Robust penalized M-estimators for generalized linear and additive models

AVELLA MEDINA, Marco Andrés

Abstract

Generalized linear models (GLM) and generalized additive models (GAM) are popular statistical methods for modelling continuous and discrete data both parametrically and nonparametrically. In this general framework, we consider the problem of variable selection by studying a wide class of penalized M-estimators that are particularly well suited for high dimensional scenarios where the number of covariates $p$ is very large relative to the sample size $n$. We focus on resistance issues in the presence of deviations from the stochastic assumptions of the postulated models and highlight the weaknesses of widely used estimators. We advocate the need for robust estimators and propose several penalized quasilikelihood estimators that achieve both good statistical properties at the assumed model and stability in a neighborhood of it. Specifically, we provide careful asymptotic analyses of our robust estimators for GLM and GAM when the number of parameters increases with the sample size. We start by revisiting the asymptotics of M-estimators for GLM with a diverging number of parameters. We establish asymptotic normality of these […]

Reference


URN : urn:nbn:ch:unige-830605
DOI : 10.13097/archive-ouverte/unige:83060

Available at:
http://archive-ouverte.unige.ch/unige:83060

Disclaimer: layout of this document may differ from the published version.
Robust penalized M-estimators for generalized linear and additive models

by Marco Avella Medina

A thesis submitted to the
Geneva School of Economics and Management,
University of Geneva, Switzerland,
in fulfillment of the requirements for the degree of
PhD in Statistics

Members of the thesis committee:
Prof. Eva CANTONI, Chair, University of Geneva
Prof. Elvezio RONCHETTI, Adviser, University of Geneva
Prof. Sylvain SARDY, University of Geneva
Prof. Roy WELSCH, Massachusetts Institute of Technology

Thesis No. 20 March 2016
Acknowledgements

Many people supported me both academically and personally during my time as a PhD student in Geneva. I would first of all like to thank my supervisor Elvezio Ronchetti for his valuable guidance, advice and availability. In addition to introducing me to academic research, his lectures increased my interest in statistics in my early days and his approach to research and teaching will certainly have a positive influence on my future career. I would also like to thank Eva Cantoni, Sylvain Sardy and Roy Welsch for accepting being part of the thesis committee, for their careful reading of this dissertation and helpful discussions.

I am very grateful to my professors, colleagues and friends at the Geneva School of Economics and Management and in particular to the members of the Research Center for Statistics. I will long remember the hard work and happy moments we have shared during seminars, lectures, informal discussions, lunches and coffee breaks.

I want to thank my great friends Daniel, Kevin and Mauro for making these years unforgettable in many ways. A special thought goes to Justine for all her support and the great moments we have spent together.

Finally, I would like to express my gratitude to my mother and to all my family for always being there for me throughout the years.
Abstract

Generalized linear models (GLM) and generalized additive models (GAM) are popular statistical methods for modelling continuous and discrete data both parametrically and nonparametrically. In this general framework, we consider the problem of variable selection by studying a wide class of penalized M-estimators that are particularly well suited for high dimensional scenarios where the number of covariates $p$ is very large relative to the sample size $n$. We focus on resistance issues in the presence of deviations from the stochastic assumptions of the postulated models and highlight the weaknesses of widely used estimators. We advocate the need for robust estimators and propose several penalized quasilikelihood estimators that achieve both good statistical properties at the assumed model and stability in a neighborhood of it. Specifically, we provide careful asymptotic analyses of our robust estimators for GLM and GAM when the number of parameters increases with the sample size. We start by revisiting the asymptotics of M-estimators for GLM with a diverging number of parameters. We establish asymptotic normality of these estimators and reexamine distributional results for likelihood ratio type and Wald type tests based on them. We then consider penalized M-estimators for high dimensional set ups where $p \gg n$. In the GLM setting we show that our estimators are consistent, asymptotically normally distributed and variable selection consistent under regularity conditions. Furthermore they have a bounded bias in a neighborhood of the model. In the GAM setting we establish an $\ell_2$-norm consistency result for the nonparametric components which achieves the optimal rates of convergence. In addition, the proposed penalized estimator is able to select the correct model consistently. We propose new algorithms for the implementation of our penalized M-estimators and illustrate the finite sample performance of our methods, at the model and under contamination, in simulation studies. An important contribution of this thesis is to formally study the local robustness properties of general nondifferentiable penalized M-estimators. In particular, we propose a framework that allows us to define rigorously the influence function as the limiting influence function of a sequence of approximating functionals. We show that this influence function can be used to characterize the robustness properties of a wide range of sparse estimators and that it can be viewed as a derivative in the sense of distribution theory. At the end of this thesis, we discuss some extensions of our work and give an overview of the future challenges of robust statistics in high dimensions.
Les modèles linéaires généralisés (GLM) et les modèles additifs généralisés (GAM) sont des méthodes statistiques populaires utilisées pour modéliser paramétriquement et non-paramétriquement des données continues et discrètes. Dans ce cadre général, nous considérons le problème de la sélection de variables en étudiant une large classe de M-estimateurs pénalisés. Cette classe est particulièrement intéressante pour des problèmes de grande dimension où le nombre de covariables $p$ est très grand par rapport au nombre d’observations $n$. Nous nous focalisons sur les propriétés de résistance de ces méthodes lorsque les conditions stochastiques des modèles postulés sont violées et nous soulignons les faiblesses de certaines classes d’estimateurs couramment utilisées en pratique. Nous montrons le besoin de développer des estimateurs robustes et, par le biais d’une quasi-vraisemblance pénalisée, nous construisons des estimateurs qui ont simultanément de bonnes propriétés statistiques au modèle adopté ainsi que dans un voisinage de celui-ci. Plus spécifiquement, nous développons une analyse asymptotique détaillée pour nos estimateurs robustes de GLM et GAM dans un cadre où le nombre de paramètres augmente avec la taille de l’échantillon. Nous commençons par revisiter les résultats asymptotiques concernant les M-estimateurs de GLM avec un nombre de paramètres divergeant. Nous montrons la normalité asymptotique de ces estimateurs et réexaminons des résultats concernant la distribution asymptotique de tests de type ratios de log-vraisemblance et de type Wald. Nous considérons ensuite des M-estimateurs pénalisés pour le scénario de grande dimension où $p \gg n$. Dans le cadre GLM, nous montrons que sous des conditions de régularité nos estimateurs sont constants, asymptotiquement normalement distribués et constants pour la sélection de variables. De surcroît, nous montrons que leur biais est borné dans un voisinage du modèle. Dans le cadre GAM, nous montrons la consistance avec vitesse de convergence optimale en norme $\ell_2$ des composantes nonparamétriques. Nous montrons également que notre estimateur pénalisé est consistant pour la sélection du vrai model. Nous proposons de nouveaux algorithmes pour l’implémentation de nos méthodes et nous illustrons à l’aide de simulations leur performance en échantillon fini, au modèle et dans un voisinage de celui-ci. Une contribution importante de cette thèse est d’étudier formellement les propriétés de robustesse des M-estimateurs pénalisés non-différentiables. En particulier, nous proposons un cadre théorique qui nous permet de définir rigoureusement la fonction d’influence comme la limite de la fonction d’influence d’une suite de fonctionnelles approximant la M-fonctionnelle pénalisée que nous voulons étudier. Nous montrons que cette fonction d’influence peut être utilisée pour caractériser les propriétés de robustesse d’un large spectre d’estimateurs parcimonieux et qu’elle peut être vue comme une dérivée dans le sens de la théorie des distributions. Nous concluons cette thèse avec une discussion portant sur des extensions de notre travail et en donnant un aperçu sur les futures défis de la statistique robuste dans le scénario des grandes dimensions.
Contents

Acknowledgements .......................................................... i
Abstract ................................................................................ iii
Résumé .................................................................................... v

1 Introduction ........................................................................... 1
  1.1 Motivation ........................................................................... 1
  1.2 Background ........................................................................ 2
  1.3 Our contribution ................................................................. 3
  1.4 Organization of this thesis .................................................... 4

2 Robust GLM with diverging number of parameters ................. 5
  2.1 Introduction ........................................................................ 5
  2.2 Robust estimation in GLM ................................................... 6
      2.2.1 Background ............................................................... 6
      2.2.2 The robust quasilikelihood ........................................ 7
      2.2.3 Bounded deviance estimation ...................................... 7
      2.2.4 Robust estimation based on Bregman divergence ......... 8
      2.2.5 Alternative approaches to robust GLM ....................... 8
  2.3 Asymptotic Analysis .......................................................... 8
      2.3.1 Regularity conditions ................................................ 8
      2.3.2 Asymptotic normality ................................................. 9
      2.3.3 Estimation of the covariance matrix ........................... 10
      2.3.4 Inference ................................................................ 10
      2.3.5 Discussion .............................................................. 11
  2.4 Concluding remarks ........................................................ 11
  2.5 Appendix ............................................................................ 12
      2.5.1 Proof of main results ............................................... 12
      2.5.2 Huberized residuals ................................................ 18

3 Robust penalized M-estimators for GLM ............................... 21
  3.1 Introduction ........................................................................ 21
  3.2 Penalized M-estimators ....................................................... 23
      3.2.1 Background ............................................................. 23
      3.2.2 The oracle estimator and its robustness properties ....... 23
      3.2.3 Some fixed parameter asymptotics ............................. 24
  3.3 Robust generalized linear models ....................................... 24
      3.3.1 A robust M-estimator ............................................... 24
3.3.2 Asymptotic analysis in high dimensions .................................. 25
3.3.3 Weak oracle properties in a contamination neighborhood ............ 27

3.4 Implementation ........................................................................ 28
3.4.1 Fisher scoring coordinate descent .......................................... 28
3.4.2 Tuning parameter selection ................................................. 28

3.5 Numerical examples .................................................................. 29
3.5.1 A simulation study: large outliers in the responses .................... 29
3.5.2 An illustration ...................................................................... 32

3.6 Appendix ................................................................................ 33
3.6.1 Regularity conditions ............................................................ 33
3.6.2 Proof of main results ............................................................ 35
3.6.3 Auxiliary Lemmas ............................................................... 40

4 The influence function of penalized M-estimators ............................. 43
4.1 Introduction ............................................................................. 43

4.2 Penalized M-estimators and Influence Functions ............................ 45
4.2.1 Background ........................................................................ 45
4.2.2 Limiting Influence Function ............................................... 46

4.3 Two-stage penalized M-estimators .............................................. 48

4.4 Numerical illustrations .............................................................. 50
4.4.1 Orthogonal design ................................................................ 50
4.4.2 High dimensional Poisson regression .................................... 52

4.5 Discussion ............................................................................... 54
4.6 Proofs ..................................................................................... 57

4.8 Appendix ................................................................................ 59
4.8.1 Auxiliary results for Berge’s maximum theorem ....................... 59
4.8.2 Auxiliary results from distribution theory ................................. 60
4.8.3 Additional Proofs ................................................................. 61
4.8.4 Example of explicit computations with distributional derivatives .. 62

5 Robust penalized M-estimators for GAM ........................................ 65
5.1 Introduction ............................................................................. 65

5.2 Notation and background .......................................................... 67
5.2.1 A class of robust M-estimators for generalized linear models .... 68
5.2.2 Robust approaches to generalized additive models .................. 69

5.3 Variable selection for robust generalized additive models ............... 69
5.3.1 Robust group and adaptive group lasso for GAM .................... 70
5.3.2 Simultaneous scale estimation .............................................. 70

5.4 Asymptotic properties .............................................................. 71
5.4.1 Conditions .......................................................................... 71
5.4.2 Consistency and oracle properties ........................................ 72

