Frobenius norm $\|\bullet\|_{p-F}$ of an $n \times m$ matrix \mathbf{A} is*

$$\|\mathbf{A}\|_{p-F} \equiv \left(\sum_{i=1}^n \sum_{j=1}^m |A_{i,j}|^p \right)^{1/p}.$$

Observe that for $p = 2$, we obtain the usual Frobenius norm.

The following theorem computes a robust counterpart for the robust optimization Problem (2.3) under the uncertainty sets

$$\mathcal{N}_1 = \{\Delta \mathbf{X} \in \mathbb{R}^{n \times m} \mid \|\Delta \mathbf{X}\|_{q,p} \leq \rho\}, \quad (2.4)$$

and

$$\mathcal{N}_2 = \{\Delta \mathbf{X} \in \mathbb{R}^{n \times m} \mid \|\Delta \mathbf{X}\|_{p-F} \leq \rho\}. \quad (2.5)$$

It proves that the robust regression problems under these uncertainty sets can be expressed as regression regularization problems.

Theorem 1.

(a) *Under uncertainty set \mathcal{N}_1 , Problem (2.3) is equivalent to problem*

$$\min_{\beta, \beta_0} \|\mathbf{y} - \mathbf{X}\beta - \beta_0 \mathbf{1}\|_p + \rho \|\beta\|_q. \quad (2.6)$$

(b) Under uncertainty set \mathcal{N}_2 , Problem (2.3) is equivalent to problem

$$\min_{\beta, \beta_0} \|\mathbf{y} - \mathbf{X}\beta - \beta_0\mathbf{1}\|_p + \rho\|\beta\|_{d(p)}, \quad (2.7)$$

where $\mathcal{N}_1, \mathcal{N}_2$ are defined in Eq. (2.4) and (2.5) respectively, and $d(p)$ is defined in Eq. (4.3).

Proof.

(a) The proof utilizes Appendix A.2 that performs certain somewhat tedious vector calculations. To prove the theorem, we shall first compute a bound for the objective function of the inner problem in Eq. (2.3) and then, calculate a $\Delta\mathbf{X}$ that achieves this bound.

Using the norm properties, we have that

$$\begin{aligned} \|\mathbf{y} - (\mathbf{X} + \Delta\mathbf{X})\beta - \beta_0\mathbf{1}\|_p &= \|\mathbf{y} - \mathbf{X}\beta - \beta_0\mathbf{1} - \Delta\mathbf{X}\beta\|_p \\ &\leq \|\mathbf{y} - \mathbf{X}\beta - \beta_0\mathbf{1}\|_p + \|\Delta\mathbf{X}\beta\|_p. \end{aligned}$$

From Golub and Van Loan [24], p. 56, we have that

$$\|\Delta\mathbf{X}\beta\|_p \leq \|\Delta\mathbf{X}\|_{q,p} \|\beta\|_q.$$

Thus, for $\|\Delta\mathbf{X}\|_{q,p} \leq \rho$,

$$\|\Delta\mathbf{X}\beta\|_p \leq \rho\|\beta\|_q,$$

and for any $\Delta\mathbf{X} \in \mathcal{N}_1$,

$$\|\mathbf{y} - (\mathbf{X} + \Delta\mathbf{X})\beta - \beta_0\mathbf{1}\|_p \leq \|\mathbf{y} - \mathbf{X}\beta - \beta_0\mathbf{1}\|_p + \rho\|\beta\|_q. \quad (2.8)$$

We next construct a solution $\Delta\mathbf{X}^\circ \in \mathcal{N}_1$ that achieves bound (2.8). Let

$$\Delta\mathbf{X}^\circ = \begin{cases} -\rho \frac{\mathbf{y} - \mathbf{X}\beta - \beta_0\mathbf{1}}{\|\mathbf{y} - \mathbf{X}\beta - \beta_0\mathbf{1}\|_p} [\mathbf{f}(\beta, q)]^T, & \text{if } \mathbf{y} - \mathbf{X}\beta - \beta_0\mathbf{1} \neq \mathbf{0}, \\ -\rho\mathbf{u}[\mathbf{f}(\beta, q)]^T, & \text{if } \mathbf{y} - \mathbf{X}\beta - \beta_0\mathbf{1} = \mathbf{0}, \end{cases} \quad (2.9)$$

where $\mathbf{f}(\mathbf{x}, p) \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^m$, $p \geq 1$, is defined in Eq. (A.1) in Appendix A.2, $\mathbf{u} \in \mathbb{R}^n$ is an arbitrary vector with $\|\mathbf{u}\|_p = 1$.

For $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1} \neq \mathbf{0}$:

$$\begin{aligned}
\|\mathbf{y} - (\mathbf{X} + \Delta\mathbf{X}^\circ)\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1} - \Delta\mathbf{X}^\circ\boldsymbol{\beta}\|_p \\
&= \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1} + \rho \frac{\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}}{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p} [\mathbf{f}(\boldsymbol{\beta}, q)]^T \boldsymbol{\beta} \right\|_p \\
&= \left\| (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}) \left(1 + \frac{\rho \|\boldsymbol{\beta}\|_q}{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p} \right) \right\|_p \quad ([\mathbf{f}(\boldsymbol{\beta}, q)]^T \boldsymbol{\beta} = \|\boldsymbol{\beta}\|_q) \\
&= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p + \rho \|\boldsymbol{\beta}\|_q.
\end{aligned}$$

Note that when $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1} = \mathbf{0}$, $\|\mathbf{y} - (\mathbf{X} + \Delta\mathbf{X}^\circ)\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p + \rho \|\boldsymbol{\beta}\|_q$ as well.

Moreover, using Propositions 1 and 3 in Appendix A.2, we have that if $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1} \neq \mathbf{0}$,

$$\|\Delta\mathbf{X}^\circ\|_{q,p} = \rho \left\| \frac{\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}}{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p} \right\|_p \|\mathbf{f}(\boldsymbol{\beta}, q)\|_{d(q)} = \rho,$$

and if $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1} = \mathbf{0}$,

$$\|\Delta\mathbf{X}^\circ\|_{q,p} = \rho \|\mathbf{u}\|_p \|\mathbf{f}(\boldsymbol{\beta}, q)\|_{d(q)} = \rho.$$

Thus, $\Delta\mathbf{X}^\circ \in \mathcal{N}_1$.

Consequently, we conclude that

$$\max_{\Delta\mathbf{X} \in \mathcal{N}_1} \|\mathbf{y} - (\mathbf{X} + \Delta\mathbf{X})\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p + \rho \|\boldsymbol{\beta}\|_q,$$

which proves the theorem.

(b) Following the same procedure, we conclude that

$$\|\mathbf{y} - (\mathbf{X} + \Delta\mathbf{X})\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p + \|\Delta\mathbf{X}\boldsymbol{\beta}\|_p.$$

From Golub and Van Loan [24], p. 56, we have that

$$\|\Delta\mathbf{X}\boldsymbol{\beta}\|_p \leq \|\Delta\mathbf{X}\|_{d(p),p} \|\boldsymbol{\beta}\|_{d(p)}.$$

Using Proposition 2 in Appendix A.2, we conclude that

$$\|\Delta\mathbf{X}\boldsymbol{\beta}\|_p \leq \|\Delta\mathbf{X}\|_{p-F} \|\boldsymbol{\beta}\|_{d(p)},$$

and thus, for any $\Delta\mathbf{X} \in \mathcal{N}_2$,

$$\|\mathbf{y} - (\mathbf{X} + \Delta\mathbf{X})\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p + \rho\|\boldsymbol{\beta}\|_{d(p)}.$$

Define $\Delta\mathbf{X}^\circ$ as in Eq. (2.9) with $q = d(p)$. Then,

$$\|\mathbf{y} - (\mathbf{X} + \Delta\mathbf{X}^\circ)\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p + \rho\|\boldsymbol{\beta}\|_{d(p)}.$$

Furthermore, using Propositions 1 and 4 in Appendix A.2, we have that if $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1} \neq \mathbf{0}$,

$$\|\Delta\mathbf{X}^\circ\|_{p-F} = \rho \left\| \frac{\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}}{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p} \right\|_p \|\mathbf{f}(\boldsymbol{\beta}, d(p))\|_p = \rho,$$

and if $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1} = \mathbf{0}$,

$$\|\Delta\mathbf{X}^\circ\|_{p-F} = \rho\|\mathbf{u}\|_p\|\mathbf{f}(\boldsymbol{\beta}, d(p))\|_p = \rho,$$

and consequently, $\Delta\mathbf{X}^\circ \in \mathcal{N}_2$.

This allows us to state that

$$\max_{\Delta\mathbf{X} \in \mathcal{N}_2} \|\mathbf{y} - (\mathbf{X} + \Delta\mathbf{X})\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p + \rho\|\boldsymbol{\beta}\|_{d(p)},$$

which proves the theorem. □

Theorem 1 states that protecting the regression estimate against errors in the independent variables data bounded according to certain matrix norms is achieved by regularizing the corresponding optimization problem. The structure of the uncertainty set determines the norm which is going to be considered in the regularized problem. The conservativeness of the estimate is affected by parameter ρ , which determines the size of the uncertainty set as well as the contribution of the norm of the coefficient vector to the objective function to be minimized in the robust counterpart.

2.3 Support Vector Machines for Regression

Consider the ϵ -insensitive regression estimate which minimizes the ϵ -insensitive loss function

$$\min_{\beta, \beta_0} \sum_{i=1}^n \max(0, |y_i - \beta^T \mathbf{x}_i - \beta_0| - \epsilon).$$

The corresponding robust regression estimate, which immunizes against errors described by the uncertainty set \mathcal{N}_3 is

$$\min_{\beta, \beta_0} \max_{\Delta \mathbf{X} \in \mathcal{N}_3} \sum_{i=1}^n \max(0, |y_i - \beta^T (\mathbf{x}_i + \Delta \mathbf{x}_i) - \beta_0| - \epsilon), \quad (2.10)$$

where \mathcal{N}_3 is defined in Eq. (3.4).

We next prove that, under Assumption 1, the robust estimate of Problem (2.10) is the same as the support vector machines regression estimate.

Definition 2. *The data set (y_i, \mathbf{x}_i) , $i \in \{1, 2, \dots, n\}$, $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^m$, is called ϵ -approximable if there exists a $(\beta, \beta_0) \in \mathbb{R}^{m+1}$ such that for any $i \in \{1, 2, \dots, n\}$, $|y_i - \beta^T \mathbf{x}_i - \beta_0| < \epsilon$. Otherwise, the data set is called non- ϵ -approximable. In this case, for any $(\beta, \beta_0) \in \mathbb{R}^{m+1}$, there exists an $i \in \{1, 2, \dots, n\}$ such that $|y_i - \beta^T \mathbf{x}_i - \beta_0| \geq \epsilon$.*

Assumption 1. *The data set (y_i, \mathbf{x}_i) , $i \in \{1, 2, \dots, n\}$, is non- ϵ -approximable.*

Theorem 2. *Under Assumption 1, Problem (2.10) has a robust counterpart*

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \sum_{i=1}^n \xi_i + \rho \|\beta\|_q \\ & \xi_i \geq y_i - \beta^T \mathbf{x}_i - \beta_0 - \epsilon, i \in \{1, 2, \dots, n\} \\ & \xi_i \geq -y_i + \beta^T \mathbf{x}_i + \beta_0 - \epsilon, i \in \{1, 2, \dots, n\} \\ & \xi_i \geq 0, i \in \{1, 2, \dots, n\}, \end{aligned} \quad (2.11)$$

which is the optimization problem defining the support vector machines regression estimate.

Proof.

Let \mathcal{R}_1 be defined as in Eq. (3.7), and \mathcal{R}_2 be defined as in Eq. (3.8). Then, similar to Theorem 3, the inner maximization problem in Eq. (2.10) is expressed as

$$\max_{(\Delta \mathbf{X}, \boldsymbol{\tau}) \in \mathcal{R}_2} \sum_{i=1}^n \max(0, |y_i - \beta^T (\mathbf{x}_i + \Delta \mathbf{x}_i) - \beta_0| - \epsilon), \quad (2.12)$$

which is equivalent to

$$\max_{r \in \mathcal{R}_1} \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} \sum_{i=1}^n \max(0, |y_i - \beta^T(\mathbf{x}_i + \Delta \mathbf{x}_i) - \beta_0| - \epsilon). \quad (2.13)$$

Consider the inner maximization problem in Eq. (2.13)

$$\max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} \sum_{i=1}^n \max(0, |y_i - \beta^T(\mathbf{x}_i + \Delta \mathbf{x}_i) - \beta_0| - \epsilon). \quad (2.14)$$

This problem has a separable objective function and separable constraints. Consequently, we can solve problem

$$\max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} \max(0, |y_i - \beta^T(\mathbf{x}_i + \Delta \mathbf{x}_i) - \beta_0| - \epsilon). \quad (2.15)$$

for any $i \in \{1, 2, \dots, n\}$ and add the optimal objectives to form the optimal objective of Problem (2.14).

Since

$$\min_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} (\beta^T \Delta \mathbf{x}_i) = -r_i \rho \|\beta\|_q,$$

$$\max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} (\beta^T \Delta \mathbf{x}_i) = r_i \rho \|\beta\|_q,$$

we conclude that:

$$\begin{aligned} & \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} \max(0, |y_i - \beta^T(\mathbf{x}_i + \Delta \mathbf{x}_i) - \beta_0| - \epsilon) \\ &= \max(0, \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} (y_i - \beta^T(\mathbf{x}_i + \Delta \mathbf{x}_i) - \beta_0 - \epsilon), \\ & \quad \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} (-y_i + \beta^T(\mathbf{x}_i + \Delta \mathbf{x}_i) + \beta_0 - \epsilon)) \\ &= \max(0, y_i - \beta^T \mathbf{x}_i - \beta_0 - \epsilon - \min_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} (\beta^T \Delta \mathbf{x}_i), \end{aligned}$$

$$\begin{aligned}
& -y_i + \boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 - \epsilon + \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} (\boldsymbol{\beta}^T \Delta \mathbf{x}_i) \\
&= \max(0, y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0 - \epsilon + r_i \rho \|\boldsymbol{\beta}\|_q, -y_i + \boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 - \epsilon + r_i \rho \|\boldsymbol{\beta}\|_q) \\
&= \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon + r_i \rho \|\boldsymbol{\beta}\|_q).
\end{aligned}$$

Thus, Problem (2.13) is equivalent to

$$\max_{\mathbf{r} \in \mathcal{R}_1} \sum_{i=1}^n \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon + r_i \rho \|\boldsymbol{\beta}\|_q). \quad (2.16)$$

To solve Problem (2.16), we will bound its objective value and then, construct an $\mathbf{r} \in \mathcal{R}_1$ that achieves this bound.

Observing that

$$\begin{aligned}
& \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon + r_i \rho \|\boldsymbol{\beta}\|_q) \\
& \leq \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon) + r_i \rho \|\boldsymbol{\beta}\|_q,
\end{aligned}$$

we conclude that for $\mathbf{r} \in \mathcal{R}_1$,

$$\begin{aligned}
& \sum_{i=1}^n \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon + r_i \rho \|\boldsymbol{\beta}\|_q) \\
& \leq \sum_{i=1}^n \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon) + \sum_{i=1}^n r_i \rho \|\boldsymbol{\beta}\|_q \\
& \leq \sum_{i=1}^n \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon) + \rho \|\boldsymbol{\beta}\|_q.
\end{aligned}$$

Using Assumption 1, the data is non- ϵ -approximable, and thus, there exists an $i_o \in \{1, 2, \dots, n\}$, such that $|y_{i_o} - \boldsymbol{\beta}^T \mathbf{x}_{i_o} - \beta_0| \geq \epsilon$. Let

$$r_i = \begin{cases} 1, & i = i_o, \\ 0, & i \neq i_o. \end{cases}$$

For this $r \in \mathcal{R}_1$,

$$\begin{aligned}
& \sum_{i=1}^n \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon + r_i \rho \|\boldsymbol{\beta}\|_q) \\
&= \sum_{i=1, i \neq i_o}^n \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon) + \max(0, |y_{i_o} - \boldsymbol{\beta}^T \mathbf{x}_{i_o} - \beta_0| - \epsilon + \rho \|\boldsymbol{\beta}\|_q) \\
&= \sum_{i=1, i \neq i_o}^n \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon) + |y_{i_o} - \boldsymbol{\beta}^T \mathbf{x}_{i_o} - \beta_0| - \epsilon + \rho \|\boldsymbol{\beta}\|_q \\
&\quad (|y_{i_o} - \boldsymbol{\beta}^T \mathbf{x}_{i_o} - \beta_0| - \epsilon + \rho \|\boldsymbol{\beta}\|_q \geq 0) \\
&= \sum_{i=1}^n \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon) + \rho \|\boldsymbol{\beta}\|_q.
\end{aligned}$$

Consequently, the optimal objective of Problem (2.12) is equal to

$$\sum_{i=1}^n \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon) + \rho \|\boldsymbol{\beta}\|_q,$$

and Problem (2.10) is equivalent to

$$\min_{\boldsymbol{\beta}, \beta_0} \sum_{i=1}^n \max(0, |y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \beta_0| - \epsilon) + \rho \|\boldsymbol{\beta}\|_q,$$

which can be expressed as Problem (2.11). □

Theorem 2 states that the support vector machines regression estimate is a robust optimization estimate, and that the contribution of the regularization term depends on the norm used to define the uncertainty sets.

2.4 Experimental Results

To compare the performance of the robust and regularized estimators for regression to the performance of their respective nominal estimators, we conducted experiments using arti-

ficial, as well as real data sets. To solve all the convex problems needed to compute the estimates, we used SeDuMi [36], [48].

The artificial data set used to evaluate the quality of the robust estimates was developed in the following way:

1. A set of 200 random points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{200}$ in \mathbb{R}^3 was produced, according to the multivariate normal distribution with mean $[1, 1, 1]^T$ and covariance matrix $5\mathbf{I}_3$, where \mathbf{I}_3 is the 3×3 identity matrix.
2. For each \mathbf{x}_i , $y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + r$ was produced, where $\beta_0 = 1$,

$$\boldsymbol{\beta} = \begin{bmatrix} 1 \\ -3 \\ 1 \end{bmatrix},$$

and r is normally distributed with mean 0 and standard deviation 1.

The data set was normalized by scaling each one of the vectors containing the data corresponding to an independent variable to make their 2-norm equal to 1.

The performance of regular regression, robust regression with $(p = 1, q = 1)$, $(p = 1, q = 2)$, $(p = 2, q = 1)$, $(p = 2, q = 2)$, ϵ -insensitive regression, and support vector machines regression with $q = 2$ was measured for various values of ρ according to the following procedure:

1. The normalized data set was divided randomly into two groups containing the 50% of the samples each, the training set and the testing set.
2. A set of 100 random data points in \mathbb{R}^3 following the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix \mathbf{I}_3 was produced. These data points were scaled by ρ and added to the training set data points to contaminate them.
3. The contaminated data were used to produce the estimates to be studied.
4. The total error of the predictions of the data in the testing set was recorded for each estimate.
5. The procedure was repeated 30 times and the average performance of each estimate was recorded.

Parameter ϵ in ϵ -insensitive regression and support vector machines regression was set to the 0.01 of the maximum absolute value of the dependent variables of the data.