5.5 Robustness properties ............................................................. 74
5.5.1 Bounded influence approach .............................................. 74
5.5.2 Regularity conditions ......................................................... 74
5.5.3 Limiting influence function ................................................. 75
5.5.4 Other approaches to the influence function ......................... 76

5.6 Implementation ....................................................................... 76
5.6.1 A pathwise algorithm ......................................................... 76
5.6.2 An iterative block thresholding algorithm for the group lasso . . . 77
5.6.3 Tuning parameter selection ................................. 78
5.7 Simulations ................................................... 78
5.8 Appendix ..................................................... 81
5.8.1 Proof of main results ....................................... 81
5.8.2 Auxiliary lemmas for consistency results ................. 85
5.8.3 Auxiliary results for Berge’s maximum theorem ........ 87

6 Final remarks and discussion 89
6.1 Extensions ..................................................... 89
6.1.1 Robust sure independence screening ...................... 89
6.1.2 Choice of the weight function ........................... 90
6.1.3 Tuning parameter selection ................................ 91
6.1.4 Further topics .............................................. 91
6.2 Some recent developments ................................... 92
6.2.1 Exponential concentration without light tails ........... 92
6.2.2 Connections between M-estimators and high dimensional statistics . 93
6.2.3 New asymptotics in high dimensions .................... 94
6.3 New challenges for robust statistics ......................... 95
6.3.1 Propagation of outliers .................................... 95
6.3.2 Characterizing robustness and regularization .......... 96
6.3.3 Revisiting old issues ...................................... 97

References 99
To the memory of my grandparents.
Chapter 1

Introduction

1.1 Motivation

Undoubtedly there is an increasing awareness of the need of statistics as large amounts of data are collected nearly in every area of industry and science. “Big data” and its widely associated need of “data scientists” are intrinsically linked to this phenomenon. From a statistician’s viewpoint, there are many new challenges and paradigms generated by the increasing availability of large data sets. At the heart of this dissertation we account for one of the main newly established paradigms in the statistics literature: the dimensionality of the statistical models can diverge possibly at an even faster rate than the number of observations.

Pioneering work establishing theory for estimators with increasing dimension appeared already in the seventies and eighties following the seminal paper of Huber [1973]. This work was partially motivated by applications where such theoretical considerations seemed necessary, specifically in X-ray crystallography, since there the typical number of parameters to be estimated ranges between 10 and 500 while the observations between 100 and 10000. In the kinds of data sets encountered nowadays both the dimensionality of the parameters and the available observations can be much larger. In the models that we study, the data analyst observes a large number of variables and therefore constructs very complex models. In particular, the number of variables at hand can be even larger than the number of observations as long as there is a relatively simple (unknown) underlying structure. This is motivated by numerous applications where the number of variables collected is very large but it is sensible to assume the existence of a simple unknown model. The hope is that we can unveil it through our data analysis. This assumption is usually referred to as sparsity and it roughly means that only a small number of variables plays an important role. One major example where we hope that such sparsity holds, is the way genes are involved in the process leading to the development of cancer. Indeed we expect that only a few of the over 30'000 genes have a direct impact in the appearance of cancer.

Statistical models are used in practice as approximations to reality. They are built under certain assumptions and, from a purely theoretical point of view, they are justified by the statistical properties that they enjoy. It is natural to wonder how a procedure behaves if the assumptions upon which the model is constructed fail to hold. This question becomes even more important with the large sizes of modern data sets since working with large models inevitably implies making more assumptions. Therefore robustness towards stochastic deviations from the assumed model is another key concept that is explored.
throughout this work.

1.2 Background

Approaches to variable selection

Variable selection is an important issue in statistics. In many statistical analyses the main interest is to find important covariates for explaining the response variable of interest. By reducing the complexity of a model and by accepting a small amount of bias, one can both improve its interpretability and its prediction accuracy and obtain a compromise between goodness of fit and parsimony. Three main approaches can achieve this goal. Subset selection chooses a model containing a subset of available variables according to some optimality criterion. In order to do so one has to explore all possible model subsets which, due to its combinatorial nature, becomes unfeasible when the number of covariates is moderate to large. For a good overview of these techniques and an extensive theoretical treatment of their properties, see Claeskens and Hjort [2008]. Stepwise procedures seek to overcome this problem by exploring the space of all possible models along some paths, thus reducing the number of models considered. However, the dependence on the chosen path, which is typically based on statistical tests, can lead to inconsistent results. Some recent theory on model selection consistency for stepwise procedures can be found in Zhang [2009] and Zhang [2011]. Note that both subset and stepwise selection are discrete in nature and therefore can be very variable and unstable as pointed out by Breiman [1995]. Note also that a sort of compromise between these two approaches is to stochastically explore the covariate space by giving promising models a higher probability of being selected. Representative work in this direction includes George and McCulloch [1993] and George and McCulloch [1997].

Finally, penalized estimators have gained popularity over the last years and have proved to be a good alternative to the traditional approaches for variable selection mentioned before. By providing estimation and variable selection simultaneously, they overcome the increasingly high computational cost of variable selection when the number of covariates is large and they reduce its variability. Since their introduction in the linear model (Breiman [1995], Tibshirani [1996]), many extensions of lasso-type procedures have been proposed; cf. for instance Tibshirani [1997], Efron et al. [2004], Zou and Hastie [2005], Yuan and Lin [2006], and the retrospective article by Tibshirani [2011]. Their asymptotic properties have been studied for the case when the number of parameters is fixed by Knight and Fu [2000], Fan and Li [2001], Zou [2006]. A large body of literature has developed for the high dimensional case, where the number of parameters is allowed to grow as the sample size increases at an even faster rate; see an overview in Bühlmann and Van De Geer [2011], Hastie et al. [2015] and references thereof. All these results have provided strong theoretical arguments in favor of penalized estimators. This third category of variable selection methods is the main topic of this thesis.

Robustness aspects

The aforementioned approaches to variable selection were initially developed under relatively restrictive stochastic assumptions which may not be satisfied on real data. Moreover, they are typically affected by the presence of a few outlying observations. Robust statistics (Huber [1981] and 2nd edition by Huber and Ronchetti [2009], Hampel et al.
1.3. Our contribution

[1986], Maronna et al. [2006]) provides a theoretical framework that allows us to take into account that models are only idealized approximations of reality and develops methods that still give reliable results when slight deviations from the stochastic assumptions on the model occur. Following this approach, robust counterparts of subset and stepwise selection procedures have been proposed for linear models; for a review, see Ronchetti [1997]. For generalized linear models (GLM; McCullagh and Nelder [1989]) a robust stepwise selection can be performed following Cantoni and Ronchetti [2001] and subset selection based on a robust $C_P$ could be implemented adapting the work of Cantoni et al. [2005] for longitudinal GLM. More recently, some robustified shrinkage methods have been developed in the wavelet framework (Sardy et al. [2001]) and in linear models (Morgensthaler et al. [2003], Khan et al. [2007], Wang et al. [2007a], Chen et al. [2010], Li et al. [2011], Lambert-Lacroix and Zwald [2011], Alfons et al. [2013]).

1.3 Our contribution

We approach the problem of robust variable selection through penalized methods in the general framework of GLM and the more flexible generalized additive models (GAM; Hastie and Tibshirani [1990], Wood [2006]). We focus on robustness issues in the presence of deviations from the stochastic assumptions of the postulated models and highlight the weaknesses of widely used estimators. We advocate the need for robust estimators and propose several penalized quasilikelihood estimators that achieve both good statistical properties at the assumed model and stability in a neighborhood of it.

We first revisit the asymptotics of a class of M-estimators for GLM with a diverging number of parameters. We establish asymptotic normality of these estimators, prove consistency of the estimated covariance matrix and reexamine distributional results for likelihood ratio type and Wald type tests based on them. All the results are shown to hold when the number of covariates $p$ diverges with the sample size $n$ in the order $o(\sqrt{n})$ or $o(n/\log n)$, depending on the assumptions on the design matrix. We greatly improve on existing results in the literature. We then study the problem of robust estimation and variable selection in GLM and GAM in high dimensional settings where $p \gg n$. We propose a class of robust penalized estimators for high dimensional GLM. We show that our estimators are consistent for variable selection, asymptotically normally distributed and can behave as well as a robust oracle estimator under regularity conditions. We extend our construction to high dimensional GAM and establish an asymptotic $\ell_2$-norm consistency result for the nonparametric components which achieves the optimal rates of convergence. In addition, the proposed penalized estimator is able to select the correct model consistently. As a side product of the derivation we establish consistency results for a group lasso estimator and oracle properties for an adaptive version of it. We give a Huber Proposal 2 approach to the estimation of the scale parameter and show that it is consistent. On the numerical side, we propose new algorithms for the implementation of our penalized M-estimators. We illustrate the finite sample performance of our proposals, at the model and under contamination, in simulation studies. A common feature to the study of GLM and GAM is the focus on estimators resulting from a robust quasilikelihood based maximization problem. This allows us to account for responses drawn the exponential family and to naturally introduce robustness considerations through this Mallow’s type of loss function. It is therefore a convenient choice, but most of the theoretical results hold more generally. Overall, our asymptotic analyses extend the currently available theory in two main direction: from linear models to GLM and GAM and from classical
likelihood based estimation to a robust one.

An important contribution of this thesis is to formally study the local robustness properties of general nondifferentiable penalized M-estimators. We show that penalized M-estimators have a bounded bias in a small contamination neighborhood of the model provided both the score function and its derivative are bounded. This characterization is in the spirit of infinitesimal robustness of Hampel et al. [1986] although it is not based on the form of the influence function. We also rigorously explore this approach by providing a formal framework that allows to define the influence function. In particular, we view it as the limiting influence function of a sequence of approximating functionals. We show that it can be used to characterize the robustness properties of a wide range of sparse estimators. An interesting feature of this construction is that the influence function of nondifferentiable penalized M-estimators becomes a derivative in the sense of distribution theory. Finally, we discuss some extensions of our work and give an overview of the future challenges of robust statistics in high dimensions.

1.4 Organization of this thesis

The core of the thesis is organized in four self contained chapters. The generality of the results in each of the four main chapters increases progressively in each of them. In Chapter 2 we start by revisiting some asymptotic results for GLM with a diverging number of parameters. In Chapter 3 we focus on robust high dimensional estimators for GLM. In Chapter 4 we propose a general approach to characterize the robustness properties of penalized M-estimators via the influence function. In Chapter 5 we consider the problem of robust high dimensional GAM estimation. Finally, we conclude this thesis with a discussion on some general issues regarding robust statistics in high dimensions.
Chapter 2

Robust generalized linear models with a diverging number of parameters

General remarks

We study the asymptotic properties of M-estimators for generalized linear models (GLM) when the number of covariates diverges. More precisely, we establish asymptotic normality of a class of Mallow’s type M-estimators and prove consistency of the estimated covariance matrix. We use these results to reexamine the asymptotic distribution of likelihood ratio type and Wald type tests based on these estimators. In the derivation of the results the dimension $p$ is allowed to increase with the sample size $n$ in the order $p = o(\sqrt{n})$. With stronger conditions on the design matrix we can weaken the dimensionality to $p = o(n/\log n)$. Our analysis improves greatly on existing results for tests in GLM that only allow for $p = o(n^{1/5})$. Finally we show that building on our asymptotic analysis for M-estimators, we can improve some existing oracle properties of penalized M-estimators in the high dimensional set up.

2.1 Introduction

Huber [1973] was the first author to study the asymptotic behavior of M-estimators in the linear model when the number of covariates diverges with the sample size. More specifically he considered the model $y_i = x_i^T \beta_0 + e_i$ with $\beta_0 \in \mathbb{R}^p$ and $e_i$ as independent errors, and M-estimators $\hat{\beta}$ that solve $\sum_i \xi(y_i - x_i^T \beta)x_i = 0$ for some score function $\xi : \mathbb{R} \rightarrow \mathbb{R}$. Huber [1973] established that if $\xi$ is smooth and $a \in \mathbb{R}^p$, $a^T \hat{\beta}$ is asymptotically normal if $p^3/n \rightarrow 0$ as $n \rightarrow \infty$.

Huber’s results have been further studied and improved upon by a number of authors. Yohai and Maronna [1979] showed that in appropriate balanced cases, asymptotic normality is achievable with $p^{5/2}/n \rightarrow 0$. Portnoy [1985] further weakened the condition to $(p \log n)^{3/2}/n \rightarrow 0$ and Mammen [1989] to $p^{3/2} \log n/n \rightarrow 0$. Welsh [1989] allowed $\xi$ to have jump discontinuities under $p^4(\log n)^2/n \rightarrow 0$. Bai and Wu [1994] viewed the condition on $p$ as an integrated part of the design conditions. He and Shao [2000] studied the asymptotic properties of M-estimators with increasing dimensionality in the general set up of parametric models.
In this paper we derive asymptotic theory for Mallow’s type GLM estimators with a parameter of increasing dimension. We therefore provide theoretical results for both classical and robust estimators. In particular, we show that assuming very mild conditions on the design matrix, \( p^2/n \to 0 \) is enough to obtain consistency and asymptotic normality as well as consistency of the sandwich formula of the covariance matrix. This also allows us to significantly weaken the condition \( p = o(n^{1/5}) \) required in Zhang et al. [2014a] for the construction of robust likelihood ratio type and Wald type tests. Strengthening our conditions on the design matrix we only need \( p \log(n)/n \to 0 \).

The rest of the paper is organized as follows. Section 2 discusses some approaches to robust estimation in GLM based on M-estimators. Section 3 provides the main asymptotic results. Section 4 revisits the oracle properties for penalized estimators in the high dimensional context and Section 5 concludes with some general comments. We relegate all the proofs to the Appendix.

### 2.2 Robust estimation in GLM

#### 2.2.1 Background

Generalized linear models (McCullagh and Nelder [1989]) provide a powerful tool that includes the linear models as a particular case. They allow us to model how the expected value of a response variable, belonging to the exponential family, changes as a function of some covariates. More precisely a GLM assumes that the response variables \( Y_i \) for \( i = 1, \ldots, n \) are drawn from independent distributions of the type

\[
    f(y; \theta_i) = \exp \left( \left( y_i \theta_i - b(\theta_i) \right)/\phi + c(y_i, \phi) \right), \quad \forall i = 1, \ldots, n
\]

where \( a(\cdot), b(\cdot) \) and \( c(\cdot) \) are some specific functions and \( \phi \) a nuisance parameter. Thus \( E(Y_i) = b'(\theta_i) \) and \( Var(Y_i) = \phi b''(\theta_i) \). Furthermore it is assumed that

\[
    E[Y_i|x_i] = \mu_i \quad \text{and} \quad Var[Y_i|x_i] = v_i = v(\mu_i).
\]

A GLM then defined by \( g(\mu_i) = \eta_i = x_i^T \beta_0 \), where \( \beta_0 \in \mathbb{R}^p \) is the vector of parameters, \( x_i \in \mathbb{R}^p \) is the set of explanatory variables and \( g(\cdot) \) the link function.

Let us recall some facts about M-estimators before discussing three robust approaches to the estimation of GLM based on them. An M-estimator \( T_n \) of \( \theta \) (Huber [1964]) is defined as a solution to

\[
    \sum_{i=1}^{n} \psi(z_i, T_n) = 0,
\]

where \( z_1, \ldots, z_n \) are i.i.d. realizations of a general parametric model \( \{F_\theta\} \). Under some regularity conditions, in a fixed parameter set up \( T_n \) is consistent at the model and asymptotically normally distributed. The robustness properties of an estimator can be characterized via the influence function. In particular a bounded influence function implies stability in the sense of local robustness; cf. Hampel et al. [1986]. For an M-estimator, it takes the form \( IF(z; \psi, F_\theta) = M(\theta, F_\theta)^{-1} \psi(z, \theta) \), where \( M(\theta, F_\theta) = -f(\partial \psi)/(\partial \theta)(z, \theta)dF_\theta(z) \). Therefore a bounded \( \psi \) function characterizes a bounded-influence, hence robust estimator.
2.2.2 The robust quasilikelihood

Cantoni and Ronchetti [2001] proposed robust estimators for GLM that can be viewed as a robustification of the quasilikelihood estimators of Wedderburn [1974]. They defined a robust quasilikelihood for GLM as

\[ \rho_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} Q_M(y_i, x_i^T \beta), \]  

where the functions \( Q_M(y_i, x_i^T \beta) \) can be written as

\[ Q_M(y_i, x_i^T \beta) = \int_{\tilde{s}}^{\mu_i} \psi(r_i) \frac{1}{\sqrt{v_i}} w(x_i) dt - \frac{1}{n} \sum_{j=1}^{n} \int \left[ \psi(r_j) \frac{1}{\sqrt{v_j}} \right] w(x_j) dt \]

with \( r_i = (y_i - \mu_i) / \sqrt{v_i} \), \( \tilde{s} \) such that \( \psi((y_i - \tilde{s}) / \sqrt{v(\tilde{s})}) = 0 \) and \( \tilde{t} \) such that \( E[\psi((y_i - \tilde{s}) / \sqrt{v(\tilde{s})})] = 0. \psi(\cdot) \) is a bounded function that protects against large outlying points in the space of the response variables and \( w(\cdot) \) downweighs leverage points. The estimator of \( \hat{\beta} \) of \( \beta_0 \), derived from the minimization of this loss function, is the solution of the estimating equations

\[ \sum_{i=1}^{n} \Psi(y_i, x_i^T \beta) = \sum_{i=1}^{n} \left[ \psi(r_i) \frac{1}{\sqrt{v_i}} w(x_i) \frac{\partial \mu_i}{\partial \beta} - a(\beta) \right] = 0, \]

where \( a(\beta) = \frac{1}{n} \sum_{i=1}^{n} E[\psi(r_i)] \frac{1}{\sqrt{v_i}} w(x_i) \partial \mu_i / \partial \beta \) is a constant ensuring Fisher consistency.

2.2.3 Bounded deviance estimation

An alternative way of constructing robust estimators is to apply a bounded function \( \rho \) on the unscaled deviance components (up to a factor 2)

\[ d_i = \log f(y_i; y_i) - \log f(y_i; \mu_i). \]

In this case the loss function minimized by the robust estimator is

\[ \rho_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \rho(\log f(y_i; y_i) - \log f(y_i; \mu_i)) + A(\beta), \]

where \( A(\beta) \) is a constant assuring Fisher consistency. This approach is in the spirit of the estimator of Bianco and Yohai [1996] for logistic regression and has been further explored in Croux and Haesbroeck [2003], Bianco et al. [2005] and Bianco et al. [2011]. Aeberhard et al. [2014] point out the close relationship between the bounded deviance estimator and the robust quasilikelihood estimator. In particular, they show the estimating equations of both estimators can be viewed as weighted versions of the maximum likelihood score functions and that the resulting estimators may be equivalent in practical terms.
2.2.4 Robust estimation based on Bregman divergence

Zhang et al. [2014a] introduced a robust estimator for GLM motivated by the Bregman divergence (BD). For a given concave function $q$, Bregman [1967] introduced a device for constructing a divergence as

$$Q_{q}(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)q'(\mu).$$

It encompasses a large class of well known functions commonly used in regression and classification; for details see Zhang et al. [2009]. For instance, choosing $q(\mu) = a\mu - \mu^2$ for some constant $a$ yields the quadratic loss $Q_{q}(y, \mu) = (y - \mu)^2$. The choice $q(\mu) = \int_{a}^{\mu} \chi(s) - \mu \chi'(s) ds$, where $a$ is a finite constant such that the integral is well defined, recovers the classical (negative) quasilikelihood.

The robust BD of Zhang et al. [2014a] is defined by

$$\rho_{q}(y, \mu) = \int_{y}^{\mu} \chi(r)q''(s)\sqrt{v(s)}ds - G(\mu),$$

where $q$ denotes the generating $q$-function of the BD, $q''$ is the second derivative of $q$, $\chi$ is a bounded function and $G(\mu)$ is a Fisher consistency term. When $\chi(r) = r$ we recover the usual BD.

2.2.5 Alternative approaches to robust GLM

The three above mentioned approaches are very closely related as they are define Mallow’s type M-estimators for GLM and our theoretical results are valid for all of them. Our analysis does not cover the general optimally bounded conditional unbiased estimators of Künnch et al. [1989], the trimmed likelihood approach Müller and Neykov [2003] nor the projection estimators of Bergesio and Yohai [2011]. In the particular case of the linear model an appealing robust approach, combining high efficiency and high breakdown point, is the MM-estimator of Yohai [1987]. Note that although these alternative GLM approaches that we are not covering have desirable robustness properties, we believe that they suffer from two main practical drawbacks. First, they do not define a simple robust loss function and they are therefore very difficult if not impossible to use for inference via a deviance type of analysis. Second, it seems very difficult to extend these constructions to high dimensional scenarios by considering appropriate penalty functions. Having a simple loss function allows to straightforwardly construct lasso type estimators for high dimensional problems. These ideas will be explored in detail in Chapters 3, 4 and 5.

2.3 Asymptotic Analysis

In this section we focus our analysis on the robust quasilikelihood estimator and the maximum likelihood estimator for GLM. The results hold for the bounded deviance and robust BD approaches introduced in the previous sections by adapting the assumptions in a straightforward way.

2.3.1 Regularity conditions

We make the following assumptions for our theoretical analysis of the robust quasilikelihood estimator.
2.3. Asymptotic Analysis

(\textbf{A1}) $\Psi_i(\beta) = \Psi(y_i, x_i^T \beta)$ is bounded as well as $|\psi'(r)|$, $|\psi'(r)r|$, $|\psi''(r)|$, $|\psi''(r)r|$ and $|\psi''(r)r^2|$, where $\psi'$ and $\psi''$ denote the first two derivatives with respect to $\eta$. In a neighborhood $\mathcal{O}$ of $\beta_0$ we have

$$M(\beta) = -E\left[\Psi_i(\beta)\right] < \infty \quad \text{and} \quad Q(\beta) = E\left[\Psi_i(\beta)\Psi_i^T(\beta)\right] < \infty,$$

where $\Psi_i(\beta) = \partial\Psi_i(\beta)/\partial\beta^T$. Furthermore $\partial^2 g^{-1}(\cdot)/\partial\eta^3$ and $\partial^2 v(\cdot)/\partial\mu^2$ are continuous, and $b_i(\beta) = E[\psi(r_i) \frac{\partial}{\partial \eta} \log h(y_i|x_i, \mu_i)] w(x_i) \frac{1}{\sqrt{\eta}} (\frac{\partial \mu_i}{\partial \eta})^2$ is positive and continuous in $\mathcal{O}$, where $h(\cdot)$ is the conditional density or probability of $y_i|x_i$.

(\textbf{A2}) $\max_{i,j} |x_{ij}| < \infty$ and $S_n = \sum_{i=1}^n x_i x_i^T$ is such that $0 < c_1 n \leq \lambda_{\min}(S_n) \leq \lambda_{\max}(S_n) \leq c_2 n < \infty$, for some positive constants $c_1$ and $c_2$.