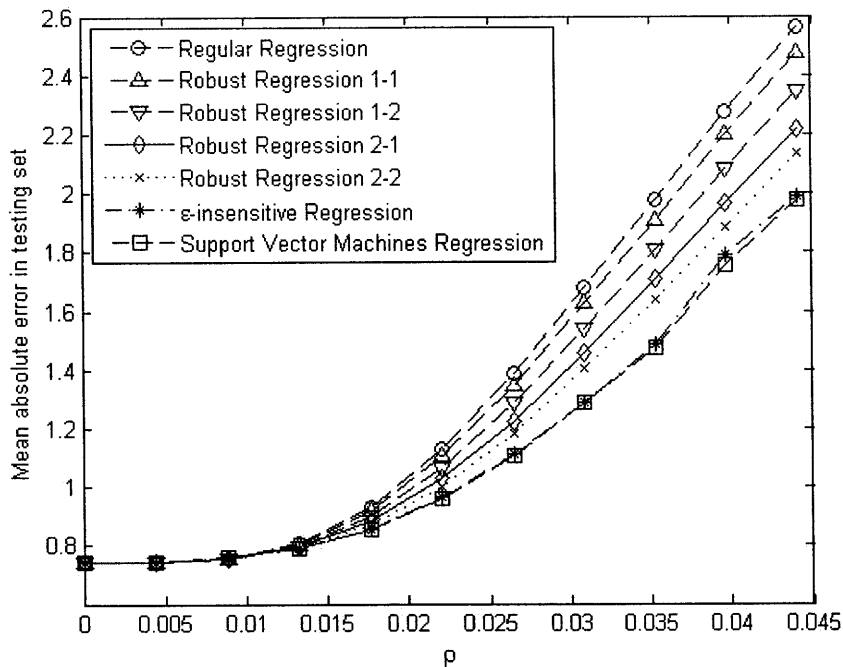


Figure 2-1: The average mean absolute error of the regression estimates according to ρ .

Using this procedure, the average mean absolute error and the average mean squared error of each estimate in the testing set was calculated for values of ρ between 0 and 0.045. The graphs of these errors according to ρ can be seen in Figures 2-1 and 2-2. The legends “Robust Regression 1-1”, “Robust Regression 1-2”, “Robust Regression 2-1”, “Robust Regression 2-2” refer to the robust regression estimates with $(p = 1, q = 1)$, $(p = 1, q = 2)$, $(p = 2, q = 1)$, and $(p = 2, q = 2)$, respectively.

We observe that as ρ increases, the difference in the out-of-sample performance between the robust and the respective classical estimates increases, with the robust estimates always yielding better results. The support vector and ϵ -insensitive regression estimates performed the best, while the ordering of the other methods in decreasing performance was: Robust 2-2, Robust 2-1, Robust 1-2, Robust 1-1.

The robust regression estimates were also tested using real data from the UCI Machine Learning Repository [3]. Again, the sets were normalized by scaling each one of the vectors containing the data corresponding to an independent variable to make their 2-norm equal to 1. The sizes of the used data sets can be seen in Table 2.1.

The evaluation procedure for each real data set was the following:

- The data set was divided in three sets, the training set, consisting of the 50% of the samples, the validating set, consisting of the 25% of the samples, and the testing set,

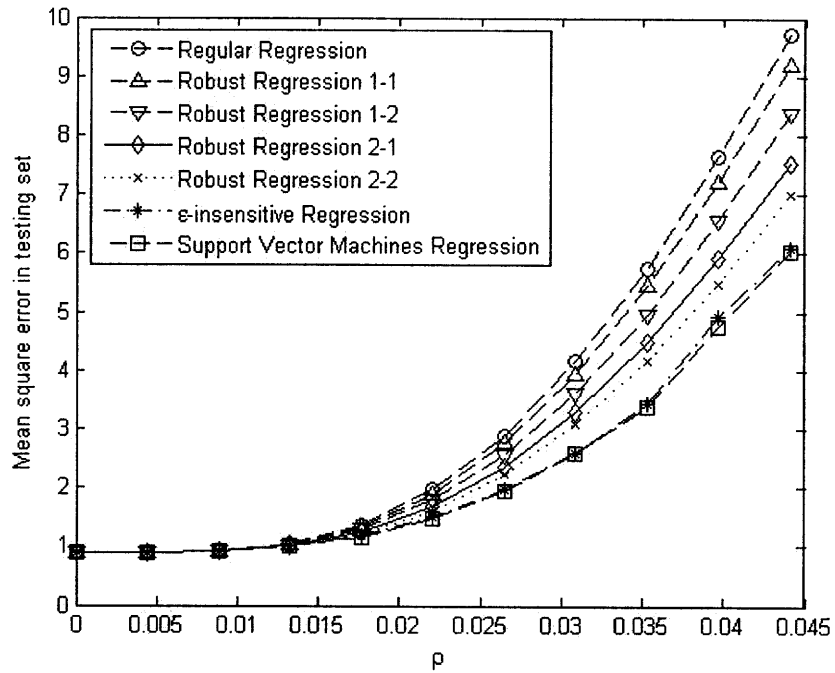


Figure 2-2: The average mean squared error of the regression estimates according to ρ .

Data set	n	m
Abalone	4177	9
Auto MPG	392	8
Comp Hard	209	7
Concrete	1030	8
Housing	506	13
Space shuttle	23	4
WPBC	46	32

Table 2.1: Sizes of real data sets for regression.

Data set	Regular	Rob 1-1	Rob 1-2	Rob 2-1	Rob 2-2	ϵ -ins	Supp Vector
Abalone	1.7517	1.7234	1.7219	1.7087	1.6700	1.6286	1.5474*
Auto MPG	3.2554	3.2412	3.2378	3.2265	3.1913	2.6661	2.5615*
Comp Hard	22.9119	21.3412	21.1217	22.6443	22.8445	19.8399	19.8239*
Concrete	9.2977	9.2387	9.1923	9.1614	8.9876	8.6809	8.3653*
Forest Fires	21.7727	21.6723	21.6321	21.7250	21.7330	18.3198	18.2675*
Housing	4.4950	4.4586	4.4489	4.4363	4.3821	3.3905	3.2341*
Space shuttle	0.5957	0.5617	0.5578	0.5865	0.5737	0.5036	0.4886*
WPBC	50.1432	49.9716	49.7623	49.2981	48.9397	49.3230	48.4037*

Table 2.2: Mean absolute error in testing set for real data sets. * denotes the estimate with the best performance.

consisting of the rest 25% of the samples. We considered 30 different partitions of the data set which were selected randomly.

- For each one of the considered partitions of the data set:
 - The regular regression estimate based on the training set was calculated.
 - The robust regression estimates based on the training set for various values of ρ were calculated. For each ρ , the total prediction error on the validating set was measured, and the ρ with the highest performance on the validating set was considered. The prediction error that this ρ yielded on the testing set was recorded.
- The prediction errors of the estimates under examination were averaged over the partitions of the data considered.

Parameter ϵ for ϵ -insensitive regression and the support vector machines regression was chosen in the same way as in the artificial data experiments.

The results of the evaluation process are summarized in Tables 2.2 and 2.3. Under the mean absolute error criterion, the support vector machines are always performing the best. In most cases, the ordering in the out-of-sample experiments is the same as in the artificial data sets. Under the mean squared error criterion, in five out of the eight real data sets, the support vector machines show the best performance, whereas in the rest three data sets, the robust regression estimates yield smaller out-of-sample errors.

Data set	Regular	Rob 1-1	Rob 1-2	Rob 2-1	Rob 2-2	ϵ -ins	Supp vector
Abalone	5.7430	5.6702	5.6543	5.6345	5.5369	5.3050	5.0483*
Auto MPG	18.7928	18.7245	18.7076	18.6981	18.5829	12.8846	12.5251*
Comp Hard	2026.00	2014.32	1978.12	1965.75	1925.13*	2463.03	2348.29
Concrete	132.47	131.46	131.32	131.08	129.31	131.53	127.09*
Forest Fires	5526.00	5312.18	5229.14	4994.81*	5266.40	5232.61	5229.52
Housing	39.8084	39.5412	39.4912	39.4257	39.0716	24.8051	24.6867*
Space shuttle	0.5323	0.5201	0.5177*	0.5225	0.5265	0.5582	0.5501
WPBC	4723.07	4676.20	4657.98	4630.19	4489.20	4498.14	4410.46*

Table 2.3: Mean squared error in testing set for real data sets. * denotes the estimate with the best performance.

2.5 Conclusions

Regularization in statistics can be interpreted as the application of robust optimization techniques in classical statistical estimates to provide protection against uncertainty in the data. The robust optimization paradigm offers a more adaptive and comprehensive control of the estimates through the use of various norms in defining uncertainty sets, while, at the same time, providing an insight of why the produced estimates yield improved performance compared to their respective classical ones.

Chapter 3

Support Vector Machines as Robust Estimators

3.1 Introduction

Vapnik et al. [10] developed Support Vector Machines (SVM), a method which provides classification estimates. Given a set of data (y_i, \mathbf{x}_i) , $i \in \{1, 2, \dots, n\}$, $y_i \in \{-1, 1\}$, $\mathbf{x}_i \in \mathbb{R}^m$, the SVM estimate is the optimal solution to optimization problem

$$\begin{aligned} \min_{\boldsymbol{\beta}, \beta_0, \boldsymbol{\xi}} \quad & \|\boldsymbol{\beta}\|_2 + \rho \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & \xi_i \geq 1 - y_i(\boldsymbol{\beta}'\mathbf{x}_i + \beta_0), i \in \{1, 2, \dots, n\} \\ & \xi_i \geq 0, i \in \{1, 2, \dots, n\}, \end{aligned} \tag{3.1}$$

where $\rho \geq 0$. The term $\|\boldsymbol{\beta}\|_2$ in the objective function is used to find a hyperplane classifier that is as far as possible from the samples and the term $\rho \sum_{i=1}^n \xi_i$ is used to allow misclassified samples. Support vector machines classifiers have been very successful in experiments (see Schölkopf [43]). Note that Problem (3.1) has a regularization term $\|\boldsymbol{\beta}\|_2$ in its objective.

The idea of applying Robust Optimization in defining classification estimators has already been explored. Lanckriet, El Ghaoui et al. defined classification estimators by minimizing the worst-case probability of misclassification [34]. The connection of SVM with robustness has already been explored. Shivaswamy et al. propose a second order cone programming problem that defines classifiers that can handle uncertainty [44].

In this chapter, we prove that Support Vector Machines are a particular case of robust optimization estimators (this last fact was also observed independently and simultaneously by Xu, Caramanis, and Mannor [52]). More specifically, in Section 3.2, we prove that the

support vector machines problem, which yields the respective estimate for classification, is the robust counterpart of some nominal problem which classifies binary data as well, and we calculate the relation between the norms of the uncertainty sets and the norm used for the regularization. In Section 3.3, we report on computational results of the Support Vector Machines estimators and their respective classical ones.

3.2 Robust Properties of Support Vector Machines

Given a set of categorical data (y_i, \mathbf{x}_i) , $y_i \in \{1, -1\}$, $\mathbf{x}_i \in \mathbb{R}^m$, $i \in \{1, 2, \dots, n\}$, we define the separation error $S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, \mathbf{X})$ of the hyperplane classifier $\boldsymbol{\beta}^T \mathbf{x} + \beta_0 = 0$, $\mathbf{x} \in \mathbb{R}^m$, by

$$S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, \mathbf{X}) = \sum_{i=1}^n \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0)), \quad (3.2)$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \text{ and } \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}.$$

According to Eq. (3.2), an observation (y_i, \mathbf{x}_i) contributes a non-zero quantity to $S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, \mathbf{X})$ only if $\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 \leq 1$, $\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 \geq -1$, for $y_i = 1$, $y_i = -1$, respectively. The amount of the contribution is the distance of $\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0$ from 1, -1 , for $y_i = 1$, $y_i = -1$, respectively.

The hyperplane which minimizes the separation error is the solution to the optimization problem

$$\min_{\boldsymbol{\beta}, \beta_0} S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, \mathbf{X}), \quad (3.3)$$

which can be expressed as the linear optimization problem

$$\begin{aligned} \min_{\boldsymbol{\beta}, \beta_0, \boldsymbol{\xi}} \quad & \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) \geq 1 - \xi_i, i \in \{1, 2, \dots, n\} \\ & \xi_i \geq 0, i \in \{1, 2, \dots, n\}. \end{aligned}$$

Consider the uncertainty set

$$\mathcal{N}_3 = \left\{ \Delta \mathbf{X} \in \mathbb{R}^{n \times m} \mid \sum_{i=1}^n \|\Delta \mathbf{x}_i\|_p \leq \rho \right\}. \quad (3.4)$$

where $\|\bullet\|_p$ is the p -norm.

The robust version of Problem (3.3), which immunizes the computed hyperplane against errors in the independent variables of the data described by the set \mathcal{N}_3 , is

$$\min_{\beta, \beta_0} \max_{\Delta \mathbf{X} \in \mathcal{N}_3} S(\beta, \beta_0, \mathbf{y}, \mathbf{X} + \Delta \mathbf{X}). \quad (3.5)$$

We next prove that under Assumption 2, this robust estimate is equivalent to the support vector machines classification estimate.

Definition 3. *The set of data (y_i, \mathbf{x}_i) , $i \in \{1, 2, \dots, n\}$, is called separable if there exists a hyperplane $\beta^T \mathbf{x} + \beta_0 = 0$, such that for any $i \in \{1, 2, \dots, n\}$, $y_i(\beta^T \mathbf{x}_i + \beta_0) \geq 0$. Otherwise, the data set is called non-separable. In this case, for any hyperplane $\beta^T \mathbf{x} + \beta_0 = 0$, there exists an $i \in \{1, 2, \dots, n\}$ with $y_i(\beta^T \mathbf{x}_i + \beta_0) < 0$.*

Assumption 2. *The data (y_i, \mathbf{x}_i) , $i \in \{1, 2, \dots, n\}$, is non-separable.*

Theorem 3. *Under Assumption 2, Problem (3.5) is equivalent to*

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \sum_{i=1}^n \xi_i + \rho \|\beta\|_{d(p)} \\ \text{s.t.} \quad & \xi_i \geq 1 - y_i(\beta^T \mathbf{x}_i + \beta_0), i \in \{1, 2, \dots, n\} \\ & \xi_i \geq 0, i \in \{1, 2, \dots, n\}, \end{aligned} \quad (3.6)$$

where $d(p)$ is defined in Eq. (4.3), i.e., $\|\bullet\|_{d(p)}$ is the dual norm of $\|\bullet\|_p$.

Proof.

To prove the theorem, we are going to introduce a new set of variables r_i , $i \in \{1, 2, \dots, n\}$, which will be used to bound each $\|\Delta \mathbf{x}_i\|_p$, and thus, to control the contribution of each $\|\Delta \mathbf{x}_i\|_p$ to $\sum_{i=1}^n \|\Delta \mathbf{x}_i\|_p$, which determines the uncertainty set.

Consider

$$\mathcal{R}_1 = \left\{ \mathbf{r} \in \mathbb{R}^n \mid \sum_{i=1}^n r_i \leq 1, r_i \geq 0, i = 1, 2, \dots, n \right\}, \quad (3.7)$$

and

$$\mathcal{R}_2 = \{(\Delta \mathbf{X}, \mathbf{r}) \in \mathbb{R}^{n \times m} \times \mathbb{R}^n \mid \mathbf{r} \in \mathcal{R}_1, \|\Delta \mathbf{x}_i\|_p \leq r_i \rho, i = 1, 2, \dots, n\}. \quad (3.8)$$

It is clear that the projection of \mathcal{R}_2 onto $\Delta \mathbf{X}$ is \mathcal{N}_3 . Thus, the inner problem in Eq. (3.5) can be expressed as

$$\max_{(\Delta \mathbf{X}, \mathbf{r}) \in \mathcal{R}_2} S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, \mathbf{X} + \Delta \mathbf{X}), \quad (3.9)$$

which is equivalent to

$$\max_{\mathbf{r} \in \mathcal{R}_1} \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, \mathbf{X} + \Delta \mathbf{X}). \quad (3.10)$$

Consider the inner maximization problem in Eq. (3.10)

$$\max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, \mathbf{X} + \Delta \mathbf{X}). \quad (3.11)$$

Since

$$S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, \mathbf{X} + \Delta \mathbf{X}) = \sum_{i=1}^n \max(0, 1 - y_i(\boldsymbol{\beta}^T(\mathbf{x}_i + \Delta \mathbf{x}_i) + \beta_0)),$$

Problem (3.11) is separable and has separable constraints. Its solution is given by solving

$$\max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} \max(0, 1 - y_i(\boldsymbol{\beta}^T(\mathbf{x}_i + \Delta \mathbf{x}_i) + \beta_0)). \quad (3.12)$$

for any $i \in \{1, 2, \dots, n\}$ and adding the optimal objectives.

Observe that:

$$\begin{aligned} & \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} \max(0, 1 - y_i(\boldsymbol{\beta}^T(\mathbf{x}_i + \Delta \mathbf{x}_i) + \beta_0)) \\ &= \max(0, \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} (1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) - y_i \boldsymbol{\beta}^T \Delta \mathbf{x}_i)) \\ &= \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) + \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} (-y_i \boldsymbol{\beta}^T \Delta \mathbf{x}_i)). \end{aligned}$$

We know that:

$$\max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} (-y_i \boldsymbol{\beta}^T \Delta \mathbf{x}_i) = \begin{cases} -\min_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} \boldsymbol{\beta}^T \Delta \mathbf{x}_i, & \text{if } y_i = 1, \\ \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} \boldsymbol{\beta}^T \Delta \mathbf{x}_i, & \text{if } y_i = -1, \end{cases}$$

$$\min_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} \boldsymbol{\beta}^T \Delta \mathbf{x}_i = -r_i \rho \|\boldsymbol{\beta}\|_q,$$

and

$$\max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} \boldsymbol{\beta}^T \Delta \mathbf{x}_i = r_i \rho \|\boldsymbol{\beta}\|_q.$$

Consequently,

$$\max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} (-y_i \boldsymbol{\beta}^T \Delta \mathbf{x}_i) = r_i \rho \|\boldsymbol{\beta}\|_q,$$

and

$$\begin{aligned} & \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} \max(0, 1 - y_i(\boldsymbol{\beta}^T(\mathbf{x}_i + \Delta \mathbf{x}_i) + \beta_0)) \\ &= \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0)) + \max_{\|\Delta \mathbf{x}_i\|_p \leq r_i \rho} (-y_i \boldsymbol{\beta}^T \Delta \mathbf{x}_i) \\ &= \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) + r_i \rho \|\boldsymbol{\beta}\|_q) \end{aligned}$$

is the optimal objective of Problem (3.12). Thus, the optimal objective of Problem (3.11) is

$$\sum_{i=1}^n \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) + r_i \rho \|\boldsymbol{\beta}\|_q).$$

Given this observation, Problem (3.10) is equivalent to

$$\max_{\mathbf{r} \in \mathcal{R}_1} \sum_{i=1}^n \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) + r_i \rho \|\boldsymbol{\beta}\|_q). \quad (3.13)$$

To solve Problem (3.13), we find an upper bound for its objective value and then, we construct

an \mathbf{r} which achieves this upper bound.