Assumption (A1) is fairly mild for a robust loss function and is satisfied with an appropriate choice of $\psi$. It is similar to condition A5 in Zhang et al. [2014a]. (A2) is a classical assumption in the linear model literature. The boundedness assumption on the covariates implies $\max_i \|x_i\| = O(\sqrt{n})$, where $\| \cdot \|$ denotes the $L_2$-norm. We can adapt our results to cover maximum likelihood estimators under usual more stringent assumptions. In particular, we require fourth moment conditions. Note that in classical GLM with canonical link, the set of estimating equations becomes $\sum_{i=1}^n \Psi_i(\beta) = \sum_{i=1}^n (y_i - \mu(x_i^T \beta))$ and will therefore in general have an unbounded influence function.

(\textbf{A1'}) $\Psi_i(\beta)$ has two derivatives and there is a neighborhood $\mathcal{O}$ of $\beta_0$ such that

$$J(\beta) = -E\left[\Psi_i(\beta)\right] = E\left[\Psi_i(\beta)\Psi_i^T(\beta)\right] < \infty$$

is positive definite. Furthermore $\partial^2 g^{-1}(\cdot)/\partial\eta^3$ and $\partial^2 v(\cdot)/\partial\mu^2$ are continuous and $E[(y_i - \mu_i)^4] < \infty$.

2.3.2 Asymptotic normality

Let $a \in \mathbb{R}^p$ with $\|a\| = 1$, $M_n(\beta) = -E\left[\sum_{i=1}^n \Psi_i(\beta)\right]$, $Q_n(\beta) = E\left[\sum_{i=1}^n \Psi_i(\beta)\Psi_i^T(\beta)\right]$, $M_n = M_n(\beta_0)$, $Q_n = Q_n(\beta_0)$ and $V_n = M_n^{-1}Q_nM_n^{-1}$. Note that $M_n(\beta) = \sum_{i=1}^n b_i(\beta) x_i x_i^T$. In the classical case $J_n = M_n = Q_n$. To simplify the notation we write $\Psi^{(n)}(\beta) = \sum_{i=1}^n \Psi_i(\beta)$.

\textbf{Theorem 1.} Suppose conditions (A1)-(A2) or (A1')-(A2) hold. Then if $p^2/n \to 0$ as $n$ goes to infinity, there exists a sequence of estimators $\hat{\beta}$ such that

$$P\left(\Psi^{(n)}(\hat{\beta}) = 0\right) \to 1 \quad (2.3)$$

and

$$a^T V_n^{-1/2}(\hat{\beta} - \beta_0) \to_d N(0, I) \quad (2.4)$$

\textbf{Remark A :} Although it is easy to construct smooth $\psi$ functions as required in (A1), obviously the frequently used Huber function $\psi_c$ does not satisfy this condition. Simulations involving this choice of $\psi$ function can be found in Zhang et al. [2014a]. We can recover an analogous version of Theorem 1 when using $\psi_c$ as long as $P(r_i(\beta_0) = \pm c) \to 0$. More details can be found in Appendix 2.5.2. Using this result, analogous versions of Theorems 2-5 follow straightforwardly. A more general treatment of nondifferentiable score functions when $p^2/n \to 0$ as $n \to \infty$ goes beyond the scope of this thesis.
2.3.3 Estimation of the covariance matrix

The asymptotic variance $V_n$ obtained in (3.16), depends on the unknown parameter $\beta_0$. A natural estimate of the covariance matrix is $\hat{V}_n = \hat{M}_n^{-1}\hat{Q}_n\hat{M}_n^{-1}$, where $\hat{M}_n = M_n(\hat{\beta})$ and $\hat{Q}_n = Q_n(\hat{\beta})$. Therefore it is interesting to know whether $\hat{V}_n$ is consistent under the conditions of Theorem 1. This issue is addressed by the following result.

**Theorem 2.** Suppose conditions (A1)-(A2) or (A1')-(A2) hold. Then if $p^2/n \to 0$ as $n$ goes to infinity we have

\[ \|\hat{V}_n - V_n\| = o_p(1/n) \] (2.5)

and

\[ n(A_n\hat{V}_n A_T^n - A_nV_nA_T^n) \to_p 0 \] (2.6)

for any $q \times p$ matrix $A_n$ such that $A_nA_T^n \to G$, where $q$ is any fixed integer.

2.3.4 Inference

We consider the problem of testing linear hypothesis, that is

\[ H_0 : A_n\beta_0 = \alpha_0 \quad \text{vs.} \quad H_1 : A_n\beta_0 \neq \alpha_0, \] (2.7)

where $A_n$ is a given $q \times p$ matrix such that $A_n A_T^n = G$ with $G$ a $q \times q$ positive definite matrix and $\alpha_0$ a known $q$ dimensional vector. We study Wald-type test statistics of the form

\[ W_n = (A_n\hat{\beta} - \alpha_0)^T(A_n\hat{M}_n^{-1}\hat{Q}_n\hat{M}_n^{-1}A_T^n)^{-1}(A_n\hat{\beta} - \alpha_0) \]

as well as likelihood ratio type tests of the form

\[ \Lambda_n = 2n\left(\min_{\beta \in \mathbb{R}^p : A_n\beta = \alpha_0} \rho_n(\beta) - \min_{\beta \in \mathbb{R}^p} \rho_n(\beta)\right). \]

The next two theorems establish respectively the asymptotic distribution of $W_n$ and $\Lambda_n$ under $H_0$. Theorem 3 shows that the Wald statistic is approximately pivotal under the null with a growing number of parameters. Theorem 4 states that the asymptotic distribution of $\Lambda_n$ under the null is in general not distribution free. It shows however that this is the case for the classical likelihood ratio test. This result is in line with Heritier and Ronchetti [1994], Cantoni and Ronchetti [2001] and Zhang et al. [2014a]. Both results greatly improve on those of Zhang et al. [2014a] who required $p = o(n^{1/5})$.

**Theorem 3.** Suppose conditions (A1)-(A2) or (A1')-(A2) hold. Then under $H_0$ in (2.7), we have

\[ W_n \to_d \chi^2_q, \]

provided $p^2/n \to 0$ as $n \to \infty$.

**Theorem 4.** Suppose conditions (A1)-(A2) or (A1')-(A2) hold. Then under $H_0$ in (2.8), if $p^2/n \to 0$ as $n \to \infty$, we have

\[ \Lambda_n = nZ_nM_n^{-1/2}P_{M_n^{-1/2}A_T^nM_n^{-1/2}}Z_n + o_p(1), \]

where $\sqrt{n}Z_n \sim \mathcal{N}(0, Q_n)$ and $P_B = B(B^T B)^{-1}B^T$ is defined for a matrix $B$ such that $(B^T B)^{-1}$ exists.
2.4. Concluding remarks

Remark B: Note that $P_{M_n^{-1/2}A_n^T}^{M_n^{-1/2}A_n^T} = C_n^T C_n$ where $C_n$ is a $q \times p$ matrix satisfying $C_n C_n^T = I_q$. Since (A1') implies $J_n = Q_n = M_n$, Theorem 4 yields in this case $\Lambda_n \to_d \chi_q^2$ because $\sqrt{n} C_n J_n^{-1/2} Z_n \sim \mathcal{N}(0, I_q)$.

We also consider contiguous alternatives of the form

$$H_{1n}: A_n \beta_0 - \alpha_0 = n^{-1/2}d + o(n^{-1/2}),$$

where $d$ is a fixed $q$ dimensional constant. We explore the local power of $W_n$ by establishing the asymptotic distribution under a contiguous alternative.

**Theorem 5.** Assume (A1)-(A2) or (A1')-(A2) hold, and $A_n M_n^{-1} Q_n M_n^{-1} A_n^T \to_p G$ where $G$ is a $q \times q$ positive definite matrix. Then under $H_{1n}$ in (2.8), if $p^2/n \to 0$ as $n \to \infty$, we have

$$W_n \to_d \chi_q^2(\tau^2)$$

with the non-centrality parameter $\tau^2 = d^T G^{-1} d$.

As pointed out in Zhang et al. [2014a], $W_n$ has some advantages over $\Lambda_n$. In particular $W_n$ is asymptotically a pivot and computationally cheaper.

### 2.3.5 Discussion

All the theorems given above depend on the arguments used to prove Theorem 1 which establishes the existence and asymptotic normality of the solution $\hat{\beta}$ of (2.2). Our proof is essentially an adaptation of the injection function argument used in Liang and Du [2012] to show that asymptotic normality for the maximum likelihood estimator in logistic regression. The major difference with our result is that they claim asymptotic normality can be reached by assuming only $p/n \to 0$. It seems to us that there might be a problem in their argument since they implicitly obtain faster rates of convergence than the $\sqrt{p/n}$ minimax lower bounds of Chen et al. [2014]. Indeed, they show the existence of $\hat{\beta}$ in the neighborhood $N_n(\delta) = \{ \beta : \|V_n^{-1/2}(\beta - \beta_0)\| \leq \delta \}$ which under (A1) shrinks as $n^{-1/2}$. Although our result is weaker, it is in line with the counterexample of Portnoy [1986] that shows the distribution of a $p$ dimensional sample mean cannot be approximated uniformly with a normal distribution if $p^2/n \to 0$ is not satisfied.

In the linear model, recent work by El Karoui et al. [2013], El Karoui [2013] and Donoho and Montanari [2016] show asymptotic normality for M-estimators assuming $p/n \to r \in (0, 1)$. Unlike us, all these papers assume a random design where all the $x_i$s are independent and identically distributed subgaussian random variables. It is possible to weaken our dimensionality conditions to $p \log(n)/n$ by assuming that $\max_i |x_i^T u| = O(\log n)$ for every $u \in \mathbb{R}^p$ with $\|u\|$ bounded. Note that this assumption was made in Portnoy [1985] and Li et al. [2011], and can be shown to hold in probability if the $x_i$s are independent with $x_i \sim N(0, s_i)$ where $s_1, \ldots, s_n$ are i.i.d. according to a distribution with compact support in $(0, \infty)$.

### 2.4 Concluding remarks

We have revisited a class of M-estimators for GLM with a diverging number of parameters. We establish that under a growing dimensionality set up, the condition $p = o(\sqrt{n})$ is
enough to obtain asymptotic normality, consistency of the sandwich formula and asymptotic \( \chi^2 \) distributions for Wald and likelihood ratio type tests. All these results are established under significantly weaker conditions on the dimensionality of the problem than the existing ones in the literature. Zhang et al. [2014a] provide a large simulation study covering the estimation and inference of robust GLM. They also illustrate the performance of the robust estimates and tests in a real data example.

## 2.5 Appendix

### 2.5.1 Proof of main results

#### Technical Lemmas

We require some additional notation and auxiliary lemmas to prove our main results. Let \( N_n(\delta) = \{ \beta : \| V_n^{-1/2}(\beta - \beta_0) \| \leq \delta \sqrt{n} \} \) and \( M_n(\beta) = - \sum_{i=1}^{n} \int_{0}^{1} \Psi_i(\beta_0 + s(\beta - \beta_0)) ds \) for \( \beta \in N_n(\delta) \). We use \( c \) as a generic constant taking different values in different places.

**Hoeffding’s inequality.** Let \( Z_1, \ldots, Z_n \) be independent random variables in some space \( Z \) and let \( \gamma \) be a real-valued function on \( Z \) satisfying \( E[\gamma(Z_i)] = 0 \) and \( |\gamma(Z_i)| \leq c_i \forall i. \) Then we have for all \( t \)

\[
P\left( \left| \sum_{i=1}^{n} \gamma(Z_i) \right| \geq t \right) \leq 2 \exp\left( - \frac{t^2}{2 \sum_{i=1}^{n} c_i^2} \right).
\]

**Proof:** See e.g. Lemma 14.11 in Bühlmann and Van De Geer [2011]. \( \square \)

**Lemma 1.** If \( g \) is continuously differentiable in a convex interval of \( \mathbb{R} \), then for \( t_1, t_2 \in \mathbb{R} \) we have

\[
g(t_2) - g(t_1) = (t_2 - t_1) \int_{0}^{1} g'(t_1 + u(t_2 - t_1)) du,
\]

where \( g' \) denotes the derivative of \( g \).

**Proof:** The equality is verified by integrating the right hand side expression. \( \square \)

**Lemma 2.** Let \( h \) be a smooth injection from \( \mathbb{R}^p \) to \( \mathbb{R}^p \) with \( h(x_0) = y_0 \) and \( \inf_{\| x - x_0 \| = \delta} \| h(x) - y_0 \| \geq r > 0 \). Then for any \( y \) with \( \| y - y_0 \| \leq r \), there is an \( x \) with \( \| x - x_0 \| \leq \delta \) such that \( h(x) = y \).

**Proof:** See Lemma A in Chen et al. [1999]. \( \square \)

**Lemma 3.** Assume (A1)-(A2) and \( p/n \to 0 \). Then \( \frac{1}{n} \| \Psi^{(n)}(\beta) + M_n(\beta) \| = o_p(1) \) for \( \beta \in \mathcal{O} \).

**Proof:** Note that

\[
a^T \left( \Psi(\beta) + M_n(\beta) \right) a = a^T \sum_{i=1}^{n} \left( \phi_i'(\beta) - b_i(\beta) \right) x_i x_i^T a = \sum_{i=1}^{n} \gamma_i,
\]

where \( E[\phi_i'(\beta)] = b_i(\beta) \). By construction \( E[\gamma_i] = 0 \) and (A1) implies \( |\gamma_i| \leq (a^T x_i)^2 \bar{c} \). Then Hoeffding’s inequality yields

\[
P\left( \left| \sum_{i=1}^{n} \gamma_i \right| \geq t n \right) \leq 2 \exp\left( - \frac{n^2 t^2}{2K^2 \max_i (a^T x_i)^2 \sum_{i=1}^{n} (a^T x_i)^2} \right) \to 0
\]
since \(\max_i (a^T x_i)^2 \leq O(p)\) and \(\frac{1}{n} a^T \sum_{i=1}^n x_i x_i^T a = O(1)\) imply
\[
\frac{n^2 I^2}{2K^2 \max_i (a^T x_i)^2 \sum_{i=1}^n (a^T x_i)^2} = O(n/p).
\]
\[
\square
\]

**Lemma 4.** Under (A1)-(A2) or (A1')-(A2), if \(p^2/n \to 0\) we have
\[
\text{(a)} \quad \sup_{\beta \in N_n(\delta)} |a^T M_n^*(\beta) M_n^{-1} a - 1| \to 0
\]
\[
\text{(b)} \quad \sup_{\beta \in N_n(\delta)} |a^T Q_n^{-1/2} M_n^*(\beta) M_n^{-1} Q_n^{1/2} a - 1| \to 0.
\]

**Proof:** We assume (A1)-(A2) throughout the proof. Assuming (A1')-(A2) instead and following the same steps, we reach the same conclusions. Let \(\beta^*\) lie between \(\beta\) and \(\beta_0\). Then for \(\beta \in N_n(\delta)\) we have
\[
a^T M_n^{-1/2} M_n(\beta) M_n^{-1/2} a - 1 = a^T M_n^{-1/2} \left( M_n(\beta) - M_n(\beta_0) \right) M_n^{-1/2} a
\]
\[
= a^T M_n^{-1/2} \sum_{i=1}^n \frac{\partial h_i(\beta^*)}{\partial \eta_i} w(x_i) x_i x_i^T (\beta - \beta_0) M_n^{-1/2} a
\]
\[
\leq c_3 \max_{1 \leq i \leq n} |x_i^T (\beta - \beta_0)| a^T M_n^{-1/2} \sum_{i=1}^n x_i x_i^T M_n^{-1/2} a
\]
\[
= O(p/\sqrt{n}),
\]
(2.9)
since
\[
a^T M_n^{-1/2} S_n M_n^{-1/2} a = O(1)
\]
and
\[
\max_{1 \leq i \leq n} |x_i^T (\beta - \beta_0)| \leq \max_{1 \leq i \leq n} \|V_n^{1/2} x_i\| \|V_n^{-1/2} (\beta - \beta_0)\| = O(\delta p \sqrt{n}).
\]
(2.10)
Then Lemma 3 and (2.9) yield
\[
-a^T M_n^{-1/2} \dot{\Psi}(n)(\beta) M_n^{-1/2} a - 1 \to_p 0
\]
Hence we have shown that
\[
\sup_{\beta \in N_n(\delta)} \left| a^T M_n^*(\beta) M_n^{-1} a - 1 \right| \to 0.
\]
This proves part (a).

Let us now turn to part (b). Part (a) implies that \(M_n^{-1/2} M_n^*(\beta) M_n^{-1/2} \to I\), where \(I\) denotes the identity matrix. Since \(M_n^{-1/2} M_n^*(\beta) M_n^{-1/2}\) have the same eigenvalues as \(Q_n^{-1/2} M_n^{1/2} M_n^{-1/2} M_n^*(\beta) M_n^{-1/2} M_n^{1/2} Q_n^{-1/2}\), we conclude that
\[
\sup_{\beta \in N_n(\delta)} \left| a^T Q_n^{-1/2} M_n^*(\beta) M_n^{-1} Q_n^{1/2} a - 1 \right| \to 0.
\]
This completes the proof. \(
\square
\)

**Lemma 5.** Assume (A1)-(A2) and \(p^2/n \to 0\). Then \(\rho_n^*(\beta) = \rho_n(\beta) - \rho_n(\beta_0)\) is asymptotically concave.
Proof: Let $\bar{T}(\beta) = \frac{1}{n} \sum_{i=1}^{n} T_i(\beta) = \rho_n(\beta) - \rho_n(\beta_0) - E[\rho_n(\beta) - \rho_n(\beta_0)]$. Note that for $\beta \in \mathcal{O}$, by (A1) and (A2) every $T_i(\beta)$ is bounded by some positive constant $\bar{c}$ because

$$\rho_n(\beta) - \rho_n(\beta_0) = \frac{1}{n} \sum_{i=1}^{n} \Psi_i^T(\beta^*)(\beta - \beta_0) = \frac{1}{n} \sum_{i=1}^{n} \phi_i^T(\beta^*) x_i^T (\beta - \beta_0)$$

$$\leq \max_{1 \leq i \leq n} |x_i^T (\beta - \beta_0)| \max_{1 \leq i \leq n} |\phi_i^T(\beta^*)|$$

$$\leq \mathcal{O}(p^2/\sqrt{n}),$$

where $\beta^*$ lies between $\beta$ and $\beta_0$, and the last inequality comes from (2.10). It follows from Hoeffding’s inequality that

$$P(|\bar{T}(\beta)| > t) = P\left(\sum_{i=1}^{n} T_i(\beta) | > tn \right) < 2 \exp \left( - \frac{nt^2}{2\sigma^2} \right) = o(1).$$

Hence $\bar{T}(\beta) = o_p(1)$ and $\rho_n^*(\beta) = E[\rho_n^*(\beta)] + o_p(1)$. Combining this with a Taylor expansion of order two of $\rho_n(\beta)$ we have

$$\rho_n^*(\beta) = \frac{1}{2} (\beta - \beta_0)^T M_n(\beta^*) (\beta - \beta_0) + o_p(1),$$

where $\beta^*$ lies between $\beta$ and $\beta_0$. Since (A1) guarantees that $M_n(\beta)$ is negative definite for $\beta \in \mathcal{O}$, this completes the proof. \qed

Proof of Theorem 1

The proof is an adaptation of the three steps of the proof of Theorem 1 in Liang and Du [2012]. In the first step we establish the asymptotic normality of $\Psi^{(n)}(\beta_0)$. Step two establishes (2.3). In step three we show that $a^T V_n^{-1/2} (\hat{\beta} - \beta_0)$ can be approximated by a linear combination of $\Psi^{(n)}(\beta_0)$ and hence establishes (2.4). We assume (A1)-(A2) throughout the proof as we can reach the same conclusions assuming instead (A1')-(A2) and following the same ideas. Special care has to be taken only in step one as indicated below.

**Step 1.** We will show that

$$a^T Q_n^{-1/2} \Psi^{(n)}(\beta_0) \rightarrow N(0,1). \quad (2.11)$$

Let $z_i = a^T Q_n^{-1/2} \Psi_i(\beta_0) = a^T Q_n^{-1/2} x_i \phi_i(\beta_0)$ and notice that $E[z_i] = 0$. It therefore suffices to check that the Lindeberg-Feller condition

$$\sum_{i=1}^{n} E[z_i^2 1_{|z_i| > \epsilon}] \rightarrow 0 \quad (2.12)$$

is satisfied for any $\epsilon > 0$. Conditions (A1)-(A2) imply that $|\phi_i(\beta_0)| \leq c_5 < \infty$,

$$\max_{1 \leq i \leq n} \left( a^T Q_n^{-1/2} x_i \right)^2 = \max_{1 \leq i \leq n} a^T Q_n^{-1/2} x_i x_i^T Q_n^{-1/2} a = O(n^{-1})$$

and

$$\sum_{i=1}^{n} (a^T Q_n^{-1/2} x_i)^2 = \sum_{i=1}^{n} a^T Q_n^{-1/2} x_i x_i^T Q_n^{-1/2} a = a^T Q_n^{-1/2} S_n Q_n^{-1/2} a \leq c_5.$$
Therefore we have \( \max_i z_i^2 \to 0 \) and \( \sum_i z_i^2 \leq c_5 \). Bearing this in mind, the Cauchy-Schwarz inequality ensures (2.12).

We verify (2.12) with a different argument when we assume (A1’) instead of (A1). In this case, from the Cauchy-Schwarz inequality we have

\[
\sum_{i=1}^n E[z_i^21_{\{z_i>\epsilon\}}] = \sum_{i=1}^n (a^T J_n^{-1/2} x_i)^2 E[(y_i - \mu_i)^2] 1_{\{a^T J_n^{-1/2} x_i (y_i - \mu_i) > \epsilon\}} \\
\leq \sum_{i=1}^n (a^T J_n^{-1/2} x_i)^2 E[(y_i - \mu_i)^4]^{1/2} P(|a^T J_n^{-1/2} x_i (y_i - \mu_i)| > \epsilon)^{1/2} \\
= O(\sqrt{\frac{p}{n}}),
\]

because \( \sum_{i=1}^n a^T J_n^{-1/2} x_i x_i^T J_n^{-1/2} a = O(1), E[(y_i - \mu_i)^4]^{1/2} \leq c_6 \) and by Markov’s inequality

\[
P(|a^T J_n^{-1/2} x_i (y_i - \mu_i)| > \epsilon) \leq \frac{(a^T J_n^{-1/2} x_i)^2 E[(y_i - \mu_i)^2]}{\epsilon^2} = O(\|J_n^{-1/2} x_i\|^2) = O(p/n).
\]

**Step 2.** We now prove that for any \( \epsilon > 0 \) there is a \( \delta > 0 \) such that when \( n \) is large enough

\[
P(\text{there is a } \hat{\beta} \in N_n(\delta) \text{ such that } \Psi^{(n)}(\hat{\beta}) = 0) > 1 - \epsilon.
\]

By the Cauchy-Schwarz inequality, for any \( \delta > 0 \),

\[
\inf_{\beta \in \partial N_n(\delta)} \delta^2 \left( \frac{1}{\delta^2} (\beta - \beta_0)^T M_n^*(\beta) Q_n^{-1/2} V_n^{-1/2} (\beta - \beta_0) \right)^2 \\
\leq \inf_{\beta \in \partial N_n(\delta)} (\beta - \beta_0)^T M_n^*(\beta) Q_n^{-1} M_n^*(\beta) (\beta - \beta_0).
\]