Since

$$\begin{aligned} & \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) + r_i \rho \|\boldsymbol{\beta}\|_q) \\ & \leq \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0)) + r_i \rho \|\boldsymbol{\beta}\|_q, \end{aligned}$$

we conclude that for $\mathbf{r} \in \mathcal{R}_1$

$$\begin{aligned} & \sum_{i=1}^n \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) + r_i \rho \|\boldsymbol{\beta}\|_q) \\ & \leq \sum_{i=1}^n \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0)) + \sum_{i=1}^n r_i \rho \|\boldsymbol{\beta}\|_q \\ & = S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, X) + \sum_{i=1}^n r_i \rho \|\boldsymbol{\beta}\|_q \leq S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, X) + \rho \|\boldsymbol{\beta}\|_q. \end{aligned}$$

Using Assumption 2, the data is non-separable, and thus, there exists an $i_m \in \{1, 2, \dots, n\}$ such that $y_{i_m}(\boldsymbol{\beta}^T \mathbf{x}_{i_m} + \beta_0) < 0$. We have:

$$1 - y_{i_m}(\boldsymbol{\beta}^T \mathbf{x}_{i_m} + \beta_0) > 0,$$

and

$$1 - y_{i_m}(\boldsymbol{\beta}^T \mathbf{x}_{i_m} + \beta_0) + \rho \|\boldsymbol{\beta}\|_q > 0.$$

Let

$$r_i = \begin{cases} 1, & i = i_m, \\ 0, & i \neq i_m. \end{cases}$$

For this \mathbf{r} ,

$$\sum_{i=1}^n \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) + r_i \rho \|\boldsymbol{\beta}\|_q)$$

$$\begin{aligned}
&= \sum_{i=1, i \neq i_m}^n \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0)) + 1 - y_{i_m}(\boldsymbol{\beta}^T \mathbf{x}_{i_m} + \beta_0) + \rho \|\boldsymbol{\beta}\|_q \\
&= \sum_{i=1}^n \max(0, 1 - y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0)) + \rho \|\boldsymbol{\beta}\|_q = S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, \mathbf{X}) + \rho \|\boldsymbol{\beta}\|_q.
\end{aligned}$$

Thus, the optimal objective value of Problem (3.13) is $S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, \mathbf{X}) + \rho \|\boldsymbol{\beta}\|_q$ and a robust counterpart of Problem (3.5) is

$$\min_{\boldsymbol{\beta}, \beta_0} S(\boldsymbol{\beta}, \beta_0, \mathbf{y}, \mathbf{X}) + \rho \|\boldsymbol{\beta}\|_q,$$

which is equivalent to Problem (3.6). □

As Theorem 3 states, the support vector machines estimate is a robust optimization estimate, attempting to provide protection against errors in the independent variables.

3.3 Experimental Results

To compare the performance of the Support Vector Machines to the performance of their respective nominal estimators, we conducted experiments using artificial, as well as real data sets. To solve all the convex problems needed to compute the estimates, we used SeDuMi [36], [48].

The artificial data set used to evaluate the support vector machines classifier was generated in the following way:

- A set of 100 points in \mathbb{R}^3 obeying the multivariate normal distribution with mean $[1, 0, 0]^T$ and covariance matrix $\frac{1}{\sqrt{2}}\mathbf{I}_3$ was generated, where \mathbf{I}_3 is the 3×3 identity matrix. The points were associated with $y = 1$, and added to the data set.
- A set of 100 points in \mathbb{R}^3 obeying the multivariate normal distribution with mean $[0, 1, 0]^T$ and standard deviation $\frac{1}{\sqrt{2}}\mathbf{I}_3$ was generated. The points were associated with $y = -1$, and added to the data set.

The performance of the separation error estimate and the support vector machines estimate was measured for values of ρ ranging between 0 and 0.045 using the generated data set and the same procedure as in the regression artificial data case described in Chapter 2. The results are summarized in Figure 3-1. The performance metric used was the classification error of the estimate. We observe that the support vector machines estimate, which is the

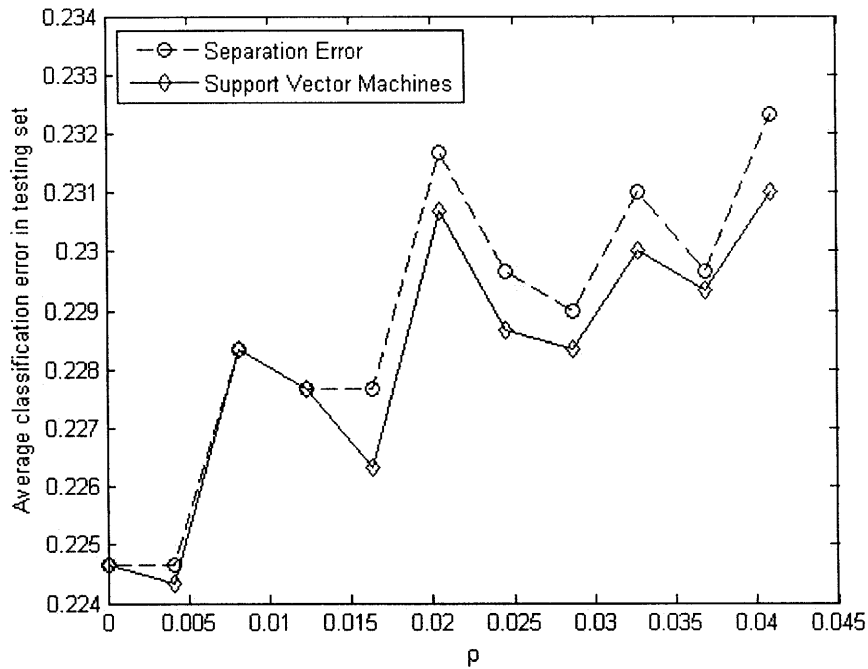


Figure 3-1: The average classification error of the classification estimates according to ρ .

robust version of the separation error estimate, yields a smaller error for most of the values of ρ , confirming the effectiveness of the robustness and regularization procedure.

The classification estimates were also tested using real data from the UCI Machine Learning Repository [3]. The procedure followed was the same as in the regression estimates case described in Chapter 2, and the performance metric applied was the classification error. The sizes of the used data sets can be seen in Table 3.1. The results are summarized in Table 3.2. The support vector machines estimate yields better results than the separation error estimate in all cases. The maximum improvement is obtained in the “Lenses” set, where the support vector machines estimate, which is characterized by the robust optimization and regularization ideas, is 6.25% better than the classical estimate.

3.4 Conclusions

Support Vector Machines for classification is a particular case of an estimator defined using the robust optimization paradigm. We proved theorems that facilitate the choice of the norm and the coefficient used in regularization based on the structure and the size of the considered uncertainty sets. The robust estimator shows improved performance in artificial and real data sets compared to its respective nominal one.

Data set	n	m
Pima	768	8
Spam	4601	57
Heart	270	13
Ionosphere	351	33
Lenses	24	5
SPECTF	267	44
TAE	151	5
WDBC	569	30
Yeast	1484	8
Wine	178	13

Table 3.1: Sizes of real data sets for classification.

Data set	Sep error	Supp vector
Pima	0.2451	0.2330*
Spam	0.0779	0.0744*
Heart	0.1676	0.1627*
Ionosphere	0.1496	0.1481*
Lenses	0.2667	0.2500*
SPECTF	0.2505	0.2413*
TAE	0.3311	0.3184*
WDBC	0.0469	0.0445*
Yeast	0.1440	0.1395*
Wine	0.0330	0.0326*

Table 3.2: Classification error in testing set for real data sets. * denotes the estimate with the best performance.

Chapter 4

Robust Logistic Regression

4.1 Logistic Regression

Logistic regression is a widely used method for analysing categorical data, and making predictions for them. Given a set of observations (y_i, \mathbf{x}_i) , $i \in \{1, 2, \dots, n\}$, $y_i \in \{0, 1\}$, $\mathbf{x}_i \in \mathbb{R}^m$, classical logistic regression calculates the maximum likelihood estimate for the parameter $(\boldsymbol{\beta}, \beta_0) \in \mathbb{R}^{m+1}$ of the logistic regression model (see Ryan [40], Hosmer [28])

$$\Pr[Y = 1 | \mathbf{X} = \mathbf{x}] = \frac{\exp(\boldsymbol{\beta}^T \mathbf{x} + \beta_0)}{1 + \exp(\boldsymbol{\beta}^T \mathbf{x} + \beta_0)}, \quad (4.1)$$

where $Y \in \{0, 1\}$ is the response variable determining the class of the sample and $\mathbf{X} \in \mathbb{R}^m$ is the independent variables vector.

Very frequently, the observations used to produce the estimate are subject to errors. The errors can be present in either the independent variables, in the response variable, or in both. For example, in predicting whether the financial condition of a company is sound, the economic indices of the company used to make the prediction might have been measured with errors. In predicting whether a patient is going to be cured from a disease given several medical tests they undertook, the response variable demonstrating whether the patient was indeed cured might contain errors.

The presence of errors affects the estimate that classical logistic regression yields and makes the predictions less accurate. It is desirable to produce estimates that are able to make accurate predictions even in the presence of such errors.

In this thesis, robust optimization techniques are applied in order to produce new robust estimates for the logistic regression model that are more immune to errors in the observations' data. The contributions achieved include:

1. New notions of robust logistic regression when the estimates are immunized against errors in only the independent variables, in only the response variable, and in both the independent and the response variables are introduced.
2. Efficient algorithms based on convex optimization methods on how to compute these robust estimates of the coefficients $(\boldsymbol{\beta}, \beta_0)$ in Eq. (4.1) are constructed.
3. Experiments in both artificial and real data illustrate that the robust estimates provide superior out-of-sample performance.

The structure of the chapter is as follows. In Sections 4.2, 4.3, and 4.4, the methodology of computing robust logistic regression estimates that are protected against errors in their independent variables, in their response variable, or in both respectively is outlined. In Section 4.5, the performance of the proposed algorithms in comparison with the classical logistic regression for both artificial, and real data sets is reported.

4.2 Robust logistic regression under independent variables uncertainty

Given a set of observations (y_i, \mathbf{x}_i) , $i \in \{1, 2, \dots, n\}$, we assume that the true value of the independent variables of the observations is $\mathbf{x}_i + \boldsymbol{\Delta}\mathbf{x}_i$, where the p -norm of $\boldsymbol{\Delta}\mathbf{x}_i$ is bounded above by some parameter ρ , that is $\|\boldsymbol{\Delta}\mathbf{x}_i\|_p \leq \rho$.

Let \mathbf{X} be the matrix in which row i is vector \mathbf{x}_i , $\boldsymbol{\Delta}\mathbf{X}$ be the matrix in which row i is vector $\boldsymbol{\Delta}\mathbf{x}_i$ and \mathbf{y} be the vector whose i -th coordinate is y_i , $i = 1, 2, \dots, n$. Let function $P(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \beta_0)$ denote the log-likelihood on the given set of observations

$$P(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \beta_0) = \sum_{i=1}^n [y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0))]. \quad (4.2)$$

Let

$$d(p) = \frac{p}{p-1}, \quad p \geq 1. \quad (4.3)$$

We also define that $d(1) = \infty$ and $d(\infty) = 1$. Note that $\|\bullet\|_{d(p)}$ is the dual norm of $\|\bullet\|_p$.

Let

$$S_1 = \{\boldsymbol{\Delta}\mathbf{X} \mid \|\boldsymbol{\Delta}\mathbf{x}_i\|_p \leq \rho, \quad i = 1, 2, \dots, n\}. \quad (4.4)$$

The robust estimate we propose is defined by:

$$\max_{\boldsymbol{\beta}, \beta_0} \min_{\Delta \mathbf{X} \in S_1} P(\mathbf{y}, \mathbf{X} + \Delta \mathbf{X}, \boldsymbol{\beta}, \beta_0). \quad (4.5)$$

In other words, $(\boldsymbol{\beta}, \beta_0)$ is evaluated by the worst case log-likelihood as the error in the measurement of the independent variables lies in the uncertainty set S_1 .

In the next theorem, we calculate an analytical formula for the optimal objective of the inner minimization problem in Eq. (4.5) as a function of $(\boldsymbol{\beta}, \beta_0)$.

Theorem 4.

(a) *An optimal solution to*

$$Z_1(\boldsymbol{\beta}, \beta_0) = \min_{\Delta \mathbf{X} \in S_1} P(\mathbf{y}, \mathbf{X} + \Delta \mathbf{X}, \boldsymbol{\beta}, \beta_0) \quad (4.6)$$

is $\Delta \mathbf{X}^o$, where the i -th row $\Delta \mathbf{x}_i^o$ of matrix $\Delta \mathbf{X}^o$, $i \in \{1, 2, \dots, n\}$, is given by

$$\Delta \mathbf{x}_i^o = (-1)^{y_i} \rho \mathbf{f}(\boldsymbol{\beta}, d(p)), \quad i \in \{1, 2, \dots, n\}, \quad (4.7)$$

and $\mathbf{f}(\mathbf{x}, p) \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^m$, $p \geq 1$, is defined in Appendix A.1.

(b) *The optimal objective value $Z_1(\boldsymbol{\beta}, \beta_0)$ is given by*

$$\begin{aligned} Z_1(\boldsymbol{\beta}, \beta_0) = & \sum_{i=1}^n [y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}) \\ & - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p})))] . \end{aligned} \quad (4.8)$$

Proof.

(a) Given that the objective function of Problem (4.6) is separable and that each of its constraints involves a single $\Delta \mathbf{x}_i$, the optimal solution for each $\Delta \mathbf{x}_i$ is the optimal solution to

$$\min_{\|\Delta \mathbf{x}_i\|_p \leq \rho} y_i[\boldsymbol{\beta}^T(\mathbf{x}_i + \Delta \mathbf{x}_i) + \beta_0] - \ln(1 + \exp(\boldsymbol{\beta}^T(\mathbf{x}_i + \Delta \mathbf{x}_i) + \beta_0)). \quad (4.9)$$

Defining

$$g_i(w) = y_i w - \ln\left(\frac{1}{1 + \exp(w)}\right), \quad i = 1, 2, \dots, n,$$

we observe that the objective function of Problem (4.9) is equal to

$$g_i(\boldsymbol{\beta}^T(\mathbf{x}_i + \Delta \mathbf{x}_i) + \beta_0).$$

The first derivative of $g_i(w)$ is

$$\frac{dg_i(w)}{dw} = y_i - \frac{\exp(w)}{1 + \exp(w)} = \begin{cases} -\frac{\exp(w)}{1 + \exp(w)} < 0, & \text{if } y_i = 0, \\ \frac{1}{1 + \exp(w)} > 0, & \text{if } y_i = 1. \end{cases}$$

Thus, for $y_i = 0$, function $g_i(w)$ is strictly decreasing, whereas for $y_i = 1$, function $g_i(w)$ is strictly increasing. This implies that to solve Problem (4.9), it suffices to maximize or minimize $\boldsymbol{\beta}^T \Delta \mathbf{x}_i$, according to whether $y_i = 0$ or $y_i = 1$ respectively.

Using Hölder's inequality (see Boyd [12], p. 78), we conclude:

$$-\|\Delta \mathbf{x}_i\|_p \|\boldsymbol{\beta}\|_{d(p)} \leq \boldsymbol{\beta}^T \Delta \mathbf{x}_i \leq \|\Delta \mathbf{x}_i\|_p \|\boldsymbol{\beta}\|_{d(p)}.$$

For any $\Delta \mathbf{X} \in S_1$, we have that $\|\Delta \mathbf{x}_i\|_p \leq \rho$ and hence,

$$-\rho \|\boldsymbol{\beta}\|_{d(p)} \leq \boldsymbol{\beta}^T \Delta \mathbf{x}_i \leq \rho \|\boldsymbol{\beta}\|_{d(p)}. \quad (4.10)$$

Let $\Delta \mathbf{X}^\circ$ be defined as in Eq. (4.7). Using Proposition 1 in Appendix A.1, we have that:

$$\boldsymbol{\beta}^T \Delta \mathbf{x}_i^\circ = \boldsymbol{\beta}^T (-1)^{y_i} \rho \mathbf{f}(\boldsymbol{\beta}, d(p)) = (-1)^{y_i} \rho \mathbf{f}(\boldsymbol{\beta}, d(p))^T \boldsymbol{\beta} = (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}.$$

Thus, if $y_i = 0$, $\boldsymbol{\beta}^T \Delta \mathbf{x}_i$ attains its maximum value $\rho \|\boldsymbol{\beta}\|_{d(p)}$ in Eq. (4.10), and, by the monotonicity of $g_i(w)$, $\Delta \mathbf{x}_i^\circ$ is an optimal solution to Problem (4.9). Similarly, if $y_i = 1$, $\boldsymbol{\beta}^T \Delta \mathbf{x}_i$ attains its minimum value $-\rho \|\boldsymbol{\beta}\|_{d(p)}$ in Eq. (4.10), and hence, it is an optimal solution to Problem (4.9). We conclude that $\Delta \mathbf{X}^\circ$ is an optimal solution to Problem (4.6).

(b) Using the observation that $\boldsymbol{\beta}^T \Delta \mathbf{x}_i^\circ = (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}$, we conclude that the optimal objective value of Problem (4.6) is given by Eq. (4.8). \square

Function $P(\mathbf{y}, \mathbf{X} + \Delta \mathbf{X}, \boldsymbol{\beta}, \beta_0)$ is concave in $(\boldsymbol{\beta}, \beta_0)$ for any $\Delta \mathbf{X} \in S_1$ (see Ryan [40], Hosmer [28]). Thus, $Z_1(\boldsymbol{\beta}, \beta_0) = \min_{\Delta \mathbf{X} \in S_1} P(\mathbf{y}, \mathbf{X} + \Delta \mathbf{X}, \boldsymbol{\beta}, \beta_0)$, which constitutes the minimum value of $P(\mathbf{y}, \mathbf{X} + \Delta \mathbf{X}, \boldsymbol{\beta}, \beta_0)$ over the feasible set S_1 , is concave in $(\boldsymbol{\beta}, \beta_0)$ (see [12], p. 81).

Using Theorem 4, the robust counterpart of Problem (4.5) is formulated as

$$\max_{\boldsymbol{\beta}, \beta_0} Z_1(\boldsymbol{\beta}, \beta_0). \quad (4.11)$$

Problem (4.11) is an unconstrained concave maximization problem. However, the function $Z_1(\boldsymbol{\beta}, \beta_0)$ is not differentiable at any $(\boldsymbol{\beta}, \beta_0) \in \mathbb{R}^{m+1}$ with $\boldsymbol{\beta} = \mathbf{0}$. For this reason, we apply the subgradient maximization method (see Shor [45]), which converges to the optimal solution of (4.11). The details of the method are outlined in Appendix B.1.

4.3 Robust logistic regression under response variable uncertainty

Given a set of observations, the response variable of each one can either be affected by an error or not. In this model, we immunize the estimate under the uncertainty set defined by an upper bound Γ on the number of errors in the response variable of the observations. In this way, we restrict “nature” to modify up to Γ observations. The parameter Γ makes a tradeoff between robustness and optimality.

Let Δy_i be a binary variable which is equal to 1, if there is an error in the nominal value of the response variable of observation i , and 0, otherwise. The true value of the response variable of observation i is then equal to $|y_i - \Delta y_i|$. Let $\boldsymbol{\Delta y}$ be the vector whose i -th coordinate is Δy_i , $i \in \{1, 2, \dots, n\}$. Each value of the parameter $(\boldsymbol{\beta}, \beta_0)$ is evaluated as the worst case of $P(|\mathbf{y} - \boldsymbol{\Delta y}|, \mathbf{X}, \boldsymbol{\beta}, \beta_0)$, when $\boldsymbol{\Delta y}$ lies in the uncertainty set

$$S_2 = \left\{ \boldsymbol{\Delta y} \in \{0, 1\}^n \mid \sum_{i=1}^n \Delta y_i \leq \Gamma \right\}. \quad (4.12)$$

Formally, the robust estimate in this case is the optimal solution to

$$\max_{\boldsymbol{\beta}, \beta_0} \min_{\boldsymbol{\Delta y} \in S_2} P(|\mathbf{y} - \boldsymbol{\Delta y}|, \mathbf{X}, \boldsymbol{\beta}, \beta_0), \quad (4.13)$$

where $P(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \beta_0)$ is defined in Eq. (4.2).

A substantial difference between the uncertainty set S_2 considered here and the uncertainty set S_1 considered in Section 4.2 is that the uncertainties in S_1 are separable, whereas in S_2 are not. In the next theorem, we calculate an equivalent maximization problem for the inner minimization problem in Eq. (4.13).