It follows from Lemma 4 that for any \( \epsilon > 0 \) and \( \delta > 0 \), there is a \( c_7 \in (0, 1) \) independent of \( \delta \), such that for a large \( n \)

\[
P\left( \inf_{|\beta| = 1, \beta \in \partial N_n(\delta)} e^T Q_n^{-1/2} M_n^*(\beta) M_n^{-1} Q_n^{1/2} e \geq c_7 \right) > 1 - \frac{\epsilon}{4}. \tag{2.14}
\]

Furthermore from Lemma 1 we obtain

\[
Q_n^{-1/2} \left( \Psi^{(n)}(\beta) - \Psi^{(n)}(\beta_0) \right) = -Q_n^{-1/2} \left( M_n^*(\beta) (\beta - \beta_0) \right). \tag{2.15}
\]

Combining (2.13)-(2.15), for any \( \delta > 0 \) and \( n \) large enough we have

\[
P\left( \inf_{\beta \in \partial N_n(\delta)} \|Q_n^{-1/2} (\Psi^{(n)}(\beta) - \Psi^{(n)}(\beta_0)) \| \geq c_7 \delta \right) > 1 - \frac{\epsilon}{4}. \tag{2.16}
\]

Taking \( \delta = (4/\epsilon)/c_7 \), for a sufficiently large \( n \), the Markov inequality and (2.11) yield

\[
P\left( |a^T Q_n^{-1/2} \Psi^{(n)}(\beta_0)| \leq c_7 \delta \right) \geq 1 - E\left[ (a^T Q_n^{-1/2} \Psi^{(n)}(\beta_0))^2 \right]/(c_7 \delta)^2 \\
\geq 1 - 1/(c_7 \delta)^2 = 1 - \frac{\epsilon}{4}. \tag{2.17}
\]

Consider \( E_n = \left\{ \|Q_n^{-1} \Psi^{(n)}(\beta_0)\| \leq \inf_{\beta \in \partial N_n(\delta)} Q_n^{-1} \left( \Psi^{(n)}(\beta) - \Psi^{(n)}(\beta_0) \right) \| \right\} \) and \( E_n^* = \{ \det M_n^*(\beta_1, \beta_2) \neq 0 \text{ for all } \beta_1, \beta_2 \in N_n(\delta) \} \), where \( M_n^*(\beta_1, \beta_2) = E\left[ \sum_{i=1}^n \int_0^1 \psi_i(\beta_1 + s(\beta_2 - \beta_1)) ds \right] \). (2.16) and (2.17) imply \( P(E_n) > 1 - \epsilon/2 \). Note that (2.15) indicates that the
map: \( \beta \mapsto a^TQ_n^{-1/2}\Psi^{(n)}(\beta) \) is an injection for \( \beta \in N_n(\delta) \) on the set \( E_n^* \). Furthermore for \( n \) large enough, Lemma 4 gives \( P(E_n^*) > 1 - \epsilon/2 \). Hence using Lemma 2 and taking \( n \) sufficiently large, there is a \( \hat{\beta}_n \) on the set \( E_n \cap E_n^* \) such that
\[
\hat{\beta} \in N_n(\delta) \text{ and } \Psi^{(n)}(\hat{\beta}) = 0. \tag{2.18}
\]

**Step 3.** (2.3) implies that
\[
\Psi^{(n)}(\beta_0) = -M_n^*(\hat{\beta} - \beta_0). \tag{2.19}
\]
Therefore
\[
V_n^{-1/2}(\hat{\beta} - \beta_0) = -V_n^{-1/2}(M_n^*(\hat{\beta}))^{−1}\Psi^{(n)}(\beta_0)
\]
\[
= -Q_n^{-1/2}M_n(\hat{\beta})^{−1}Q_n^{1/2}M_n^{-1/2}\Psi^{(n)}(\beta_0)
\]
\[
= -Q_n^{-1/2}\Psi^{(n)}(\beta_0) + o_p(1), \tag{2.20}
\]
where the last equality is obtained using Lemma 4. (2.11) and (2.20) show (2.4).

**Proof of Theorem 2**

We assume (A1)-(A2) throughout the proof as we can reach the same conclusions assuming instead (A1')-(A2) and following the same ideas. Note that \( \|A_n(\hat{V}_n - V_n)A_n^T\| \leq \|\hat{V}_n - V_n\|_{\text{F}} \|A_n\|^2_{\text{F}} \), where \( \| \cdot \|_{\text{F}} \) denotes the Frobenious norm. Since \( \|A_n\|^2_{\text{F}} \to \text{tr}(G) \), it suffices to prove that \( \|\hat{V}_n - V_n\| = o_p(1/n) \).

First, we show that \( \frac{1}{n}\|\hat{M}_n - M_n\| = o_p(1) \). This can be done by noting that (2.10) and the consistency of \( \hat{\beta} \) imply that for large \( n \) we have
\[
\frac{1}{n}\|\hat{M}_n - M_n\| = \left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T (b_i(\hat{\beta}) - b_i(\beta_0)) \right\|
\]
\[
= \left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \frac{\partial b_i(\hat{\beta})}{\partial \eta_i} x_i^T (\hat{\beta} - \beta_0) \right\|
\]
\[
\leq c_3 \max_{1 \leq i \leq n} |x_i^T (\hat{\beta} - \beta_0)| \left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right\|
\]
\[
= O_p(p/\sqrt{n}). \tag{2.21}
\]

Second, we show that \( \frac{1}{n}\|\hat{Q}_n - Q_n\| = o_p(1) \). Note that
\[
Q_n(\beta) = \sum_{i=1}^{n} \left( x_i x_i^T (E[a_i^2(\beta)] - E^2[a_i(\beta)]) \right),
\]
where \( a_i(\beta) = \psi(r_i)w(x_i)\frac{1}{\sqrt{1/2(\mu_i)}} \frac{\partial \mu_i}{\partial \eta_i} \). Therefore as in (2.21) we have
\[
\frac{1}{n}\|\hat{Q}_n - Q_n\| = \left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \left( \text{Var}[a_i(\hat{\beta})] - \text{Var}[a_i(\beta_0)] \right) \right\|
\]
\[
\leq c_4 \max_{1 \leq i \leq n} |x_i^T(\hat{\beta} - \beta_0)| \left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right\|
\]
\[
= O_p(p/\sqrt{n}). \tag{2.22}
\]
Finally, we show that \( \| \hat{V}_n - V_n \| = o_p(1/n) \). Note that \( \hat{V}_n - V_n = T_1 + T_2 + T_3 \), where \( T_1 = \hat{M}_n^{-1}(\hat{Q}_n - Q_n)\hat{M}_n^{-1} \), \( T_2 = \hat{M}_n^{-1}(M_n - \hat{M}_n)\hat{M}_n^{-1}Q_n\hat{M}_n^{-1} = \hat{M}_n^{-1}(M_n - \hat{M}_n)\hat{M}_n^{-1}Q_n\hat{M}_n^{-1} + \hat{M}_n^{-1}(M_n - \hat{M}_n)\hat{M}_n^{-1}Q_n(\hat{M}_n^{-1} - M_n^{-1}) + T_2 + T_3 \) and \( T_3 = M_n^{-1}Q_n(\hat{M}_n^{-1} - M_n^{-1}) \). By (A1) and (2.21) we have \( \| \hat{M}_n^{-1} \| = O_p(1/n) \). Therefore (2.22) implies

\[
\| T_1 \| \leq \| \hat{M}_n^{-1} \| \| Q_n - Q_n \| \| \hat{M}_n^{-1} \| = o_p(1/n).
\]

From the same type of arguments we obtain \( \| T_2 \| = o_p(1/n) \) and \( \| T_3 \| = o_p(1/n) \). □

**Proof of Theorem 3**

Denote \( G_n = A_n M_n^{-1}Q_n M_n^{-1} A_n^T \) and \( \hat{G}_n = A_n \hat{M}_n^{-1}Q_n \hat{M}_n^{-1} A_n^T \). It suffices to show that

\[
\hat{G}_n^{-1/2}A_n(\hat{\beta} - \beta_0) \to_d \mathcal{N}(0, I_q). \tag{2.23}
\]

From (2.11) we have

\[
G_n^{-1/2}A_n M_n^{-1} \Psi(n)(\beta_0) \to \mathcal{N}(0, I_q). \tag{2.24}
\]

Then it follows from (2.19), consistency of \( \hat{\beta} \) and Lemma 4 that

\[
G_n^{-1/2}A_n(\hat{\beta} - \beta_0) = G_n^{-1/2}A_n M_n^{-1} \Psi(n)(\beta_0) + o_p(1) \to_d \mathcal{N}(0, I_q). \tag{2.25}
\]

It is immediate from Theorem 2 that \( \frac{1}{n}\| \hat{G}_n - G_n \| = o_p(1) \) and that the eigenvalues of \( \hat{M}_n \) are uniformly bounded away from 0 and \( \infty \) with probability tending to one. Therefore

\[
\| \hat{G}_n^{-1/2}G_n^{1/2} - I_q \| = o_p(1). \tag{2.26}
\]

Combining (2.24), (2.25) and Slutsky’s theorem yields (2.23). □

**Proof of Theorem 4**

In this proof we integrate the arguments provided for Theorem 1 to the ideas of Theorem 3 in Zhang et al. [2014a].

For \( A_n \) in (2.7), there is a \((p-q) \times p\) matrix \( B_n \) such that \( B_n B_n^T = I_{p-q} \) and \( A_n B_n^T = 0 \). Therefore \( A_n^\alpha = A_0^\alpha \) is equivalent to \( \beta = B_n^T \gamma + c_0 \), where \( \gamma \) is a \((p-q) \times 1\) vector and \( c_0 = A_n^T G_n^{-1} A_0 \). Then minimizing \( \rho_n(\beta) \) subject to \( A_n^\alpha = A_0^\alpha \) is equivalent to minimizing \( \rho_n(B_n^T \gamma + c_0) \) with respect to \( \gamma \). Denote by \( \hat{\gamma} \) one such minimizer. By Lemma 5 \( \hat{\rho}_n \) is asymptotically concave. Thus \( \beta \) is the unique minimizer of \( \rho_n(\beta) \) and \( \Lambda_n = 2n(\rho_n(B_n^T \hat{\gamma} + c_0) - \rho_n(\hat{\beta})) \) for large \( n \).

By similar arguments to those used in step 3 of the proof of Theorem 1 we have

\[
-B_n M_n^*(B_n^T \hat{\gamma} + c_0)B_n^T(\hat{\gamma} - \gamma_0) = B_n \Psi(n)(B_n^T \gamma_0 + c_0).
\]

and

\[
(\hat{\gamma} - \gamma_0) = -(B_n M_n(B_n^T \gamma_0 + c_0)B_n^T)^{-1}B_n \Psi(n)(B_n^T \gamma_0 + c_0) + o_p(1/\sqrt{n}). \tag{2.27}
\]

Define \( \hat{P}_n = I_p - P_{M_n}^{-1/2}B_n^T = P_{M_n^{-1/2}A_n^T} \). Then \( M_n^{-1} - B_n^T(B_n M_n B_n^T)^{-1}B_n = M_n^{-1/2} \hat{P}_n M_n^{-1/2} \). Furthermore (2.26) and (2.20) yield

\[
B_n^T \hat{\gamma} + c_0 - \hat{\beta} = B_n^T(\hat{\gamma} - \gamma_0) - (\hat{\beta} - \beta_0)
\]

\[
= -M_n^{-1/2} \hat{P}_n M_n^{-1/2} \Psi(n)(\beta_0) + o_p(1/\sqrt{n}). \tag{2.28}
\]
Hence (2.26)-(2.28) and Lemma 3 lead to
\[ \Psi_n(B_n^T \hat{\gamma} + c_0) - \Psi_n(\hat{\beta}) = \frac{1}{2}(B_n^T \hat{\gamma} + c_0 - \hat{\beta})^T \frac{1}{n} \hat{\Psi}_n(\beta^*)(B_n^T \hat{\gamma} + c_0 - \hat{\beta}), \] (2.28)
where \( \beta^* \) lies between \( B_n^T \hat{\gamma} + c_0 \) and \( \hat{\beta} \). By consistency of \( \hat{\beta} \) and \( \hat{\gamma} \), we have \( \beta^* \to_p \beta_0 \). Hence (2.26)-(2.28) and Lemma 3 lead to
\[ A_n = \left( M_n^{-1/2} \tilde{P}_n M_n^{-1/2} \Psi_n(\beta_0) \right)^T M_n(\beta_0) \left( M_n^{-1/2} \tilde{P}_n M_n^{-1/2} \Psi_n(\beta_0) \right) + o_p(1) \]
\[ = \left( \Psi_n(\beta_0) \right)^T M_n^{-1/2} \tilde{P}_n M_n^{-1/2} \left( \Psi_n(\beta_0) \right) + o_p(1). \]
Because \( \tilde{P}_n \) is idempotent it can be written as \( \tilde{P}_n = C_n^T C_n \), where \( C_n \) is such that \( C_n C_n^T = I_q \). Hence
\[ A_n = \left( C_n M_n^{-1/2} \Psi_n(\beta_0) \right)^T \left( C_n M_n^{-1/2} \Psi_n(\beta_0) \right) + o_p(1). \]
This completes the proof since by (2.11) \( \Psi_n(\beta_0) \sim N(0, Q_n) \).

**Proof of Theorem 5**

Following the same arguments given for Theorem 3 we obtain
\[ (A_n \hat{M}_n^{-1} \hat{\Phi}_n \hat{M}_n^{-1} A_n^T)^{-1/2} (A_n \hat{\beta} - \alpha_0) \to_d N(G^{-1/2} d, I_q). \]
This completes the proof. \( \square \)

### 2.5.2 Huberized residuals

Theorem 7 establishes the existence and asymptotic normality of the robust estimator given by the solution of (2.2) when \( \psi(r) = \psi_c(r) \), i.e. the Huber function defined by
\[ \psi_c(r) = \begin{cases} r & \text{if } |r| \leq c \\ c \cdot \text{sign}(r) & \text{otherwise.} \end{cases} \]
Before showing the theorem we require two auxiliary lemmas. Lemma 6 shows that \( \Psi_n(\beta) \) can be approximated by a linear function. Lemma 7 can be seen as an adaptation of Lemma 4 for this setting.

**Lemma 6.** Let \( r_{i0} = r_i(\beta_0) \), \( P(r_{i0}) = \pm c \) \( \to 0 \) and suppose condition (A2) holds. Then if \( \| \hat{\beta} - \beta_0 \| = O_p(\sqrt{p/n}) \) and \( p/n \to 0 \) as \( n \to \infty \), we have
\[ \Psi_n(\hat{\beta}) = \Psi_n(\beta_0) + \hat{M}_n(\hat{\beta} - \beta_0) + o_p(\sqrt{n}), \]
where
\[ \hat{M}_n = \hat{M}_n(\hat{\beta}, \beta', \beta'', \beta''') = \sum_{i=1}^n \left[ \psi_c(r_{i0}) \frac{\partial}{\partial \beta^T} \left( \frac{1}{\sqrt{v_i^T \beta}} \frac{\partial \mu_i}{\partial \beta} \right)_{\beta=\beta'} + \psi_c(r_{i0}) \frac{\partial \mu_i}{\sqrt{v_i^T \beta}} \frac{\partial^2 \mu_i}{\partial \beta^T \partial \beta} \right] \]
\( \beta', \beta'' \) and \( \beta''' \) lie between \( \hat{\beta} \) and \( \beta_0 \), and \( \psi_c(r) = 1 \{ |r| < c \} \).

**Lemma 7.** Under the conditions of Lemma 6 and \( p^2/n \to 0 \) as \( n \to \infty \), we have
where $H$.

Hence writing $h = \hat{r}_i - r_{i0}$, for $n$ large enough we have

$$\Psi^{(n)}(\hat{\beta}) - \Psi^{(n)}(\beta_0) = \sum_{i=1}^{n} \left[ \left( \psi_c(\hat{r}_i) - \psi_c(r_{i0}) \right) w(x_i) \frac{1}{\sqrt{v_{i0}}} \frac{\partial \mu_{i0}}{\partial \beta} + \psi_c(\hat{r}_i) w(x_i) \frac{1}{\sqrt{v_{i0}}} \frac{\partial \mu_{i0}}{\partial \beta} \right]$$

Proof of Lemma 6: Note that for $h \in \mathbb{R}$ small enough, Huber’s $\psi_c(\cdot)$ satisfies

$$\psi_c(r + h) = \psi_c(r) + h(1 \{ |r| < c \} + 1 \{ r = c, h \geq 0 \} + 1 \{ r = -c, h < 0 \}).$$

Hence writing $h = \hat{r}_i - r_{i0}$, for $n$ large enough we have

$$\Psi^{(n)}(\hat{\beta}) - \Psi^{(n)}(\beta_0) = \sum_{i=1}^{n} \left[ \left( \psi_c(\hat{r}_i) - \psi_c(r_{i0}) \right) w(x_i) \frac{1}{\sqrt{v_{i0}}} \frac{\partial \mu_{i0}}{\partial \beta} + \psi_c(\hat{r}_i) w(x_i) \frac{1}{\sqrt{v_{i0}}} \frac{\partial \mu_{i0}}{\partial \beta} \right] + T$$

where $\beta'$ and $\beta''$ lie between $\hat{\beta}$ and $\beta_0$, and $T = \sum_{i=1}^{n} \frac{\partial \mu_{i0}}{\partial \beta} \frac{\partial r_i}{\partial \beta'} \frac{\partial \mu_{i0}}{\partial \beta'} |_{\beta = \beta''} + \frac{\partial \mu_{i0}}{\partial \beta'} |_{\beta = \beta''}$.

The proof is completed by verifying that $T = o_p(\sqrt{n})$. Consider the $j$th component of $z^{(n)} = \sum_{i=1}^{n} z_i$ denoted by $z_j^{(n)}$. For any $j = 1, \ldots, p$ and $\epsilon > 0$, Markov’s inequality gives

$$P(|z_j^{(n)}| > \epsilon \sqrt{n}) \leq E[|z_j^{(n)}|] \epsilon \sqrt{n} \leq K \sum_{i=1}^{n} |x_{ij}| E[|\hat{r}_i - r_{i0}|(1 \{ r_{i0} = -c, h < 0 \} + 1 \{ r_{i0} = -c, h < 0 \})] \frac{\epsilon \sqrt{n}}{\epsilon \sqrt{n}} \leq K \sum_{i=1}^{n} |x_{ij}| E[|\hat{r}_i - r_{i0}|1 \{ r_{i0} = \pm c \}] \frac{\epsilon \sqrt{n}}{\epsilon \sqrt{n}} \leq K \max_{1 \leq i \leq n} P(r_{i0} = \pm c) \sum_{i=1}^{n} |x_{ij}| \sqrt{n} = o(1)$$

where $K = \max_{1 \leq i \leq n} \left\{ w(x_i) \frac{1}{\sqrt{v_{i0}}} \frac{\partial \mu_{i0}}{\partial \beta} \right\}$. □

Proof of Lemma 7: We will show that for $\beta_1, \beta_2, \beta_3, \beta_4 \in N_n(\delta)$ we have

$$a^T M_n^{-1/2} \tilde{M}_n(\beta_1, \beta_2, \beta_3) M_n^{-1/2} a - 1 \to_p 0.$$
Since $P(r_{i0} = \pm c) \to 0$, the same arguments given in Lemma 3 yield

$$\frac{1}{n} a^T \left( \hat{M}_n(\beta_1, \beta_2, \beta_3, \beta_4) - M_n \right) a = \frac{1}{n} a^T \left( E[\hat{M}_n(\beta_1, \beta_2, \beta_3, \beta_4)] - M_n \right) a + o_p(1)$$

$$= T + o_p(1). \quad (2.30)$$

Using $P(r_{i0} = \pm c) \to 0$ and (2.10), we can write

$$T = \frac{1}{n} a^T \sum_{i=1}^{n} x_i x_i^T \left( \varphi_i(\beta_1, \beta_2, \beta_3, \beta_4) - \varphi_i(\beta_0, \beta_0, \beta_0, \beta_0) \right) a$$

$$\leq \frac{1}{n} \max_{1 \leq i \leq n} |x_i^T (\beta - \beta_0)| a^T \sum_{i=1}^{n} x_i x_i^T a$$

$$\leq O(p/\sqrt{n}),$$

where the first inequality is obtained by the boundedness $\psi_c(\cdot)$ and the continuity of $\hat{M}_n$ outside a set of zero measure for $n$ large enough. As $M_n = O(n)$, it follows that

$$a^T M_n^{-1/2} \hat{M}_n M_n^{-1/2} a - 1 = a^T M_n^{-1/2} (\hat{M}_n - M_n) M_n^{-1/2} a = o_p(1)$$

which proves part (a). Part (b) follows by the same arguments used in the proof of Lemma 4. □

**Theorem 7.** Let $r_{i0} = r_i(\beta_0)$, $P(r_{i0} = \pm c) \to 0$ and suppose condition (A2) holds. Then if $p^2/n \to 0$ as $n$ goes to infinity, there exists a sequence of Huberized estimators $\hat{\beta}$ such that

$$P\left( \Psi^{(n)}(\hat{\beta}) = 0 \right) \to 1 \quad (2.31)$$

and

$$a^T V_n^{-1/2}(\hat{\beta} - \beta_0) \to_d N(0, 1) \quad (2.32)$$

**Proof:** The proof follows by essentially the same three steps provided in the proof of Theorem 1. Step 1 is immediate. In Step 2 the only change needed in the proof is to replace Lemma 4 by Lemma 7, and $M_n(\beta)$ by $\hat{M}_n(\beta)$. We now prove Step 3. Since Step 2 established $\Psi^{(n)}(\hat{\beta}) = 0$ with probability tending to one, Lemmas 6 and 7 imply that for large enough $n$

$$a^T V_n^{-1/2}(\hat{\beta} - \beta_0) = -a^T V_n^{-1/2} M_n^{-1} \Psi^{(n)}(\beta_0) + o_p(1)$$

$$= -a^T Q_n^{-1/2} M_n \hat{M}_n^{-1} \Psi^{(n)}(\beta_0) + o_p(1)$$

$$= -a^T Q_n^{-1/2} \Psi^{(n)}(\beta_0) + o_p(1). \quad (2.33)$$

□
Chapter 3

Robust penalized M-estimators for high dimensional generalized linear models

General remarks

Generalized linear models are popular statistical methods for modeling continuous and discrete data parametrically. In this general framework we consider the problem of variable selection through penalized methods by focusing on resistance issues in the presence of outlying data and other deviations from the stochastic assumptions. We highlight the weaknesses of widely used penalized M-estimators and advocate the need for robust estimators. In particular we propose a robust penalized quasilikelihood estimator and show that it enjoys the oracle properties in a high dimensional set up and is stable in a neighborhood of the model. Finally, we illustrate the finite sample performance of the estimator on simulated and real data.