Theorem 5. *Problem*

$$\min_{\Delta \mathbf{y} \in S_2} P(|\mathbf{y} - \Delta \mathbf{y}|, \mathbf{X}, \boldsymbol{\beta}, \beta_0) \quad (4.14)$$

has the same objective value as problem

$$\begin{aligned} \max \quad & \Gamma p + \sum_{i=1}^n q_i + \sum_{i=1}^n [y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0))] \\ \text{s.t.} \quad & p + q_i + (-1)^{1-y_i}(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) \leq 0, \quad i \in \{1, 2, \dots, n\} \\ & p \leq 0, \mathbf{q} \leq \mathbf{0}. \end{aligned} \quad (4.15)$$

Proof. Since $|y_i - \Delta y_i| = (-1)^{y_i} \Delta y_i + y_i$, the objective function of (4.14) is expressed as

$$\begin{aligned} P(|\mathbf{y} - \Delta \mathbf{y}|, \mathbf{X}, \boldsymbol{\beta}, \beta_0) &= \sum_{i=1}^n [|y_i - \Delta y_i| (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0))] \\ &= \sum_{i=1}^n (-1)^{y_i} (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) \Delta y_i + \sum_{i=1}^n [y_i (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0))]. \end{aligned} \quad (4.16)$$

The only term in Eq. (4.16) affected by Δy_i is $\sum_{i=1}^n (-1)^{y_i} (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) \Delta y_i$, implying that the optimal solution of Problem (4.14) is the optimal solution of the integer optimization problem

$$\begin{aligned} \min \quad & \sum_{i=1}^n (-1)^{y_i} (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) \Delta y_i \\ \text{s.t.} \quad & \Delta \mathbf{y} \in S_2. \end{aligned} \quad (4.17)$$

Since the polyhedron $\{\Delta \mathbf{y} \mid \sum_{i=1}^n \Delta y_i \leq \Gamma, 0 \leq \Delta y_i \leq 1\}$ has integer extreme points, Problem (4.17) has the same optimal solution as its linear optimization relaxation

$$\begin{aligned} \min \quad & \sum_{i=1}^n (-1)^{y_i} (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) \Delta y_i \\ \text{s.t.} \quad & \sum_{i=1}^n \Delta y_i \leq \Gamma \\ & 0 \leq \Delta y_i \leq 1, \quad i \in \{1, 2, \dots, n\}. \end{aligned} \quad (4.18)$$

By strong duality in linear optimization, the optimal objective value of Problem (4.18)

is equal to the optimal objective value of its dual

$$\begin{aligned}
& \max \quad \Gamma p + \sum_{i=1}^n q_i \\
& \text{s.t.} \quad p + q_i \leq (-1)^{y_i} (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) \\
& \quad \quad p \leq 0, \mathbf{q} \leq \mathbf{0},
\end{aligned} \tag{4.19}$$

and hence, Problem (4.15) has the same optimal objective value as Problem (4.14). \square

Using Theorem 5, we express the robust counterpart of Problem (4.13) as

$$\begin{aligned}
& \max \quad Z_2(p, \mathbf{q}, \boldsymbol{\beta}, \beta_0) \\
& \text{s.t.} \quad p + q_i + (-1)^{1-y_i} (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) \leq 0, \quad i \in \{1, 2, \dots, n\} \\
& \quad \quad p \leq 0, \mathbf{q} \leq \mathbf{0},
\end{aligned} \tag{4.20}$$

where

$$Z_2(p, \mathbf{q}, \boldsymbol{\beta}, \beta_0) = \Gamma p + \sum_{i=1}^n q_i + \sum_{i=1}^n [y_i (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0))].$$

The objective function of Problem (4.20) is concave in $(p, \mathbf{q}, \boldsymbol{\beta}, \beta_0)$ subject to linear constraints. Problem (4.20) is a concave maximization problem with twice continuously differentiable objective function and twice continuously differentiable constraints, solvable with an interior point method (see Bertsekas [7], Boyd [12]). The details of the method are outlined in Appendix B.2.

4.4 Globally robust logistic regression

Globally robust logistic regression provides an estimate for the parameters $(\boldsymbol{\beta}, \beta_0)$ of the logistic regression model taking into consideration that both the independent and the response variables of the observations are subject to errors.

The proposed estimate is defined as the solution to

$$\max_{\boldsymbol{\beta}, \beta_0} \min_{\Delta \mathbf{X} \in S_1, \Delta \mathbf{y} \in S_2} P(|\mathbf{y} - \Delta \mathbf{y}|, \mathbf{X} + \Delta \mathbf{X}, \boldsymbol{\beta}, \beta_0), \tag{4.21}$$

where $P(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \beta_0)$ is defined in Eq. (4.2), and S_1, S_2 are defined in Eq. (4.4), Eq. (4.12), respectively.

In Theorem 6 below, we calculate a formula for the optimal value of the inner minimization problem in Eq. (4.21). To present this formula we need to introduce some notation.

Definition 4. If $\mathbf{v}^1 = (v_1^1, v_2^1, \dots, v_n^1)$, $\mathbf{v}^2 = (v_1^2, v_2^2, \dots, v_n^2)$ are length n binary sequences, \mathbf{v}^1 is lexicographically smaller than \mathbf{v}^2 , or equivalently $\mathbf{v}^1 <_{\text{lex}} \mathbf{v}^2$, if there exists some $i_0 \in \{1, 2, \dots, n\}$ such that for any $i \in \{1, 2, \dots, i_0 - 1\}$, $v_i^1 = v_i^2$ and $v_{i_0}^1 < v_{i_0}^2$.

If $\mathbf{a} = (a_1, a_2, \dots, a_n)$, let $S(\mathbf{a})$ be the set of optimal basic feasible solutions to problem

$$\begin{aligned} \min \quad & \sum_{i=1}^n a_i x_i \\ \text{s.t.} \quad & \sum_{i=1}^n x_i \leq \Gamma \\ & 0 \leq x_i \leq 1, \quad i = 1, 2, \dots, n. \end{aligned} \tag{4.22}$$

Let

$$n_S(\mathbf{a}) = |S(\mathbf{a})|,$$

and $\mathbf{s}^q(\mathbf{a})$, $q = 1, 2, \dots, n_S(\mathbf{a})$, be the binary sequence which is placed at position q if we order the binary sequences of set $S(\mathbf{a})$ according to relation “ $<_{\text{lex}}$ ”, i.e.,

$$\mathbf{s}^1(\mathbf{a}) <_{\text{lex}} \mathbf{s}^2(\mathbf{a}) <_{\text{lex}} \dots <_{\text{lex}} \mathbf{s}^{n_S(\mathbf{a})}(\mathbf{a}),$$

where relation “ $<_{\text{lex}}$ ” is defined in Definition 4.

For example, if $\mathbf{a} = (a_1, a_2, \dots, a_6)$ with

$$a_1 < a_2 = a_3 = a_4 < 0 = a_5 = a_6$$

and $\Gamma = 3$, then

$$S(\mathbf{a}) = \{(1, 1, 1, 0, 0, 0), (1, 1, 0, 1, 0, 0), (1, 0, 1, 1, 0, 0)\},$$

$$n_S(\mathbf{a}) = 3,$$

$$\begin{aligned} \mathbf{s}^1(\mathbf{a}) &= (1, 0, 1, 1, 0, 0), \\ \mathbf{s}^2(\mathbf{a}) &= (1, 1, 0, 1, 0, 0), \\ \mathbf{s}^3(\mathbf{a}) &= (1, 1, 1, 0, 0, 0). \end{aligned}$$

Note that if h is a strictly increasing function with $h(0) = 0$, then

$$\begin{aligned} S(\mathbf{h}(\mathbf{a})) &= S(\mathbf{a}), \\ n_S(\mathbf{h}(\mathbf{a})) &= n_S(\mathbf{a}), \\ \mathbf{s}^q(\mathbf{h}(\mathbf{a})) &= \mathbf{s}^q(\mathbf{a}), \quad q = 1, 2, \dots, n_S(\mathbf{a}), \end{aligned}$$

where $\mathbf{h}(\mathbf{a})$ is a vector whose i -th coordinate, $i = 1, 2, \dots, n$, is $h(a_i)$.

We also define functions $U(\boldsymbol{\beta}, \beta_0)$, $n_U(\boldsymbol{\beta}, \beta_0)$, $\mathbf{u}^q(\boldsymbol{\beta}, \beta_0)$, $q = 1, 2, \dots, n_U(\boldsymbol{\beta}, \beta_0)$, $(\boldsymbol{\beta}, \beta_0) \in \mathbb{R}^{m+1}$, as

$$\begin{aligned} U(\boldsymbol{\beta}, \beta_0) &= S(\mathbf{a}), \\ n_U(\boldsymbol{\beta}, \beta_0) &= n_S(\mathbf{a}), \\ \mathbf{u}^q(\boldsymbol{\beta}, \beta_0) &= \mathbf{s}^q(\mathbf{a}), \quad q = 1, 2, \dots, n_S(\mathbf{a}), \end{aligned} \tag{4.23}$$

where $\mathbf{a} = (a_1, a_2, \dots, a_n)$ and

$$a_i = (-1)^{y_i} (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0), \quad i = 1, 2, \dots, n.$$

The next theorem solves the inner minimization problem in Eq. (4.21).

Theorem 6.

(a) *An optimal solution to*

$$Z_3(\boldsymbol{\beta}, \beta_0) = \min_{\Delta \mathbf{X} \in S_1, \Delta \mathbf{y} \in S_2} P(|\mathbf{y} - \Delta \mathbf{y}|, \mathbf{X} + \Delta \mathbf{X}, \boldsymbol{\beta}, \beta_0), \tag{4.24}$$

is $(\Delta \mathbf{X}^o, \Delta \mathbf{y}^o) = (\mathbf{R}(\mathbf{u}^1(\boldsymbol{\beta}, \beta_0)), \mathbf{u}^1(\boldsymbol{\beta}, \beta_0))$,

where $\mathbf{u}^1(\boldsymbol{\beta}, \beta_0)$ is defined in Eq. (4.23), $\mathbf{R}(\mathbf{v})$ is a matrix whose i -th row is vector

$$\mathbf{r}_i(\mathbf{v}) = (-1)^{|y_i - v_i|} \rho \mathbf{f}(\boldsymbol{\beta}, d(p)), \quad i \in \{1, 2, \dots, n\}, \tag{4.25}$$

\mathbf{v} is a length n binary sequence, and $\mathbf{f}(\mathbf{x}, p)$, $\mathbf{x} \in \mathbb{R}^m$, $p \geq 1$, is defined in Appendix A.1.

(b) *The optimal objective value $Z_3(\boldsymbol{\beta}, \beta_0)$ is given by*

$$\begin{aligned} Z_3(\boldsymbol{\beta}, \beta_0) &= \sum_{i=1}^n \left[|y_i - u_i^1(\boldsymbol{\beta}, \beta_0)| \left(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - u_i^1(\boldsymbol{\beta}, \beta_0)|} \rho \|\boldsymbol{\beta}\|_{d(p)} \right) \right. \\ &\quad \left. - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - u_i^1(\boldsymbol{\beta}, \beta_0)|} \rho \|\boldsymbol{\beta}\|_{d(p}))) \right]. \end{aligned} \tag{4.26}$$

Proof.

(a) Problem (4.24) can be expressed as

$$\min_{\Delta \mathbf{y} \in \mathcal{S}_2} \min_{\Delta \mathbf{X} \in \mathcal{S}_1} P(|\mathbf{y} - \Delta \mathbf{y}|, \mathbf{X} + \Delta \mathbf{X}, \boldsymbol{\beta}, \beta_0). \quad (4.27)$$

Using Theorem 4, we conclude that the optimal solution to the inner minimization problem in Eq. (4.27) is $\mathbf{R}(\Delta \mathbf{y})$, and its optimal objective value is

$$\begin{aligned} & \sum_{i=1}^n \left[|y_i - \Delta y_i| (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - \Delta y_i|} \rho \|\boldsymbol{\beta}\|_{d(p)}) \right. \\ & \left. - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - \Delta y_i|} \rho \|\boldsymbol{\beta}\|_{d(p)})) \right]. \end{aligned}$$

Thus, Problem (4.24) is equivalent to

$$\min_{\Delta \mathbf{y} \in \mathcal{S}_2} \sum_{i=1}^n \left[|y_i - \Delta y_i| (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - \Delta y_i|} \rho \|\boldsymbol{\beta}\|_{d(p)}) \right. \\ \left. - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - \Delta y_i|} \rho \|\boldsymbol{\beta}\|_{d(p)})) \right]. \quad (4.28)$$

Defining

$$\begin{aligned} F_{i,0}(\boldsymbol{\beta}, \beta_0) &= y_i (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}) \\ &- \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)})), \quad i \in \{1, 2, \dots, n\}, \end{aligned}$$

$$\begin{aligned} F_{i,1}(\boldsymbol{\beta}, \beta_0) &= (1 - y_i) (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{1-y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}) \\ &- \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{1-y_i} \rho \|\boldsymbol{\beta}\|_{d(p)})), \quad i \in \{1, 2, \dots, n\}, \end{aligned}$$

we observe that

$$\begin{aligned} & |y_i - \Delta y_i| (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - \Delta y_i|} \rho \|\boldsymbol{\beta}\|_{d(p)}) \\ & - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - \Delta y_i|} \rho \|\boldsymbol{\beta}\|_{d(p)})) \\ &= \begin{cases} F_{i,0}(\boldsymbol{\beta}, \beta_0), & \text{if } \Delta y_i = 0, \\ F_{i,1}(\boldsymbol{\beta}, \beta_0), & \text{if } \Delta y_i = 1, \end{cases} \end{aligned}$$

$$= \sum_{i=1}^n F_{i,0}(\boldsymbol{\beta}, \beta_0) + \sum_{i=1}^n (F_{i,1}(\boldsymbol{\beta}, \beta_0) - F_{i,0}(\boldsymbol{\beta}, \beta_0)) \Delta y_i.$$

Problem (4.28) has the same solution as

$$\min_{\Delta \mathbf{y} \in \mathcal{S}_2} \sum_{i=1}^n (F_{i,1}(\boldsymbol{\beta}, \beta_0) - F_{i,0}(\boldsymbol{\beta}, \beta_0)) \Delta y_i. \quad (4.29)$$

Defining

$$h_i(w) = w - (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)} + \ln \left(\frac{1 + \exp((-1)^{y_i} w + (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)})}{1 + \exp((-1)^{y_i} w - (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)})} \right),$$

we express

$$F_{i,1}(\boldsymbol{\beta}, \beta_0) - F_{i,0}(\boldsymbol{\beta}, \beta_0) = h_i((-1)^{y_i} (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0)). \quad (4.30)$$

We observe that

$$\begin{aligned} & \frac{dh_i(w)}{dw} \\ &= 1 + \frac{(-1)^{y_i} (\exp((-1)^{y_i} w + (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}) - \exp((-1)^{y_i} w - (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}))}{(1 + \exp((-1)^{y_i} w + (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}))(1 + \exp((-1)^{y_i} w - (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}))} \\ &= 1 + \frac{\exp((-1)^{y_i} w - (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}) (-1)^{y_i} [\exp(2(-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}) - 1]}{(1 + \exp((-1)^{y_i} w + (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}))(1 + \exp((-1)^{y_i} w - (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}))}. \end{aligned}$$

Since for $y_i = 0$,

$$\exp(2(-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}) - 1 \geq 0,$$

and for $y_i = 1$,

$$\exp(2(-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}) - 1 \leq 0,$$

we have that

$$\frac{dh_i(w)}{dw} > 0,$$

implying that function $h_i(w)$ is strictly increasing in w .

Furthermore,

$$h_i(0) = -(-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)} + \ln \left(\frac{1 + \exp((-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)})}{1 + \exp(-(-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)})} \right) = 0.$$

Given Eq. (4.30), the fact that $h_i(w)$ is strictly increasing in w , and the fact that $h_i(0) = 0$, we conclude that $\mathbf{u}^1(\boldsymbol{\beta}, \beta_0)$, defined in Eq. (4.23), is an optimal solution to Problem (4.29) and that $(\Delta \mathbf{X}^o, \Delta \mathbf{y}^o) = \mathbf{R}(\mathbf{u}^1(\boldsymbol{\beta}, \beta_0), \mathbf{u}^1(\boldsymbol{\beta}, \beta_0))$ is an optimal solution to Problem (4.24).

(b) If we apply $(\Delta \mathbf{X}, \Delta \mathbf{y}) = (\mathbf{R}(\mathbf{u}^1(\boldsymbol{\beta}, \beta_0)), \mathbf{u}^1(\boldsymbol{\beta}, \beta_0))$ to the objective function of (4.24), we obtain the expression in Eq. (4.26) for the optimal objective. \square

Given Theorem 6, Problem (4.21) has a robust counterpart

$$\max_{\boldsymbol{\beta}, \beta_0} Z_3(\boldsymbol{\beta}, \beta_0). \quad (4.31)$$

Function $P(|\mathbf{y} - \Delta \mathbf{y}|, \mathbf{X} + \Delta \mathbf{X}, \boldsymbol{\beta}, \beta_0)$ is concave in $(\boldsymbol{\beta}, \beta_0)$ for any $\Delta \mathbf{X} \in S_1$, $\Delta \mathbf{y} \in S_2$ (see Ryan [40], Hosmer [28]). Thus, function

$$Z_3(\boldsymbol{\beta}, \beta_0) = \min_{\Delta \mathbf{x} \in S_1, \Delta \mathbf{y} \in S_2} P(|\mathbf{y} - \Delta \mathbf{y}|, \mathbf{X} + \Delta \mathbf{X}, \boldsymbol{\beta}, \beta_0),$$

which is equal to the objective function of Problem (4.31), is concave in $(\boldsymbol{\beta}, \beta_0)$ (see Boyd [12], p. 81). Problem (4.31) is an unconstrained concave maximization problem.

Since function $Z_3(\boldsymbol{\beta}, \beta_0)$ depends on $\mathbf{u}^1(\boldsymbol{\beta}, \beta_0)$, which in turn depends on the ordering of $(-1)^{y_i}(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0)$, $i \in \{1, 2, \dots, n\}$, it is not differentiable at all $(\boldsymbol{\beta}, \beta_0) \in \mathbb{R}^{m+1}$. Hence, we need to calculate left and right derivatives, which is conceptually simple, but notationally cumbersome, and therefore, we present the details of the calculation in Appendix C.

Since Problem (4.31) is an unconstrained maximization problem with a concave objective function, it can be solved using the subgradient method. The details of the method are outlined in Appendix B.3. Note that if $\rho = 0$, we have another method to calculate the robust logistic regression under response variable uncertainty estimate. In this method, we calculate the objective function of the inner minimization problem in Eq. (4.13), and apply the subgradient method to find its optimal value, whereas the method in Section 4.3 computes the dual of the inner minimization problem in Eq. (4.13), and then, unifies the outer with the inner maximization problems.

4.5 Experimental Results

In this section, we report computational results on the performance of the methods outlined in Sections 4.2, 4.3 and 4.4 involving both artificial and real data sets. The results were produced using Matlab 7.0.1, where the algorithms for calculating the robust and the classical estimates were coded.

4.5.1 Artificial Data Sets

To evaluate the proposed robust estimates, we produced an artificial data set in the following way:

- A set of 100 points in \mathbb{R}^3 obeying the multivariate normal distribution with mean $[1, 0, 0]^T$ and covariance matrix $\frac{1}{\sqrt{2}}\mathbf{I}_3$ was generated, where \mathbf{I}_3 is the 3×3 identity matrix. The points were associated with $y = 1$, and added to the data set.
- A set of 100 points in \mathbb{R}^3 obeying the multivariate normal distribution with mean $[0, 1, 0]^T$ and standard deviation $\frac{1}{\sqrt{2}}\mathbf{I}_3$ was generated. The points were associated with $y = 0$, and added to the data set.

The generated data set was normalized by scaling each one of the vectors containing the data corresponding to an independent variable to make their 2-norm equal to 1.

The performances of the classical logistic regression and the robust logistic regression under independent variables uncertainty estimates were measured for various values of ρ according to the following procedure:

1. The normalized data set was divided randomly into two groups containing the 50% of the samples each, the training set and the testing set.
2. A set of 100 random data points in \mathbb{R}^3 following the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix \mathbf{I}_3 was produced. These data points were scaled by ρ and added to the training set data points to contaminate them.
3. The contaminated data were used to produce the estimates to be studied.
4. The prediction success rate in the testing set was recorded for each estimate.
5. The procedure was repeated 30 times and the average performance of each estimate was recorded.

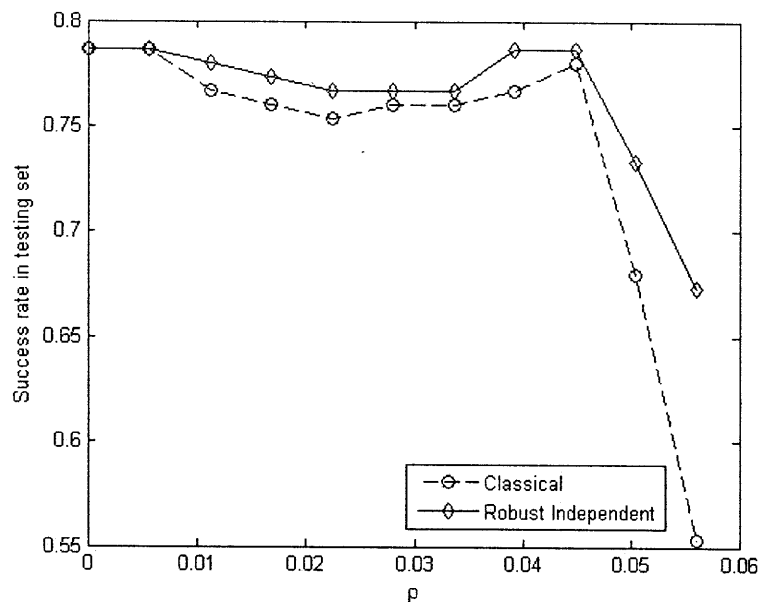


Figure 4-1: Success rate of the robust logistic regression under independent variables uncertainty estimate in the testing set.

Figure 4-1 illustrates the results of the experiment. We observe that the robust estimate is always superior to the classical one. As ρ increases, the success rate for both estimates drops, and the difference between the estimates grows.

To evaluate the performance of the robust logistic regression under response variable uncertainty estimate for various values of Γ , we followed the same procedure. The contamination of the response variable data was simulated by producing a random permutation of the 200 samples of the training set and changing the value of the response variable for the first Γ samples in the permutation. Figure 4-2 illustrates the results. The robust estimate yields a better success rate and its performance drops at a lower rate as Γ increases.

To evaluate the performance of the globally robust logistic regression estimate for various values of ρ and Γ , we followed the same procedure as for the other robust estimates and produced contamination of both the independent and response variables in the same way. Figures 4-3 and 4-4 illustrate the results. For both values of Γ , the robust estimates have a higher success rate than the classical one.

4.5.2 Real Data Sets

The following real data sets were used to test the performance of the robust logistic regression estimates:

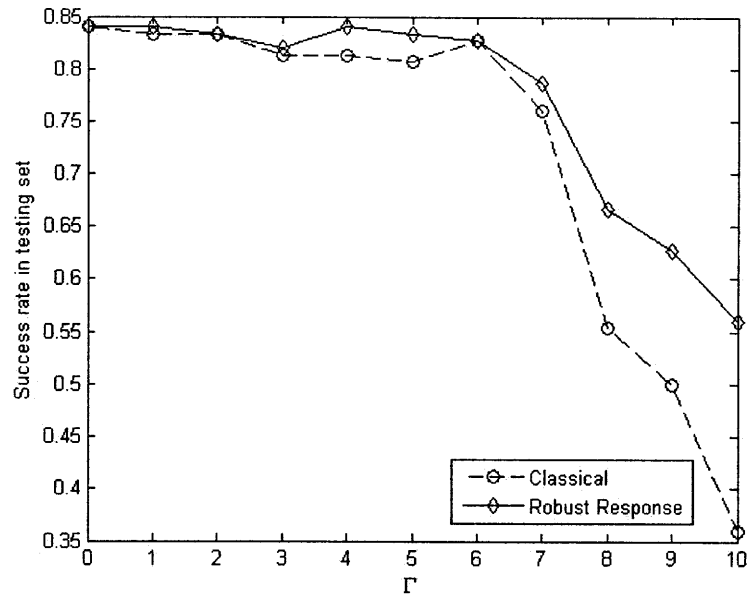


Figure 4-2: Success rate of the robust logistic regression under response variable uncertainty estimate in the testing set.

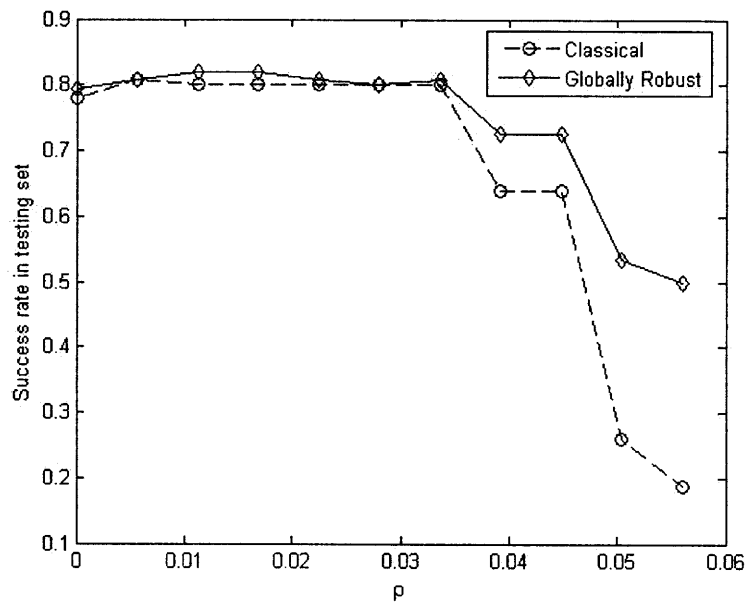


Figure 4-3: Success rate of the globally robust logistic regression estimate in the testing set ($\Gamma = 1$).

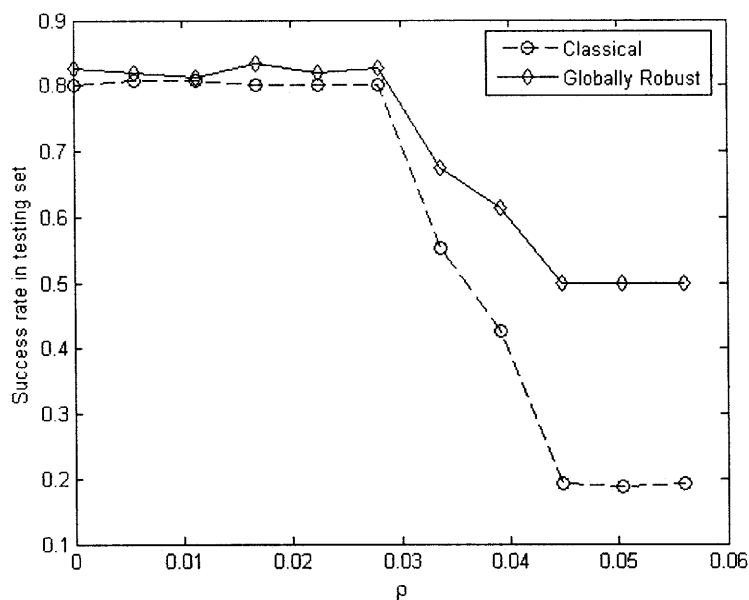


Figure 4-4: Success rate of the globally robust logistic regression estimate in the testing set ($\Gamma = 2$).

1. “Banks” from N. R. Patel [35]. The independent variables are the ratio of the total loans and leases over the total assets and the ratio of the total expenses over the total assets of the bank. The response variable is equal to 1 or 0, according to whether the bank is financially strong or weak respectively. It consists of 20 samples.
2. “WDBC”, the Breast Cancer Wisconsin Diagnostic Data Set, from the UCI Repository [3]. The data set consists of 569 samples, each one referring to a patient who was examined for breast cancer. The independent variables are 30, the mean, the standard deviation and the mean of the 3 largest values for 10 real-valued characteristics of the cell nucleus. The response variable is 0 or 1, according to whether the person suffers from the disease or not respectively.
3. “Chocs” from SAS Real Data Sets [2], [42]. To produce the data, 10 people were given 8 chocolate bars with varying characteristics. After eating the bars, they were asked to choose their favorite. The independent variables are DARK, which denotes if it is milk or dark chocolate, SOFT, which denotes if it has a hard or soft center, and NUTS, which denotes whether it has nuts. Each person were given a chocolate for each one of the eight combinations of those 3 variables, and their preference or no preference was listed. There is a sample for each combination of a person and a chocolate category, making $8 \cdot 10 = 80$ samples totally.

Data set	n	m
Banks	20	2
Breast	569	30
Chocs	80	3
Crash	58	3
PTSD	948	5

Table 4.1: Sizes of real data sets.

4. “Crash dummies” from SAS Real Data Sets [2], [42]. It provides information for the calibration of crash dummies in automobile safety tests. There are 3 independent variables, the acceleration, the velocity, and the designed “age” of the crash dummy. It consists of 58 samples.
5. “PTSD” from SAS Real Data Sets, [2], [42]. The samples were generated by interviewing 316 people who survived residential fires in the Philadelphia area at 3, 6, and 12 months after the fire. The dependent variable PTSD is coded 1, if the person had symptoms of post-traumatic stress disorder, and 0, otherwise. The independent variables are CONTROL, a scale of a person’s perceived control over several areas of life, PROBLEMS, the total number of problems reported in several areas of life, SEVENT, the number of stressful events reported since the last interview, COHES, a scale of family cohesion and TIME, equal to 1, 2, or 3 if the interview is 3, 6 or 12 months after the fire respectively. There is a sample for each combination of a person and an interview, making $3 \cdot 316 = 948$ samples in total.

The sets were normalized by scaling each one of the vectors containing the data corresponding to an independent variable to make their 2-norm equal to 1. The sizes of the used data sets can be seen in Table 4.1.

The evaluation procedure for each real data set was the following:

- The data set was divided in three sets, the training set, consisting of the 50% of the samples, the validating set, consisting of the 25% of the samples, and the testing set, consisting of the rest 25% of the samples. We randomly selected 30 different partitions of the data set.
- For each one of the considered partitions of the data set:
 - The classical logistic regression estimate based on the training set was calculated.
 - The robust logistic regression under independent variables uncertainty estimates based on the training set for various values of ρ were calculated. For each ρ ,

Data set	Classical	Robust Indep	Robust Resp	Glob Robust
Banks	0.7120	0.7133	0.7667*	0.7133
Breast	0.6182	0.6589*	0.6378	0.6421
Chocs	0.7833	0.8509	0.8502	0.8517*
Crash	0.8098	0.8133	0.8444*	0.8133
PTSD	0.6861	0.6861	0.6987*	0.6861

Table 4.2: Success rate in testing set for real data sets. * denotes the estimate with the best performance.

the success rate on the validating set was measured, and the ρ with the highest performance on the validating set was considered. The success rate that this ρ yielded on the testing set was recorded.

- The robust logistic regression under response variable uncertainty estimates based on the training set for various values of Γ were calculated and the same procedure was followed to determine the Γ which was used to calculate the success rate of this estimate in the testing set.
- The same procedure involving combinations of values of ρ and Γ was used to determine the success rate of the globally robust logistic regression estimate.
- The success of the estimates under examination were averaged over the partitions of the data considered.

The results of the evaluation process are summarized in Table 4.2. The robust estimates have a higher success rate than the classical ones. In three of the data sets, the estimate that is shielded against errors in the response variable has the highest success rate. The improvement that the robust estimates yield ranges from 1.84% in the “PTSD” data set to 8.73% in the “Chocs” data set.

4.6 Conclusions

In this chapter, the robust optimization paradigm was applied in producing estimates for the logistic regression model. Uncertainty sets for the independent variables, the response variable, and both of them were considered. The robust logistic regression estimates improved the success rate in the out-of-sample prediction compared to the classical one in both artificial and real data sets.

Chapter 5

Robust Maximum Likelihood In Normally Distributed Data

5.1 Introduction

In Chapter 4, we defined robust estimators for logistic regression. In this chapter, we develop a robust maximum likelihood estimator for the multivariate normal distribution. We develop an efficient algorithm to calculate the robust estimator. More specifically, in Section 5.2, we formally define the estimator, and provide a first order gradient descent method to calculate it. In Section 5.3, we use computer generated data to show the efficiency of the robust estimator in providing accurate predictions even in the presence of errors. We test and compare the nominal and the robust estimator under various circumstances.

5.2 Method

Consider samples \mathbf{x}_i , $i = 1, 2, \dots, n$, for which it is known that they follow a multivariate normal distribution parametrized by its mean $\boldsymbol{\mu} \in \mathbb{R}^m$ and its covariance matrix $\boldsymbol{\Sigma}$, a symmetric positive semidefinite $m \times m$ matrix. Multivariate normal distributions arise in many practical situations, such as in studying large populations. The maximum likelihood method receives this set of samples and returns estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ for parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Let $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x})$ be the probability density function for the data:

$$f(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x}) = \frac{1}{(2\pi)^{m/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (5.1)$$

If the values of the parameters used to generate the samples are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and the samples

\mathbf{x}_i are independent, the probability density function evaluated at them is $\prod_{i=1}^n f(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x}_i)$. The maximum likelihood method returns the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that maximize this density, or equivalently the logarithm of this density.

Let \mathbf{X} be the $n \times m$ matrix having sample \mathbf{x}_i in its i -th row. If we define

$$\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}) \equiv \log \left(\prod_{i=1}^n f(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x}_i) \right) = \max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \sum_{i=1}^n \log(f(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x}_i)), \quad (5.2)$$

the maximum likelihood estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the solution to

$$\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}). \quad (5.3)$$

The solution to Problem (5.3) is proved to be given by

$$\hat{\boldsymbol{\mu}}_{\text{nom}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i, \quad (5.4)$$

$$\hat{\boldsymbol{\Sigma}}_{\text{nom}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{nom}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{nom}})^T, \quad (5.5)$$

see [31].

Maximum likelihood seeks the values of the parameters which maximize the density evaluated at the observed samples. It is a Bayesian method, assuming no prior knowledge on the parameters, and thus, concluding that the parameters maximizing the density of the data have the highest probability to be the true parameters. Following the robust optimization paradigm, we consider an uncertainty set for the samples and define the robust normal distribution estimate to be the value of the parameter which maximizes the worst-case density. In this way, we seek an estimate which is secured against errors residing in the uncertainty set. Let $\mathbf{x}_i, i = 1, 2, \dots, n$, be the samples contaminated with errors, and \mathbf{X} be the $n \times m$ matrix having the contaminated sample \mathbf{x}_i in its i -th row. We use the decision variable $\boldsymbol{\Delta}\mathbf{x}_i, i = 1, 2, \dots, n$ to model the error of the i -th sample. The $n \times m$ matrix $\boldsymbol{\Delta}\mathbf{X}$ has $\boldsymbol{\Delta}\mathbf{x}_i$ as its i -th row. Then, the true value of sample i is $\mathbf{x}_i + \boldsymbol{\Delta}\mathbf{x}_i$. Formally, the robust normal distribution estimator is the solution to

$$\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \min_{\boldsymbol{\Delta}\mathbf{X} \in \mathcal{N}} \psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \boldsymbol{\Delta}\mathbf{X}) = \max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \min_{\boldsymbol{\Delta}\mathbf{X} \in \mathcal{N}} \sum_{i=1}^n \log(f(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x}_i + \boldsymbol{\Delta}\mathbf{x}_i)), \quad (5.6)$$

where \mathcal{N} is given by:

$$\mathcal{N} = \left\{ \Delta \mathbf{X} = \begin{bmatrix} \Delta \mathbf{x}_1 \\ \Delta \mathbf{x}_2 \\ \vdots \\ \Delta \mathbf{x}_n \end{bmatrix} \mid \|\Delta \mathbf{x}_i\|_2 \leq \rho, i = 1, 2, \dots, n \right\}. \quad (5.7)$$

Through this uncertainty set, the error for each sample is restricted to be in a ball with radius ρ . Notice that there is no correlation among the errors of the samples, which is reasonable for errors in reality.

We define $\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})$ to be the worst-case density, which is the objective value of the inner minimization problem. Formally,

$$\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}) \equiv \min_{\Delta \mathbf{X} \in \mathcal{N}} \psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}). \quad (5.8)$$

Knowing how to calculate $\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})$ and at least its first derivative at any $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ enables us to compute the robust estimates, using a gradient ascent method.

Consider the inner minimization problem in Eq. (5.6) where uncertainty set \mathcal{N} is defined in Eq. (5.7):

$$\begin{aligned} \phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}) &= \min_{\Delta \mathbf{X} \in \mathcal{N}} \psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}) \\ &= \min_{\|\Delta \mathbf{x}_i\|_2 \leq \rho} -\frac{nm}{2} \log(2\pi) - \frac{n}{2} \log |\boldsymbol{\Sigma}| \\ &\quad + \sum_{i=1}^n -\frac{1}{2} (\mathbf{x}_i + \Delta \mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i + \Delta \mathbf{x}_i - \boldsymbol{\mu}). \end{aligned} \quad (5.9)$$

Problem (5.9) has separable objective function and separable constraints in $\Delta \mathbf{x}_i$, $i = 1, 2, \dots, n$. Thus, to solve it, it suffices to solve

$$\min_{\|\Delta \mathbf{x}_i\|_2 \leq \rho} -\frac{1}{2} (\mathbf{x}_i + \Delta \mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i + \Delta \mathbf{x}_i - \boldsymbol{\mu}) \quad (5.10)$$

for each $i = 1, 2, \dots, n$.

Since the objective function of Problem (5.10) can be written as

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x}_i + \Delta \mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i + \Delta \mathbf{x}_i - \boldsymbol{\mu}) \\ &= -\frac{1}{2} \Delta \mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \Delta \mathbf{x}_i - [\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})]^T \Delta \mathbf{x}_i - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}), \end{aligned}$$

it is a trust region problem. This is solved via its convex dual, which has 0 duality gap [12]. To express the dual, we introduce some special notation. Consider the spectral decomposition of matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \sum_{j=1}^m \eta_j \mathbf{q}_j \mathbf{q}_j^T. \quad (5.11)$$

Let η_{\min} be the minimum eigenvalue of $\boldsymbol{\Sigma}$, or, more formally,

$$\eta_{\min} \equiv \min_{j \in \{1, 2, \dots, m\}} \{\eta_j\}. \quad (5.12)$$

Let H_{\min} be the set of indices corresponding to the minimum eigenvalues of $\boldsymbol{\Sigma}$

$$H_{\min} \equiv \{j \in \{1, 2, \dots, m\} \mid \eta_j = \eta_{\min}\}. \quad (5.13)$$

Define

$$\Lambda_i = \begin{cases} \left[\frac{1}{2\eta_{\min}}, +\infty \right), & \text{if for all } j \in H_{\min}, (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{q}_j = 0, \\ \left(\frac{1}{2\eta_{\min}}, +\infty \right), & \text{otherwise,} \end{cases} \quad (5.14)$$

and

$$w_i(\lambda) = -\sum_{j=1}^m \frac{[(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{q}_j]^2}{2\eta_j(2\lambda\eta_j - 1)} - \lambda\rho^2 - \sum_{j=1}^m \frac{[(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{q}_j]^2}{2\eta_j}, \quad \lambda \in \left[\frac{1}{2\eta_{\min}}, +\infty \right), \quad (5.15)$$

$$i = 1, 2, \dots, n,$$

where

$$\frac{[(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{q}_j]^2}{2\eta_j(2\lambda\eta_j - 1)} = \begin{cases} 0, & \text{if } (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{q}_j = 0 \text{ and } 2\lambda\eta_j - 1 = 0, \\ +\infty, & \text{if } (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{q}_j \neq 0 \text{ and } 2\lambda\eta_j - 1 = 0. \end{cases}$$

Using this notation, a dual to Problem (5.10) is

$$\begin{aligned} \max \quad & w_i(\lambda) \\ \text{s.t.} \quad & \lambda \in \Lambda_i, \end{aligned} \tag{5.16}$$

see [12]. Problem (5.16) can be solved using a gradient ascent method to obtain the optimal solution λ_i^* .

Since the derivative of the objective function of Problem (5.10) is given by

$$\nabla_{\Delta \mathbf{x}_i} \left(-\frac{1}{2}(\mathbf{x}_i + \Delta \mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i + \Delta \mathbf{x}_i - \boldsymbol{\mu}) \right) = -\boldsymbol{\Sigma}^{-1} \Delta \mathbf{x}_i - \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}),$$

and the derivative of its constraint is $2\Delta \mathbf{x}_i$, using the Karush-Kuhn-Tucker conditions [12], we conclude that any $\mathbf{z} \in \mathbb{R}^m$ that satisfies

$$-\boldsymbol{\Sigma}^{-1} \mathbf{z} - \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) + 2\kappa \mathbf{z} = 0, \tag{5.17}$$

$$\mathbf{z}^T \mathbf{z} = \rho^2, \tag{5.18}$$

$$\kappa \geq 0 \tag{5.19}$$

is an optimal solution to Problem (5.10). We will use these conditions to express an optimal solution to Problem (5.10). To achieve this, we are going to introduce some special notation. Define

$$K_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \{j \in \{1, 2, \dots, m\} \mid 2\lambda_i^* \eta_j - 1 > 0\}, \quad i = 1, 2, \dots, n, \tag{5.20}$$

$$u_{i,j} = \begin{cases} \frac{(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{q}_j}{2\lambda_i^* \eta_j - 1}, & j \in K_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ \left(\rho^2 - \sum_{j \in K_i(\boldsymbol{\mu}, \boldsymbol{\Sigma})} u_{i,j}^2 \right)^{1/2}, & j = \min(\{1, 2, \dots, m\} \setminus K_i(\boldsymbol{\mu}, \boldsymbol{\Sigma})), \\ 0, & \text{otherwise,} \end{cases} \tag{5.21}$$

$$i = 1, 2, \dots, n, \quad j = 1, 2, \dots, m,$$

and

$$\Delta \mathbf{x}_i^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{j=1}^m u_{i,j} \mathbf{q}_j, \quad i = 1, 2, \dots, n. \tag{5.22}$$

We can easily verify that $\Delta \mathbf{x}_i^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ satisfies the Karush-Kuhn-Tucker conditions (5.17), (5.18), (5.19), for $\kappa = \lambda_i^*$ and, thus, is an optimal solution to Problem (5.10). Let $\Delta x_{i,j}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the j -th element of $\Delta \mathbf{x}_i^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Delta \mathbf{X}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the $n \times m$ matrix having $\Delta \mathbf{x}_i^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in its rows. Since Problem (5.9) is separable in $\Delta \mathbf{x}_i$, $\Delta \mathbf{X}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is an optimal solution to Problem (5.9).

We have developed a method to calculate $\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})$ using

$$\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}) = \psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})). \quad (5.23)$$

We now show how we can use $\Delta \mathbf{X}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \Delta \mathbf{X})$ to calculate the derivatives $\nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}))$. We will prove that:

$$\nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})) = \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*)), \quad (5.24)$$

where $\Delta \mathbf{X}^* = \Delta \mathbf{X}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\Delta \mathbf{x}_i^*$ be the i -th row of $\Delta \mathbf{X}^*$ and $\Delta x_{i,j}$ the j -th element of the i -th row of it.

Using the chain rule of derivation, we have that:

$$\begin{aligned} \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})) &= \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}))) \\ &= \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*)) + \sum_{i=1}^n \sum_{j=1}^m \frac{\partial \psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*)}{\partial \Delta x_{i,j}} \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\Delta x_{i,j}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \end{aligned}$$

Since $\Delta \mathbf{X}^*$ is an optimal solution to the inner Problem (5.9), there exist $\nu_1, \nu_2, \dots, \nu_n \geq 0$, such that:

$$\frac{\partial \psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*)}{\partial \Delta x_{i,j}} + 2\nu_i \Delta x_{i,j}^* = 0.$$

Thus,

$$\begin{aligned} \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})) &= \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*)) - \sum_{i=1}^n \nu_i \sum_{j=1}^m 2\Delta x_{i,j}^* \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\Delta x_{i,j}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \\ &= \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*)) - \sum_{i=1}^n \nu_i \sum_{j=1}^m \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\Delta x_{i,j}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}))^2 \end{aligned}$$

$$\begin{aligned}
&= \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*)) - \sum_{i=1}^n \nu_i \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \|\Delta \mathbf{x}_i^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})\|^2 \quad (\|\Delta \mathbf{x}_i^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})\|^2 = \rho^2) \\
&= \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*)).
\end{aligned}$$

In this way, we can calculate the derivatives of $\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})$ using $\Delta \mathbf{X}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and Eq. (5.24), and use a gradient ascent method to calculate the robust estimator.

The objective function $\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})$ is not convex in both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, because its second derivative is not a positive semidefinite matrix. Thus, the solution algorithm converges to a local minimum. In summary, the algorithm to calculate the robust normal distribution estimator defined in Eq. (5.6) is:

1. Start with some initial $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.
2. For each $i = 1, 2, \dots, n$, find the optimal solution $\Delta \mathbf{x}_i^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of Problem 5.10, via its dual Problem (5.16).
3. Use

$$\Delta \mathbf{X}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \begin{bmatrix} \Delta \mathbf{x}_1^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \Delta \mathbf{x}_2^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \dots \\ \Delta \mathbf{x}_n^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{bmatrix}$$

and Eqs. (5.23) and (5.24) to calculate $\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})$ and its derivative.

4. Update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ using the descent direction given by the derivative of $\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})$, until the norm of the derivative is smaller than some tolerance parameter ϵ .

The convergence of the algorithm is linear, since it is a first order method.

We investigate the possibility of using second order methods. If we apply the chain rule to obtain the second derivatives of $\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})$, we obtain:

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}^2(\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})) &= \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}^2(\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}))) \\
&= \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}^2(\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*)) \\
&+ \sum_{i=1}^n \sum_{j=1}^m \frac{\partial \nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X} + \Delta \mathbf{X}^*))}{\partial \Delta x_{i,j}} (\nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\Delta x_{i,j}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma})))^T.
\end{aligned}$$

We observe that the second derivative consists of two terms. The first term is the second derivative of the objective function of the nominal problem and the second term accounts for the dependence of the worst case error $\Delta \mathbf{X}^*$ on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The calculation of $\nabla_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\Delta x_{i,j}^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}))$ complicates the calculation of the second derivatives of $\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X})$.

5.3 Experiments

To evaluate the robust estimator, we conduct experiments using computer generated random data. We generate samples following a multivariate normal distribution. As our estimators are designed to deal with errors in the samples, we generate errors following both a normal and a uniform distribution, and use them to contaminate our samples. First of all, the worst-case and the average value of the probability density function is calculated for both the nominal and the robust estimator. Furthermore, the nominal and the robust estimator are calculated on the true and the contaminated samples and their performance in prediction is evaluated. Finally, a comparison in the performance between the cases of normally and uniformly distributed errors is conducted.

More specifically, the following process is followed in conducting the experiments. A number of $n = 400$ samples in \mathbb{R}^4 following the multivariate normal distribution with some random mean and some random covariance matrix is generated randomly. Let \mathbf{X}^{true} be the 400×4 matrix having the samples in its rows, and $\mathbf{x}_i^{\text{true}}$, $i = 1, 2, \dots, 400$, be the samples. $\mathbf{x}_i^{\text{true}}$ are the true samples, which are not affected by any errors.

Also, we generate errors for the samples in the following way. $\Delta \mathbf{X}_k$, $k = 1, 2, \dots, 40$, is a 400×4 matrix containing errors corresponding to the samples in the 400×4 matrix \mathbf{X} . The errors in $\Delta \mathbf{X}_k$ follow the normal distribution with mean $\mathbf{0}$ and covariance matrix \mathbf{I}_4 , where $\mathbf{0}$ is the zero vector in \mathbb{R}^4 and \mathbf{I}_4 is the 4×4 identity matrix. The reason we use normally distributed errors is that most real errors are closely related to this distribution.

We are going to evaluate the performance of the estimators using the worst-case and average value of the probability density, as well as their distance from the value of the nominal estimator on the true data. Initially, we are going to use the normally distributed errors. In the end, we are going to compare the results with the case that we have uniformly distributed errors.

The experimental section is organized as follows. In Section 5.3.1, we evaluate the estimators based on the worst-case and average value of the probability density. In Section 5.3.2, we evaluate the estimators based on their distance from the nominal estimator on the true data. In Section 5.3.3, we compare to the case that we have uniformly distributed errors.

5.3.1 Worst-Case and Average Probability Density

As defined in Eq. (5.6), the robust estimator maximizes the worst-case density evaluated on the observed samples. Therefore, in order to check the efficiency of the estimator, we are going to check the worst-case and the average value of the probability density, as we add properly scaled errors from the set denoted by $\Delta\mathbf{X}_k$ to the true values of the data.

In particular, we calculate the robust maximum likelihood estimate $\hat{\boldsymbol{\mu}}_{\text{rob}}(\mathbf{X}^{\text{true}}, \rho)$, $\hat{\boldsymbol{\Sigma}}_{\text{rob}}(\mathbf{X}^{\text{true}}, \rho)$, on the true data \mathbf{X}^{true} , for the values of ρ varying between 0 and 3 with a step of 0.1. For $\rho = 0$, we have the nominal estimates $\hat{\boldsymbol{\mu}}_{\text{nom}}(\mathbf{X}^{\text{true}})$, $\hat{\boldsymbol{\Sigma}}_{\text{nom}}(\mathbf{X}^{\text{true}})$. To calculate the nominal estimates, we use Eqs. (5.4) and (5.5). To calculate the robust estimates, we use a first order gradient descent method with initial point the robust estimate for the previous value of ρ , in the considered ρ sequence.

For each estimate $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ that we calculate, we compute the probability density of the observed samples $\psi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}; \mathbf{X}^{\text{true}} + \alpha\rho\Delta\mathbf{X}_k)$, $k = 1, 2, \dots, 40$, where ρ is the same parameter as the one used to compute the robust estimate. We record the worst-case value, as well as the average value over the set of errors indexed by k . We consider the cases $\alpha = 0.5$, $\alpha = 1.0$, and $\alpha = 1.5$.

Figure 5-1 shows the results. As we observe, for small values of ρ , the nominal and the robust estimator depict almost the same performance. As ρ grows, the difference between them increases, with the robust always showing a better performance than the nominal. This is true for both the worst-case and the average value of the probability density. It can be explained by the observation that for larger errors, the robust has an advantage, because it always considers the worst-case. The superiority of the robust is detected for values of ρ greater than or equal to 1. The robust is better than the nominal up to a factor of 10%. As α increases, both nominal and robust performances deteriorate at a higher rate.

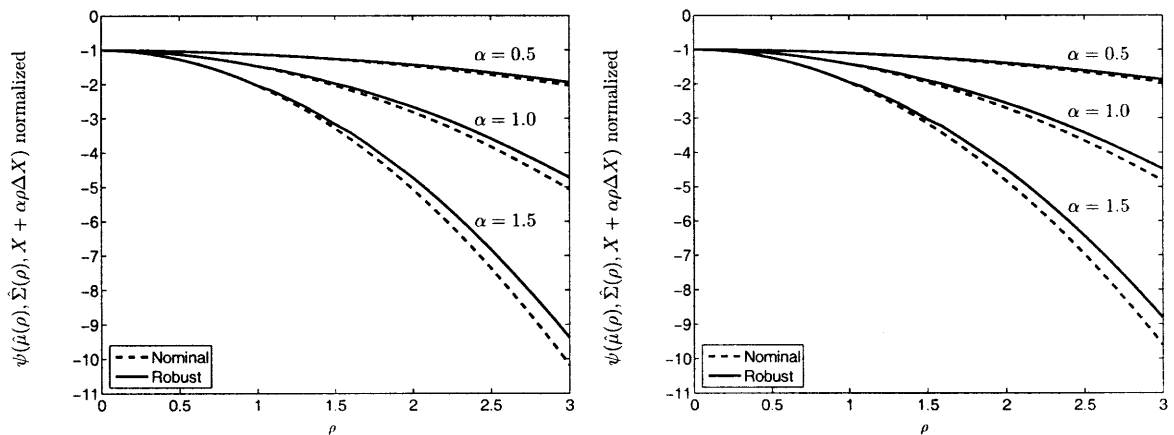


Figure 5-1: Worst-case (left) and average (right) value of ψ , normal errors

5.3.2 Distance From the Nominal Estimator

The purpose of defining the robust estimator is to be able to deal with errors. Thus, to evaluate its performance, we compute both the nominal and the robust estimator on contaminated data having errors of various magnitudes. Then, we compare it to the nominal estimator computed on the true data, which is the estimator we would get if there were no errors.

More specifically, we compute the robust maximum likelihood estimators $\hat{\boldsymbol{\mu}}_{\text{rob}}(\mathbf{X}^{\text{true}} + \delta\Delta\mathbf{X}_k, \rho)$, $\hat{\boldsymbol{\Sigma}}_{\text{rob}}(\mathbf{X}^{\text{true}} + \delta\Delta\mathbf{X}_k, \rho)$ on the contaminated data $\mathbf{X}^{\text{true}} + \delta\Delta\mathbf{X}_k$, for the sets of errors $k = 1, 2, \dots, 40$, for the values of δ ranging between 0 and 1 with a step of 0.05, and for the values of ρ ranging between 0 and 3 with a step of 0.1. For $\rho = 0$, we have the nominal estimators $\hat{\boldsymbol{\mu}}_{\text{nom}}(\mathbf{X}^{\text{true}} + \delta\Delta\mathbf{X}_k)$, $\hat{\boldsymbol{\Sigma}}_{\text{nom}}(\mathbf{X}^{\text{true}} + \delta\Delta\mathbf{X}_k)$.

For each estimate that we computed, we calculate its distance from the nominal estimate on the true data

$$\|\hat{\boldsymbol{\mu}}_{\text{rob}}(\mathbf{X}^{\text{true}} + \delta\Delta\mathbf{X}_k, \rho) - \hat{\boldsymbol{\mu}}_{\text{nom}}(\mathbf{X}^{\text{true}})\|_2$$

and

$$\|\hat{\boldsymbol{\Sigma}}_{\text{rob}}(\mathbf{X}^{\text{true}} + \delta\Delta\mathbf{X}_k, \rho) - \hat{\boldsymbol{\Sigma}}_{\text{nom}}(\mathbf{X}^{\text{true}})\|_{\text{fro}}.$$

Note that $\|A\|_{\text{fro}}$ is the Frobenius norm of an $n \times m$ matrix A defined by

$$\|A\|_{\text{fro}} = \sqrt{\sum_{i=1}^n \sum_{j=1}^m A_{i,j}^2},$$

where $A_{i,j}$ is the (i, j) -element of matrix A . We average the calculated distances over the error sets k , $k = 1, 2, \dots, 40$. We use the Frobenius norm because it takes into consideration the differences of the variances of the variables, as well as their cross-correlation terms.

Figure 5-2 shows the results. The performance of the nominal estimator is the one showed for $\rho = 0$. In all cases, for an interval of ρ starting from 0, the robust shows an almost stable performance, equal to the nominal one. As ρ grows, the performance of the robust improves compared to the nominal one up to some point. There is an interval of ρ , where the robust is up to 15% better than the nominal. Then, the performance of the robust deteriorates significantly. This can be expected, because, in this interval, ρ is big compared to the magnitudes of the errors that are added to the true samples and the robust estimator becomes very conservative.

The area where the robust outperforms the nominal depends on δ , the size of the errors. As δ increases, the interval where the robust shows increased performance moves to the right. This is explained by the fact that the robust estimator is secured against errors with norm up to some ρ , and thus, it cannot deal with higher errors. The errors in μ and Σ estimation show the same qualitative patterns, as we would expect.

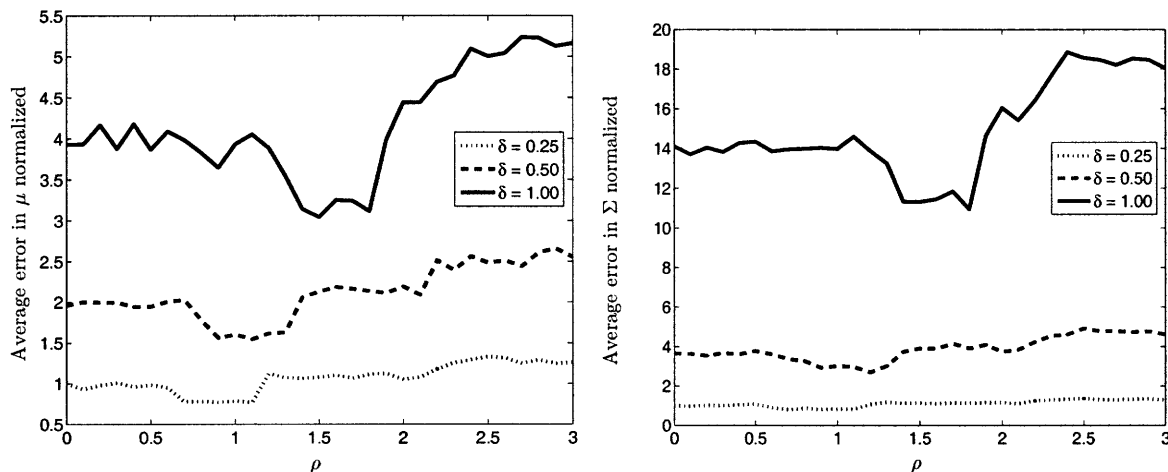


Figure 5-2: Error in μ (left) and Σ (right), normal errors

5.3.3 Comparison of the Error Distributions

To check the dependance of our observations on the distribution that the errors follow, we conduct the same experiments using uniformly distributed errors. Now, $\Delta \mathbf{X}_k$, $k = 1, 2, \dots, 40$, is a 400×4 matrix, where each of its rows follows the uniform distribution in the ball with radius 1. The uniform distribution is also closely related to real errors.

In Figure 5-3, we can see the performances of the nominal and the robust estimators in the case of uniformly distributed errors. The same patterns, as in the case of the normally distributed errors, apply. However, by comparing Figure 5-3 to the respective one for normally distributed errors Figure 5-2, we observe that for the same δ , the region where the robust is superior is moved to the left. This is explained by the fact that the uniform distribution has its samples concentrated in the ball with radius δ , whereas the normal distribution can have samples outside of this region.

5.4 Conclusions

In this chapter, we defined a robust maximum likelihood estimator in normally distributed data to deal with errors in the observed samples, based on the robust optimization principles.

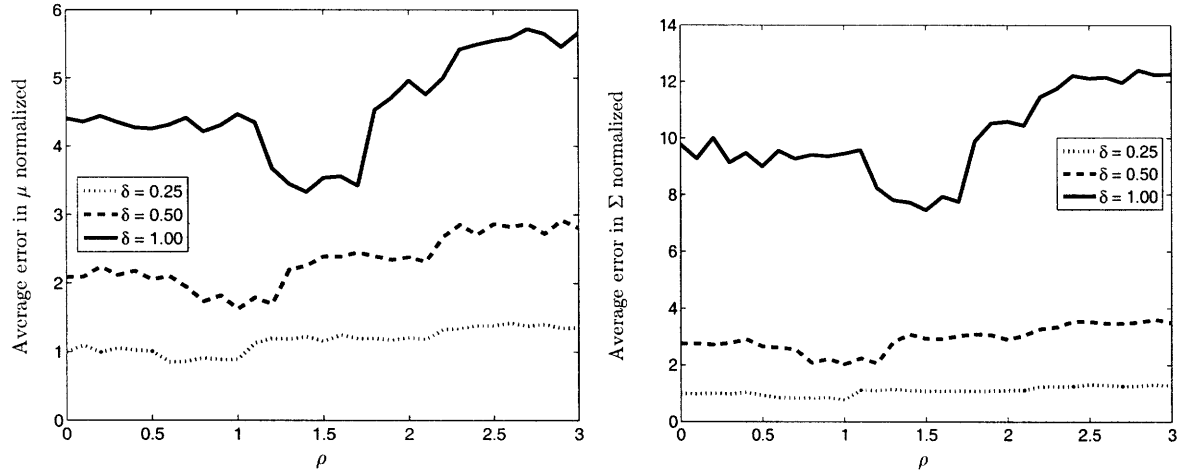


Figure 5-3: Error in μ (left) and Σ (right), uniform errors

We developed an algorithm to efficiently calculate the robust estimator. We conducted extensive experiments to compare the performance of the robust estimator to their respective nominal one, which confirmed that the robust estimator, when correctly tuned, can be resistant to errors.

Appendix A

Algebraic Propositions

We define function $\mathbf{f}(\mathbf{x}, p) \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^m$, $p \geq 1$, which is used in vector calculations, and prove some propositions on the matrix norms which are considered in the uncertainty sets.

A.1 Properties of function $\mathbf{f}(\mathbf{x}, p)$

Definition 5. Function $\mathbf{f}(\mathbf{x}, p) \in \mathbb{R}^m$, where $\mathbf{x} \in \mathbb{R}^m$, and $p \geq 1$, is defined by

$$[\mathbf{f}(\mathbf{x}, p)]_j = \begin{cases} \operatorname{sgn}(x_j) \left(\frac{|x_j|}{\|\mathbf{x}\|_p} \right)^{p-1}, & \text{if } \mathbf{x} \neq \mathbf{0}, \\ 0, & \text{if } \mathbf{x} = \mathbf{0}, \end{cases} \quad j = 1, 2, \dots, m, \quad (\text{A.1})$$

where function $\operatorname{sgn}(x)$, $x \in \mathbb{R}$, is defined by

$$\operatorname{sgn}(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ -1, & \text{if } x < 0. \end{cases}$$

Function $\mathbf{f}(\mathbf{x}, p)$ has some interesting properties.

Proposition 1. For any $\mathbf{x} \in \mathbb{R}^m$ and $p \geq 1$:

(a) $[\mathbf{f}(\mathbf{x}, p)]^T \mathbf{x} = \|\mathbf{x}\|_p$.

(b) $\|\mathbf{f}(\mathbf{x}, p)\|_{d(p)} = 1$, where $d(p)$ is defined in Eq. (4.3), i.e. $\|\bullet\|_{d(p)}$ is the dual norm of $\|\bullet\|_p$.

Proof.

(a) For $\mathbf{x} \neq \mathbf{0}$:

$$\begin{aligned}
[\mathbf{f}(\mathbf{x}, p)]^T \mathbf{x} &= \sum_{j=1}^m \operatorname{sgn}(x_j) \left(\frac{|x_j|}{\|\mathbf{x}\|_p} \right)^{p-1} x_j \\
&= \frac{1}{\|\mathbf{x}\|_p^{p-1}} \sum_{j=1}^m |x_j|^p \quad (\operatorname{sgn}(x_j)x_j = |x_j|) \\
&= \frac{\|\mathbf{x}\|_p^p}{\|\mathbf{x}\|_p^{p-1}} = \|\mathbf{x}\|_p.
\end{aligned}$$

Note that for $\mathbf{x} = \mathbf{0}$, $[\mathbf{f}(\mathbf{x}, p)]^T \mathbf{x} = 0 = \|\mathbf{x}\|_p$. Also, note that if $p = \infty$,

$$\left(\frac{|x_j|}{\|\mathbf{x}\|_p} \right)^{p-1} = \begin{cases} 1, & \text{if } j = \operatorname{argmax}_{j_1} (|x_{j_1}|), \\ 0, & \text{if } j \neq \operatorname{argmax}_{j_1} (|x_{j_1}|), \end{cases}$$

and if $p = 1$,

$$\left(\frac{|x_j|}{\|\mathbf{x}\|_p} \right)^{p-1} = 1.$$

(b) We observe that:

$$\begin{aligned}
\|\mathbf{f}(\mathbf{x}, p)\|_{d(p)} &= \left(\sum_{j=1}^m |[\mathbf{f}(\mathbf{x}, p)]_j|^{d(p)} \right)^{1/d(p)} \\
&= \left(\sum_{j=1}^m \frac{|x_j|^p}{\|\mathbf{x}\|_p^p} \right)^{1/d(p)} = \left(\frac{1}{\|\mathbf{x}\|_p^p} \sum_{j=1}^m |x_j|^p \right)^{1/d(p)} = 1.
\end{aligned}$$

□

A.2 Propositions on matrix norms

The following proposition connects the norm $\|\bullet\|_{d(p),p}$ with the p -Frobenius norm $\|\bullet\|_{p-F}$.

Proposition 2. *For any $n \times m$ matrix \mathbf{A} ,*

$$\|\mathbf{A}\|_{d(p),p} \leq \|\mathbf{A}\|_{p-F},$$

where $d(p)$ is defined in Eq. (4.3).

Proof.

For some $\mathbf{x} \in \mathbb{R}^m$ with $\|\mathbf{x}\|_{d(p)} = 1$,

$$\|\mathbf{A}\|_{d(p),p} = \|\mathbf{A}\mathbf{x}\|_p = \left(\sum_{i=1}^n |\mathbf{A}_i\mathbf{x}|^p \right)^{1/p},$$

where \mathbf{A}_i is row i of matrix \mathbf{A} .

Using Hölder's inequality,

$$|\mathbf{A}_i\mathbf{x}| \leq \|\mathbf{A}_i\|_p \|\mathbf{x}\|_{d(p)} = \|\mathbf{A}_i\|_p.$$

Thus,

$$\|\mathbf{A}\|_{d(p),p} \leq \left(\sum_{i=1}^n \|\mathbf{A}_i\|_p^p \right)^{1/p} = \left(\sum_{i=1}^n \sum_{j=1}^m |A_{i,j}|^p \right)^{1/p} = \|\mathbf{A}\|_{p-F}.$$

□

Norms $\|\bullet\|_{q,p}$ and $\|\bullet\|_{p-F}$ can be expressed through the product of norms of vectors for special kinds of matrices, as the following propositions state.

Proposition 3. For $\mathbf{u}_1 \in \mathbb{R}^n$, $\mathbf{u}_2 \in \mathbb{R}^m$, $p, q \geq 1$,

$$\|\mathbf{u}_1\mathbf{u}_2^T\|_{q,p} = \|\mathbf{u}_1\|_p \|\mathbf{u}_2\|_{d(q)}.$$

Proof.

Since for any $\mathbf{x} \in \mathbb{R}^m$,

$$\|\mathbf{u}_1\mathbf{u}_2^T\mathbf{x}\|_p = \|(\mathbf{u}_2^T\mathbf{x})\mathbf{u}_1\|_p = |\mathbf{u}_2^T\mathbf{x}| \|\mathbf{u}_1\|_p,$$

we have that:

$$\|\mathbf{u}_1\mathbf{u}_2^T\|_{q,p} = \max_{\|\mathbf{x}\|_q=1} \|\mathbf{u}_1\mathbf{u}_2^T\mathbf{x}\|_p = \max_{\|\mathbf{x}\|_q=1} (|\mathbf{u}_2^T\mathbf{x}| \|\mathbf{u}_1\|_p) = \|\mathbf{u}_1\|_p \max_{\|\mathbf{x}\|_q=1} |\mathbf{u}_2^T\mathbf{x}|.$$

Using Hölder's inequality,

$$|\mathbf{u}_2^T\mathbf{x}| = |\mathbf{x}^T\mathbf{u}_2| \leq \|\mathbf{x}\|_q \|\mathbf{u}_2\|_{d(q)},$$

and for any $\mathbf{x} \in \mathbb{R}^m$ with $\|\mathbf{x}\|_q = 1$,

$$|\mathbf{u}_2^T \mathbf{x}| \leq \|\mathbf{u}_2\|_{d(q)}.$$

We have that:

$$|\mathbf{u}_2^T \mathbf{f}(\mathbf{u}_2, d(q))| = |\mathbf{f}(\mathbf{u}_2, d(q))^T \mathbf{u}_2| = \|\mathbf{u}_2\|_{d(q)}.$$

Thus,

$$\max_{\|\mathbf{x}\|_q=1} |\mathbf{u}_2^T \mathbf{x}| = \|\mathbf{u}_2\|_{d(q)},$$

and

$$\|\mathbf{u}_1 \mathbf{u}_2^T\|_{q,p} = \|\mathbf{u}_1\|_p \max_{\|\mathbf{x}\|_q=1} |\mathbf{u}_2^T \mathbf{x}| = \|\mathbf{u}_1\|_p \|\mathbf{u}_2\|_{d(q)}.$$

□

Proposition 4. For $\mathbf{u}_1 \in \mathbb{R}^n$, $\mathbf{u}_2 \in \mathbb{R}^m$, $p \geq 1$,

$$\|\mathbf{u}_1 \mathbf{u}_2^T\|_{p-F} = \|\mathbf{u}_1\|_p \|\mathbf{u}_2\|_p.$$

Proof.

We have that:

$$\begin{aligned} \|\mathbf{u}_1 \mathbf{u}_2^T\|_{p-F} &= \left(\sum_{i=1}^n \sum_{j=1}^m |u_{1,i} u_{2,j}|^p \right)^{1/p} \\ &= \left(\sum_{i=1}^n |u_{1,i}|^p \right)^{1/p} \left(\sum_{j=1}^m |u_{2,j}|^p \right)^{1/p} = \|\mathbf{u}_1\|_p \|\mathbf{u}_2\|_p. \end{aligned}$$

□

Appendix B

Robust Logistic Regression Algorithms

B.1 Robust logistic regression under independent variables uncertainty solution algorithm

At any $(\boldsymbol{\beta}, \beta_0) \in \mathbb{R}^{m+1}$ with $\boldsymbol{\beta} \neq \mathbf{0}$, the subdifferential of Z_1 contains $\nabla Z_1(\boldsymbol{\beta}, \beta_0)$. At any $(\boldsymbol{\beta}, \beta_0) \in \mathbb{R}^{m+1}$ with $\boldsymbol{\beta} = \mathbf{0}$, the subdifferential of $Z_1(\boldsymbol{\beta}, \beta_0)$ contains any vector $(\mathbf{s}, s_0) \in \mathbb{R}^{m+1}$, $\mathbf{s} = (s_1, s_2, \dots, s_m)$, in the box defined by $\left(\frac{\partial Z_1}{\partial \beta_j}\right)^+ \leq s_j \leq \left(\frac{\partial Z_1}{\partial \beta_j}\right)^-$, $j \in \{1, 2, \dots, m\}$ and $s_0 = \frac{\partial Z_1}{\partial \beta_0}$.

Let ϵ_1 be a convergence parameter. The subgradient method used to find the optimal solution of Problem (4.11) is (see Shor [45]):

1. Initiate $(\boldsymbol{\beta}, \beta_0) := (\mathbf{0}, 0)$.
2. If there exists a vector (\mathbf{s}, s_0) in the subdifferential of $Z_1(\boldsymbol{\beta}, \beta_0)$ such that $\|(\mathbf{s}, s_0)\|_\infty \leq \epsilon_1$, then, terminate.
3. Determine a descent direction (\mathbf{s}, s_0) , $\mathbf{s} = (s_1, s_2, \dots, s_m)$, in the following way:

- If $\boldsymbol{\beta} \neq \mathbf{0}$,

$$s_j = \frac{\partial Z_1}{\partial \beta_j}, \quad j = 1, 2, \dots, m.$$

- If $\boldsymbol{\beta} = \mathbf{0}$,

$$s_j = \frac{1}{2} \left(\left(\frac{\partial Z_1}{\partial \beta_j} \right)^- + \left(\frac{\partial Z_1}{\partial \beta_j} \right)^+ \right), \quad j = 1, 2, \dots, m.$$

- $s_0 = \frac{\partial Z_1}{\partial \beta_0}$.

4. Apply the Armijo rule on direction (\mathbf{s}, s_0) to update $(\boldsymbol{\beta}, \beta_0)$ (see Bertsekas [7], p. 29).

5. Go to Step 2.

B.2 Robust logistic regression under response variable uncertainty solution algorithm

The interior point method for solving Problem (4.20) involves the solution of the unconstrained problem

$$\max_{p, \mathbf{q}, \boldsymbol{\beta}, \beta_0} H(p, \mathbf{q}, \boldsymbol{\beta}, \beta_0), \quad (\text{B.1})$$

where

$$\begin{aligned} H(p, \mathbf{q}, \boldsymbol{\beta}, \beta_0) &= Z_2(p, \mathbf{q}, \boldsymbol{\beta}, \beta_0) + \frac{1}{t} \sum_{i=1}^n \ln(-p - q_i + (-1)^{y_i} (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0)) \\ &\quad + \frac{1}{t} \ln(-p) + \frac{1}{t} \sum_{i=1}^n \ln(-q_i), \end{aligned} \quad (\text{B.2})$$

for various values of $t > 0$, through the Newton method.

Let ϵ_2 be a convergence parameter and μ a running parameter. The interior point algorithm used to find the optimal solution of Problem (4.20) is (see Bertsekas [7], Boyd [12]):

1. Initiate $(p, \mathbf{q}, \boldsymbol{\beta}, \beta_0) := (-1, -\mathbf{1}, \mathbf{0}, 0)$, $t := 1$.
2. Calculate

$$(\Delta p, \Delta \mathbf{q}, \Delta \boldsymbol{\beta}, \Delta \beta_0) = -(\nabla^2 H(p, \mathbf{q}, \boldsymbol{\beta}, \beta_0))^{-1} \nabla H(p, \mathbf{q}, \boldsymbol{\beta}, \beta_0).$$

3. If $\|\nabla H(p, \mathbf{q}, \boldsymbol{\beta}, \beta_0)\|_\infty \leq \epsilon_2$, go to Step 6.

4. Apply the Armijo rule on direction $(\Delta p, \Delta \mathbf{q}, \Delta \boldsymbol{\beta}, \Delta \beta_0)$ to update $(p, \mathbf{q}, \boldsymbol{\beta}, \beta_0)$ (see Bertsekas [7], p. 29).
5. Go to Step 2.
6. If $\frac{n+m+2}{t} \leq \epsilon_2$, terminate.
7. $t := \mu \cdot t$.
8. Go to Step 2.

B.3 Globally robust logistic regression solution algorithm

At any $(\boldsymbol{\beta}, \beta_0) \in \mathbb{R}^{m+1}$, the subdifferential of $Z_3(\boldsymbol{\beta}, \beta_0)$ contains any vector $(\mathbf{s}, s_0) \in \mathbb{R}^{m+1}$, $\mathbf{s} = (s_1, s_2, \dots, s_m)$, in the box defined by $\left(\frac{\partial Z_3}{\partial \beta_j}\right)^+ \leq s_j \leq \left(\frac{\partial Z_3}{\partial \beta_j}\right)^-$, $j \in \{1, 2, \dots, m\}$, and $\left(\frac{\partial Z_3}{\partial \beta_0}\right)^+ \leq s_0 \leq \left(\frac{\partial Z_3}{\partial \beta_0}\right)^-$.

Let ϵ_3 be a convergence parameter. The subgradient method used to find the optimal solution of Problem (4.31) is (see Shor [45]):

1. Initiate $(\boldsymbol{\beta}, \beta_0) := (\mathbf{0}, 0)$.
2. If there exists a vector (\mathbf{s}, s_0) in the subdifferential of $Z_3(\boldsymbol{\beta}, \beta_0)$ such that $\|(\mathbf{s}, s_0)\|_\infty \leq \epsilon_3$, then, terminate.
3. Determine a descent direction (\mathbf{s}, s_0) , $\mathbf{s} = (s_1, s_2, \dots, s_m)$, in the following way:
 - $s_j = \frac{1}{2} \left(\left(\frac{\partial Z_3}{\partial \beta_j}\right)^- + \left(\frac{\partial Z_3}{\partial \beta_j}\right)^+ \right)$, $j = 1, 2, \dots, m$.
 - $s_0 = \frac{1}{2} \left(\left(\frac{\partial Z_3}{\partial \beta_0}\right)^- + \left(\frac{\partial Z_3}{\partial \beta_0}\right)^+ \right)$.
4. Apply the Armijo rule on direction (\mathbf{s}, s_0) to update $(\boldsymbol{\beta}, \beta_0)$ (see Bertsekas [7], p. 29).
5. Go to Step 2.

Appendix C

The partial derivatives of $Z_1(\boldsymbol{\beta}, \beta_0)$ and $Z_3(\boldsymbol{\beta}, \beta_0)$

In this section of the Appendix, we describe the way to calculate the partial derivatives of functions $Z_1(\boldsymbol{\beta}, \beta_0)$ and $Z_3(\boldsymbol{\beta}, \beta_0)$, that return the optimal objective values of the inner minimization problems in robust logistic regression.

Define:

$$D_j^\mp(y, \mathbf{x}, \rho, p, \boldsymbol{\beta}, \beta_0) = \left(\frac{\partial[\boldsymbol{\beta}^T \mathbf{x} + \beta_0 + (-1)^y \rho \|\boldsymbol{\beta}\|_{d(p)}]}{\partial \beta_j} \right)^\mp$$

$$= \begin{cases} x_j + (-1)^y \rho f(\boldsymbol{\beta}, p), & \boldsymbol{\beta} \neq \mathbf{0}, j \geq 1, \\ x_j \mp (-1)^y \rho & \boldsymbol{\beta} = \mathbf{0}, j \geq 1, \\ 1, & j = 0, \end{cases}$$

$$j \geq 0, y \in \{0, 1\}, \mathbf{x} \in \mathbb{R}^m, \rho \geq 0, p \geq 1, \boldsymbol{\beta} \in \mathbb{R}^m, \beta_0 \in \mathbb{R}.$$

Using $D_j^\mp(y, \mathbf{x}, \rho, p, \boldsymbol{\beta}, \beta_0)$, we express

$$\left(\frac{\partial Z_1}{\partial \beta_j} \right)^\mp$$

$$= \sum_{i=1}^n D_j^\mp(y_i, \mathbf{x}_i, \rho, p, \boldsymbol{\beta}, \beta_0) \frac{(-1)^{1-y_i} \exp((1-y_i)(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)}))}{1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{y_i} \rho \|\boldsymbol{\beta}\|_{d(p)})},$$

$$j \in \{1, 2, \dots, m\}.$$

To calculate the partial derivatives of $Z_3(\boldsymbol{\beta}, \beta_0)$, we need to introduce some special nota-

tion. Consider vector $\mathbf{a} = (a_1, a_2, \dots, a_n)$.

Definition 6. $\mathcal{O}(\mathbf{a})$ is the set of permutations of $\{1, 2, \dots, n\}$, such that if i_1 precedes i_2 in the permutation, then, $a_{i_1} \leq a_{i_2}$.

We observe that $\mathcal{O}(\mathbf{a})$ contains all the permutations of the members of $\{1, 2, \dots, n\}$, such that as the index i of the permutation increases, the value a_i stays the same or increases.

Definition 7.

$$n_l(\mathbf{a}) \equiv \{i \in \{1, 2, \dots, n\} \mid a_i < 0\}.$$

Definition 8.

$$n_0(\mathbf{a}) \equiv \{i \in \{1, 2, \dots, n\} \mid a_i \leq 0\}.$$

Recall that $S(\mathbf{a})$ is the set of optimal basic feasible solutions to problem

$$\begin{aligned} \min \quad & \sum_{i=1}^n a_i x_i \\ \text{s.t.} \quad & \sum_{i=1}^n x_i \leq \Gamma \\ & 0 \leq x_i \leq 1, \quad i = 1, 2, \dots, n. \end{aligned} \tag{C.1}$$

Lemma 7. $S(\mathbf{a})$ is the set of length n binary sequences \mathbf{v} , such that there exists a permutation $\mathbf{o} \in \mathcal{O}(\mathbf{a})$, and $i_1, \min(n_l(\mathbf{a}), \Gamma) \leq i_1 \leq \min(n_0(\mathbf{a}), \Gamma)$, with $v_{o_i} = 1$, for $i \leq i_1$, and $v_{o_i} = 0$, for $i > i_1$.

Proof. Consider the following algorithm that returns a set of length n binary sequences:

1. $T := \emptyset$.
2. For each $\mathbf{o} \in \mathcal{O}(\mathbf{a})$ and for $i_1 = \min(n_l(\mathbf{a}), \Gamma), \min(n_l(\mathbf{a}), \Gamma) + 1, \dots, \min(n_0(\mathbf{a}), \Gamma)$, construct binary sequence \mathbf{v} , such that $v_{o_i} = 1$, for $i \leq i_1$, and $v_{o_i} = 0$, for $i > i_1$. Add \mathbf{v} to set T .
3. Return T .

It is obvious that T contains the optimal basic feasible solutions of Problem (C.1) and that $T = S(\mathbf{a})$. □

Using the same principles, we now define another set of length n binary sequences that takes into consideration the ordering according to length n vectors \mathbf{a} and $\mathbf{b} = (b_1, b_2, \dots, b_n)$. This is a hierarchical ordering. In the case that a_{i_1} and a_{i_2} are equal, their ordering is determined by the relation between b_{i_1} and b_{i_2} .

Definition 9. $\mathcal{O}(\mathbf{a}, \mathbf{b})$ is the set of permutations of $\{1, 2, \dots, n\}$, such that if i_1 precedes i_2 in the permutation, then, either $a_{i_1} < a_{i_2}$ holds, or $a_{i_1} = a_{i_2}$ and $b_{i_1} \leq b_{i_2}$ both hold.

Definition 10.

$$n_l(\mathbf{a}, \mathbf{b}) \equiv \{i \in \{1, 2, \dots, n\} \mid a_i < 0 \text{ or } (a_i = 0 \text{ and } b_i < 0)\}.$$

Definition 11.

$$n_0(\mathbf{a}, \mathbf{b}) \equiv \{i \in \{1, 2, \dots, n\} \mid a_i < 0 \text{ or } (a_i = 0 \text{ and } b_i \leq 0)\}.$$

Definition 12. $S(\mathbf{a}, \mathbf{b})$ is the set of length n binary sequences \mathbf{v} , such that there exists a permutation $\mathbf{o} \in \mathcal{O}(\mathbf{a}, \mathbf{b})$, and $i_1, \min(n_l(\mathbf{a}, \mathbf{b}), \Gamma) \leq i_1 \leq \min(n_0(\mathbf{a}, \mathbf{b}), \Gamma)$, with $v_{o_i} = 1$, for $i \leq i_1$, and $v_{o_i} = 0$, for $i > i_1$.

Let

$$n_S(\mathbf{a}, \mathbf{b}) = |S(\mathbf{a}, \mathbf{b})|,$$

and $\mathbf{s}^q(\mathbf{a}, \mathbf{b})$, $q = 1, 2, \dots, n_S(\mathbf{a}, \mathbf{b})$, be the binary sequence which is placed at position q if we order the binary sequences of set $S(\mathbf{a}, \mathbf{b})$ according to relation “ $<_{\text{lex}}$ ”, i.e.,

$$\mathbf{s}^1(\mathbf{a}, \mathbf{b}) <_{\text{lex}} \mathbf{s}^2(\mathbf{a}, \mathbf{b}) <_{\text{lex}} \dots <_{\text{lex}} \mathbf{s}^{n_S(\mathbf{a}, \mathbf{b})}(\mathbf{a}, \mathbf{b}),$$

where relation “ $<_{\text{lex}}$ ” is defined in Definition 4.

We also define functions $W^{j,\mp}(\boldsymbol{\beta}, \beta_0)$, $n_W^{j,\mp}(\boldsymbol{\beta}, \beta_0)$, $\mathbf{w}^{q,j,\mp}(\boldsymbol{\beta}, \beta_0)$, $q = 1, 2, \dots, n_W^{j,\mp}(\boldsymbol{\beta}, \beta_0)$, $(\boldsymbol{\beta}, \beta_0) \in \mathbb{R}^{m+1}$, as

$$\begin{aligned} W^{j,\mp}(\boldsymbol{\beta}, \beta_0) &= S(\mathbf{a}, \mathbf{b}), \\ n_W^{j,\mp}(\boldsymbol{\beta}, \beta_0) &= n_S(\mathbf{a}, \mathbf{b}), \\ \mathbf{w}^{q,j,\mp}(\boldsymbol{\beta}, \beta_0) &= \mathbf{s}^q(\mathbf{a}, \mathbf{b}), \quad q = 1, 2, \dots, n_W^{j,\mp}(\mathbf{a}, \mathbf{b}), \end{aligned}$$

where $\mathbf{a} = (a_1, a_2, \dots, a_n)$, $\mathbf{b} = (b_1, b_2, \dots, b_n)$,

$$a_i = (-1)^{y_i}(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0), \quad i = 1, 2, \dots, n,$$

and

$$b_i = \mp(-1)^{y_i} \begin{cases} x_{i,j}, & \text{if } j \geq 1, \\ 1, & \text{if } j = 0, \end{cases}, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, m.$$

It is obvious that

$$W^{j,\mp}(\boldsymbol{\beta}, \beta_0) \subseteq U(\boldsymbol{\beta}, \beta_0), \quad j = 1, 2, \dots, m,$$

because $W^{j,\mp}(\boldsymbol{\beta}, \beta_0)$ is produced using a “stricter” ordering than the one used to produce $U(\boldsymbol{\beta}, \beta_0)$.

Definition 13. $\mathbf{d}^{j,-} \in \mathbb{R}^{m+1}$ is the vector having -1 at the coordinate corresponding to β_j and 0 at all other coordinates, $j = 1, 2, \dots, m$. $\mathbf{d}^{j,+} \in \mathbb{R}^{m+1}$ is the vector having 1 at the coordinate corresponding to β_j and 0 at all other coordinates, $j = 1, 2, \dots, m$. $\mathbf{d}_{\boldsymbol{\beta}}^{j,\mp}$ is the component of $\mathbf{d}^{j,-}$ corresponding to $\boldsymbol{\beta}$ and $d_{\beta_0}^{j,\mp}$ is the component of $\mathbf{d}^{j,\mp}$ corresponding to β_0 .

The following theorem is used to calculate the partial derivatives of $Z_3(\boldsymbol{\beta}, \beta_0)$ with respect to any member of $\boldsymbol{\beta}$ and β_0 .

Theorem 8. For any $(\boldsymbol{\beta}, \beta_0) \in \mathbb{R}^{m+1}$ and any direction $\mathbf{d}^{j,\mp}$, there exists an $\epsilon > 0$ such that for any $0 \leq t \leq \epsilon$,

$$\begin{aligned} Z_3((\boldsymbol{\beta}, \beta_0) + t\mathbf{d}^{j,\mp}) &= \sum_{i=1}^n \left[|y_i - w_i^{1,j,\mp}(\boldsymbol{\beta}, \beta_0)| (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - w_i^{1,j,\mp}(\boldsymbol{\beta}, \beta_0)|} \rho \|\boldsymbol{\beta}\|_{d(p)}) \right. \\ &\quad \left. - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - w_i^{1,j,\mp}(\boldsymbol{\beta}, \beta_0)|} \rho \|\boldsymbol{\beta}\|_{d(p)})) \right]. \end{aligned}$$

Proof. Let

$$a_i(t) = (-1)^{y_i} [(\boldsymbol{\beta} + t\mathbf{d}_{\boldsymbol{\beta}}^{j,\mp})^T \mathbf{x}_i + (\beta_0 + td_{\beta_0}^{j,\mp})],$$

and

$$b_i = \mp(-1)^{y_i} \begin{cases} x_{i,j}, & \text{if } j \geq 1, \\ 1, & \text{if } j = 0, \end{cases}, \quad i = 1, 2, \dots, n.$$

Since

$$W^{j,\mp}(\boldsymbol{\beta}, \beta_0) \subseteq U(\boldsymbol{\beta}, \beta_0),$$

we conclude that

$$Z_3(\boldsymbol{\beta}, \beta_0) = \sum_{i=1}^n \left[|y_i - w_i^{1,j,\mp}(\boldsymbol{\beta}, \beta_0)| (\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - w_i^{1,j,\mp}(\boldsymbol{\beta}, \beta_0)|} \rho \|\boldsymbol{\beta}\|_{d(p)}) \right. \\ \left. - \ln(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - w_i^{1,j,\mp}(\boldsymbol{\beta}, \beta_0)|} \rho \|\boldsymbol{\beta}\|_{d(p)})) \right].$$

It is obvious that there exists an $\epsilon > 0$ that preserves the ordering of $i \in \{1, 2, \dots, n\}$, according to a_i , i.e. if $a_{i_1}(0) < a_{i_2}(0)$, then, $a_{i_1}(t) < a_{i_2}(t)$. We observe that:

$$b_i = \frac{\partial a_i(t)}{\partial t}, \quad i = 1, 2, \dots, n.$$

Thus, if $a_{i_1}(0) = a_{i_2}(0)$, the ordering of $a_{i_1}(t)$ and $a_{i_2}(t)$ is the same as the ordering of b_{i_1} and b_{i_2} , for $0 < t \leq \epsilon$. This implies, that

$$W^{j,\mp}(\boldsymbol{\beta}, \beta_0) \subseteq U((\boldsymbol{\beta}, \beta_0) + t\mathbf{d}^{j,\mp}),$$

which completes the proof. □

The following is a direct corollary from Theorem 8.

Corollary 9. *The partial derivatives of $Z_3(\boldsymbol{\beta}, \beta_0)$ are:*

$$\left(\frac{\partial Z_3}{\partial \beta_j} \right)^\mp = \sum_{i=1}^n D_j^\mp (|y_i - w_i^{1,j,\mp}(\boldsymbol{\beta}, \beta_0)|, \mathbf{x}_i, \rho, p, \boldsymbol{\beta}, \beta_0) \\ (-1)^{1 - |y_i - w_i^{1,j,\mp}(\boldsymbol{\beta}, \beta_0)|} \\ \frac{\exp((1 - |y_i - w_i^{1,j,\mp}(\boldsymbol{\beta}, \beta_0)|)(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - w_i^{1,j,\mp}(\boldsymbol{\beta}, \beta_0)|} \rho \|\boldsymbol{\beta}\|_{d(p)}))}{1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 + (-1)^{|y_i - w_i^{1,j,\mp}(\boldsymbol{\beta}, \beta_0)|} \rho \|\boldsymbol{\beta}\|_{d(p)})}, \\ j \in \{1, 2, \dots, m\}.$$

Bibliography

- [1] J. Aldrich, R. A. Fisher and the Making of Maximum Likelihood 1912-1922, *Statistical Science*, Volume 12, Number 3, pp. 162-176, 1997.
- [2] P. D. Allison, *Logistic Regression Using the SAS System: theory and application*, SAS Institute, Cary, NC, 1999.
- [3] A. Asuncion, D.J. Newman, UCI Machine Learning Repository, <http://www.ics.uci.edu/~mllearn/MLRepository.html>, University of California, School of Information and Computer Science, Irvine, CA, 2007.
- [4] A. Ben-Tal, A. Nemirovski, Robust convex optimization, *Mathematics of Operations Research*, Volume 23, Issue 4, pp. 769-805, 1998.
- [5] A. Ben-Tal, A. Nemirovski, Robust solutions of Linear Programming problems contaminated with uncertain data, *Mathematical Programming*, Volume 88, Number 3, pp. 411-424, 2000.
- [6] A. Ben-Tal, A. Nemirovski, Robust solutions to uncertain programs, *Operations Research Letters*, Volume 25, Issue 1, pp. 1-13, 1999.
- [7] D. P. Bertsekas, *Nonlinear Programming*, Second Edition, Athena Scientific, Belmont, Massachusetts, 1995.
- [8] D. Bertsimas, M. Sim, Robust Discrete Optimization and Network Flows, *Mathematical Programming*, Volume 98, Numbers 1-3, pp. 49-71, 2003.
- [9] D. Bertsimas, M. Sim, The Price of Robustness, *Operations Research*, Volume 52, Issue 1, pp. 35-53, 2004.
- [10] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pp. 144-152, New York, NY, 1992.

- [11] S. Boyd, C. Barratt, Linear Controller Design: Limits of Performance, Prentice-Hall, 1991.
- [12] S. Boyd, L. Vandenberghe, Convex optimization, Cambridge University Press, 2004.
- [13] E. J. Candes, Y. Plan, Near-ideal model selection by \mathcal{L}_1 minimization, Technical Report, California Institute of Technology, 2007.
- [14] D. R. Cox, Analysis of Binary Data, Chapman and Hall, London, 1969.
- [15] J. M. Danskin, The theory of Max-Min, With Applications, SIAM Journal of Applied Mathematics, Volume 14, Number 4, July 1966, USA.
- [16] A. S. Deif, Advanced Matrix Theory for Scientists and Engineers, Gordon and Breach Science Publishers, 1991.
- [17] R. A. Fisher, Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population, Biometrika, Volume 10, pp. 507-521, 1915.
- [18] R. A. Fisher, On an absolute criterion for fitting frequency curves, Messenger of Mathematics, Volume 41, pp. 155-160, 1912.
- [19] M. Foster, An application of the Wiener-Kolmogorov smoothing theory to matrix inversion, Journal of SIAM, Volume 9, pp. 387-392, 1961.
- [20] L. E. Ghaoui, H. Le Bret, Robust Solutions to Least-Squares Problems with Uncertain Data, SIAM Journal on Matrix Analysis and Applications, Volume 18, Issue 4, pp. 1035-1064, 1997.
- [21] G. Giorgi, A. Guerraggio, J. Thierfelder, Mathematics of Optimization: Smooth and Nonsmooth Case, Elsevier, 2004.
- [22] A. S. Goldberger, Econometric Theory, J. Wiley & Sons, Madison, Wisconsin, 1963.
- [23] G. H. Golub, P. C. Hansen, D. P. O'Leary, Tikhonov Regularization and Total Least Squares, SIAM J. Matrix Anal. Appl., Volume 21, pp. 185-194.
- [24] G. H. Golub, C. F. Van Loan, Matrix Computations, The Johns Hopkins University Press, Baltimore and London, 1996.
- [25] W. H. Greene, Econometric Analysis, 5th edition, Prentice Hall, 2003.

- [26] F. R. Hampel, The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, Volume 62, pp. 1179-1186.
- [27] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer, New York, 2001.
- [28] D. W. Hosmer, S. Lemeshow, *Applied Logistic Regression*, Wiley Series in Probability and Statistics, New York, 2000.
- [29] R. J. Huber, *Robust Estimation of a Location Parameter*, University of California, Berkeley, 1963.
- [30] P. J. Huber, *Robust Statistics*, John Wiley & Sons, Cambridge, Massachusetts, 1980.
- [31] R. A. Johnson, D. W. Wichern, *Applied Multivariate Statistical Analysis*, Pearson Prentice Hall, Upper Saddle River, New Jersey, 2007.
- [32] E. L. Lehmann, G. Casella, *Theory of Point Estimation*, Springer, 2nd edition, 1998.
- [33] A. J. McCoy, New Applications of Maximum Likelihood and Bayesian Statistics in Macromolecular Crystallography, *Current Opinion in Structural Biology*, Volume 12, Issue 5, pp. 670-673, October 2002.
- [34] G. R. G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, M. I. Jordan, A Robust Minimax Approach to Classification, *Journal of Machine Learning Research*, Volume 3, 2002, pp. 555-582.
- [35] N. R. Patel, Logistic Regression, <http://ocw.mit.edu/NR/rdonlyres/Sloan-School-of-Management/15-062Data-MiningSpring2003/B2EC3803-F8A7-46CF-8B9E-D0D080E52A6B/0/logreg.pdf>.
- [36] I. Polik, Addendum to the SeDuMi User Guide, version 1.1.
- [37] I. Popescu, Robust Mean-Covariance Solutions for Stochastic Optimization, *Operations Research*, Volume 55, Number 1, pp. 98-112, January-February 2007.
- [38] F. Rendl, H. Wolkowicz, A Semidefinite Framework for Trust Region Subproblems with Applications to Large Scale Optimization, *CORR Report 94-32*, 1996.
- [39] W. W. Rogosinsky, Moments of Non-Negative Mass, *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, Volume 245, Number 1240, pp. 1-27, 1958.

- [40] T. P. Ryan, *Modern Regression Methods*, Wiley Series in Probability and Statistics, New York, 1997.
- [41] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2nd edition, 2003.
- [42] SAS Real data sets, <http://ftp.sas.com/samples/A55770>.
- [43] B. Schölkopf, A. J. Smola, *Learning with Kernels*, The MIT Press, Cambridge, Massachusetts, 2002.
- [44] P. K. Shivaswamy, C. Bhattacharyya, A. J. Smola, Second Order Cone Programming Approaches for Handling Missing and Uncertain Data, *Journal of Machine Learning Research*, Volume 7, 2006, pp. 1283-1314.
- [45] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, Translated from the Russian by K. C. Kiwiel and A. Ruszczyński, Springer Series in Computational Mathematics, Berlin, 1985.
- [46] J. E. Smith, Generalized Chebychev Inequalities: Theory and Applications in Decision Analysis, *Operations Research*, Volume 43, Number 5, September-October 1995.
- [47] A. L. Soyster, Convex Programming with Set-Inclusive Constraints and Applications to Inexact Linear Programming, *Operations Research*, Volume 21, Issue 5, pp. 1154-1157, 1973.
- [48] J. F. Sturm, Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, Special Issue on Interior Point Methods (CD supplement with software), 1999.
- [49] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B (Methodological)*, Volume 57, Number 1, pp. 267-288, 1995.
- [50] A. N. Tikhonov, V. Y. Arsenin, *Solutions of Ill-Posed Problems*, V. H. Winston & Sons, 1977.
- [51] R. L. Wu, C. X. Ma, M. Lin, G. Casella, A general framework for analyzing the genetic architecture of developmental characteristics, *Genetics*, Volume 166, pp. 1541-1551, 2004.

- [52] H. Xu, C. Caramanis, S. Mannor, Robustness, Risk, and Regularization in Support Vector Machines, *Journal of Machine Learning Research*, 2009.