3.1 Introduction

Penalized methods have gained popularity over the last years and have proved to be a good alternative to the traditional approaches for variable selection, in particular in high dimensional problems. By providing estimation and variable selection simultaneously, they overcome the increasingly high computational cost of variable selection when the number of covariates is large and they reduce its variability. Since their introduction for the linear model, (Breiman [1995], Tibshirani [1996]) many extensions of lasso-type procedures have been proposed; cf. for instance Efron et al. [2004], Zou and Hastie [2005], Yuan and Lin [2006], and the retrospective article by Tibshirani [2011]. Their asymptotic properties have been studied for the case when the number of parameters is fixed by Knight and Fu [2000], Fan and Li [2001], Zou [2006]. A large body of literature is available for the high dimensional case, where the number of parameters is allowed to grow as the sample size increases at an even faster rate; see an overview in Bühlmann and Van De Geer [2011] and references thereof. These results have provided strong theoretical arguments in favor of these procedures.

As for any statistical procedure, these approaches to variable selection rely on stochastic assumptions which may not be satisfied on real data. Moreover, they are typically
affected by the presence of a few outlying observations. As an illustration consider for instance the parametric Poisson regression setting of Section 5.1 where the true underlying generating process is not the assumed model $P(\mu_i)$ but the slightly perturbed $(1-\epsilon)P(\mu_i) + \epsilon P(\nu \mu_i)$. The latter represents a situation where the distribution of the data lies in a small neighborhood of the model that can produce for instance overdispersion. When say $\epsilon = 0.05$ and $\nu = 5, 10$, the performance of standard penalized methods such as the lasso deteriorates drastically.

Robust statistics (Huber [1981] and 2nd edition by Huber and Ronchetti [2009], Hampel et al. [1986], Maronna et al. [2006]) provides a theoretical framework that allows to take into account that models are only idealized approximations of reality and develops methods that still give reliable results when slight deviations from the stochastic assumptions on the model occur. Following this approach, many authors have suggested sparse estimators that limit the impact of contamination in the data (e.g. Sardy et al. [2001], Wang et al. [2007a], Li et al. [2011], Lozano and Meinshausen [2013], and Fan et al. [2014a] among others). These procedures rely on the intuition that a loss function that defines robust estimates in the well understood unpenalized fixed dimensional M-estimation set up, should also define a robust estimator when it is penalized by a deterministic function. In the linear model for instance, Fan et al. [2014a] show that under very mild conditions on the error term, their estimator satisfies the oracle properties. Wang et al. [2013] and Alfons et al. [2013] studied the finite sample breakdown of their proposals for the linear model. The derivation of the influence function has been explored in Wang et al. [2013] and Öllerer et al. [2015]. In Chapter 4 we provide a rigorous treatment of this question in the general framework of penalized M-estimators.

One of the main contributions of our work is to propose a class of robust penalized estimators for high dimensional generalized linear models. We show that our estimators are consistent for variable selection, asymptotically normally distributed and can behave as well as a robust oracle estimator under regularity conditions. A second main contribution of this article is to characterize the robustness of penalized M-estimators by showing that they have a bounded bias in a small contamination neighborhood of the model. This is a characterization in the spirit of infinitesimal robustness of Hampel et al. [1986], but unlike the usual approach we do not establish our results based on the form of the influence function. From a technical standpoint, the conditions required for the derivations of our results in high dimensions are similar to the conditions required to obtain higher order robustness as in La Vecchia et al. [2012].

The rest of the paper is organized as follows. In Section 2 we present general results on the robustness of penalized M-estimators in a fixed parameters set up. We discuss the properties of the oracle estimators and show that penalized robust estimators have a bounded bias in a contamination neighborhood. In Section 3 we introduce our robust estimator for generalized linear models and provide oracle properties results including a non-asymptotic analysis in a small neighborhood of the model. In Section 4 we describe an algorithm for the implementation of our method and the criterion used to choose the regularization parameter. Finally, Section 5 concludes with numerical illustrations on both simulated and real data. All the proofs and conditions are given in the Appendix.
3.2 Penalized M-estimators

3.2.1 Background

Consider a sample of \( n \) observations \( Z^{(n)} = \{z_1, \ldots, z_n\} \) drawn from a common distribution \( F \) over the space \( Z \) and a loss function \( \rho_n : \mathbb{R}^d \times Z \rightarrow \mathbb{R} \). The quantity \( \rho_n(\beta; Z^{(n)}) = \frac{1}{n} \sum_{i=1}^{n} \rho(\beta; z_i) \) serves as a measure of fit between a parameter vector \( \beta \in \mathbb{R}^d \) and the observed data. This empirical function can be seen as an estimator of the unknown population risk function \( E_F[\rho_n(\beta; Z^{(n)})] \). We study the estimators resulting from the minimization of the regularized risk

\[
\Lambda_n(\beta) = \rho_n(\beta; Z^{(n)}) + \sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|)
\]

with respect to \( \beta \), where \( p_{\lambda}(\cdot) \) is a continuous penalty function with regularization parameter \( \lambda_n \). We suppose that the true underlying parameter vector of interest is sparse and without loss of generality we write \( \beta_0 = (\beta_1^T, \beta_2^T)^T \), where \( \beta_1 \in \mathbb{R}^k \), \( \beta_2 = 0 \in \mathbb{R}^{d-k} \), \( k < n \) and \( k < d \).

3.2.2 The oracle estimator and its robustness properties

The oracle estimator plays a key role in the theoretical analysis of many penalized estimators. It is the ideal unpenalized estimator we would use if we knew beforehand the support \( \mathcal{A} \) of the true parameter \( \beta_0 \). A typical oracle properties statement would say that the penalized likelihood estimator behaves asymptotically as well as the oracle maximum likelihood estimator in the sense that with probability 1 it sets to zero all the parameters in \( \mathcal{A}^c \), whereas those in \( \mathcal{A} \) are estimated consistently and are asymptotically normal; see for instance Fan and Li [2001] and Fan et al. [2014c].

However, if we consider the behavior of an ideal estimator not only at the model but in a neighborhood of the model, this characterization is not satisfactory. This follows from well established results in robust statistics that state that even such an oracle is in general unstable in a neighborhood of the model. To stress this point, let us consider a setup where the cardinality of \( \mathcal{A} \) is a fixed number \( k \) and let us first recall some facts. An M-estimator \( T(F_n) \) of \( \beta \) (Huber [1964], Huber and Ronchetti [2009]), where \( F_n \) denotes the empirical distribution function, is defined as a solution to

\[
\sum_{i=1}^{n} \Psi\{z_i, T(F_n)\} = 0,
\]

where \( z_1, \ldots, z_n \) are \( i.i.d. \) realizations of a general parametric model \( F \). Note that this class of estimators is a strict generalization of the class of regular maximum likelihood estimators. If a statistical functional \( T(F) \) is sufficiently regular, a von Mises expansion (von Mises [1947]) yields

\[
T(G) \approx T(F) + \int IF(z; F, T)d(G - F)(z),
\]

where \( IF(z; F, T) \) denotes the influence function of the functional \( T \) at the distribution \( F \); see Hampel [1974] and Hampel et al. [1986]. Considering the approximation (3.2) over an \( \epsilon \) neighborhood of the model \( \mathcal{F}_\epsilon = \{F(\epsilon) | F(\epsilon) = (1-\epsilon)F + \epsilon G, G \text{ an arbitrary distribution}\} \), we see that the influence function can be used to linearize the asymptotic bias in a
neighborhood of the ideal model. Therefore, a bounded influence function implies a bounded approximate bias and in that sense an M-estimator characterized by a bounded score equation implies infinitesimal robustness. Throughout the paper we call such M-estimators and their penalized counterparts robust. It is well known that likelihood based estimators are in general not robust in this sense. It follows that we could expect a penalized M-estimator based on a loss function with a bounded derivative to behave better in a neighborhood of the model than the classical oracle estimator, i.e. a robust penalized M-estimator could 'beat' the oracle estimator under contamination. An illustration of this phenomenon is given in Section 5.1. Clearly an appropriate benchmark estimator under contamination will only be given by a robust estimator that remains stable in $F_\epsilon$.

### 3.2.3 Some fixed parameter asymptotics

We now provide an asymptotic analysis of estimators obtained as the solution to problem (3.1) when the dimensionality of the parameter is fixed and the sample size $n$ goes to infinity. In particular we study the asymptotic behavior of robust penalized M-estimators under the $\epsilon$-contamination model $F_\epsilon$.

Theorem 1 shows that for a given tuning parameter under some regularity conditions, the robust penalized M-estimators have a bounded asymptotic bias in an $\epsilon$-contamination neighborhood. Theorem 2 states oracle properties results for general penalized M-estimators in a shrinking $\epsilon$ neighborhood. We use the notation $\Psi_i(\beta) = \partial \rho(z_i; \beta)/\partial \beta$, $M(\beta) = \partial E_{F}\{\Psi_i(\beta)\}/\partial \beta$ and $Q(\beta) = E_{F}\{\Psi_i(\beta)\Psi_i^T(\beta)\}$. We also let $M_{11}$ and $Q_{11}$ denote partitions of $M(\beta_0)$ and $Q(\beta_0)$ corresponding to derivatives with respect to $\beta_1$.

**Theorem 1.** Under conditions A, B in Appendix 3.6 and any $\lambda > 0$, the asymptotic bias of the penalized M-estimator is of order $O(\epsilon)$ in an $\epsilon$ contamination neighborhood of the model.

**Theorem 2.** Let $\epsilon = o(\lambda_n)$ with $\lambda_n \sqrt{n} \to 0$ and $\lambda_n n \to \infty$. Further, let $a$ be a $k$ dimensional vector with $\|a\|_2 = 1$. Then if the data is generated under the contamination model $F_\epsilon$ and condition A holds, there is a minimizer of (3.1) satisfying the following oracle properties as $n \to \infty$

1. Sparsity : $\hat{\beta}_2 = 0$
2. Asymptotic normality : $\sqrt{n}a^T M_{11} Q_{11}^{-1/2}(\hat{\beta}_1 - \beta_1) \to_d N(0, 1)$.

### 3.3 Robust generalized linear models

#### 3.3.1 A robust M-estimator

Generalized linear models (McCullagh and Nelder [1989]) provide a powerful tool that includes standard linear models as a particular case and allow to model both discrete and continuous responses, belonging to the exponential family. More precisely it is assumed that the response variables $Y_i$ for $i = 1, \ldots, n$ are drawn from independent distributions of the type

$$f(y_i; \theta_i) = \exp\left[\{y_i\theta_i - b(\theta_i)\}/\phi + c(y_i, \phi)\right],$$
where \( a(\cdot), b(\cdot) \) and \( c(\cdot) \) are some specific functions and \( \phi \) a nuisance parameter. Thus 
\[
E(Y_i) = b'(\theta_i) \text{ and } Var(Y_i) = \phi b''(\theta_i).
\]
Furthermore it is assumed that 
\[
E(Y_i | x_i) = \mu_i \text{ and } Var(Y_i | x_i) = v_i = v(\mu_i).
\]
Generalized linear models are then defined by \( g(\mu_i) = \eta_i = x_i^T \beta_0 \), where \( \beta_0 \in \mathbb{R}^d \) is the vector of parameters, \( x_i \in \mathbb{R}^d \) is the set of explanatory variables and \( g(\cdot) \) the link function.

Several robust M-estimates have been proposed for generalized linear models. Here we focus on the class of estimates proposed by Cantoni and Ronchetti [2001] because they can be viewed as a natural robustification of the quasilikelihood estimators of Wedderburn [1974]. The robust quasilikelihood is defined by
\[
\rho_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} Q_M(y_i, x_i^T \beta),
\]
where the functions \( Q_M(y_i, x_i^T \beta) \) can be written as
\[
Q_M(y_i, x_i^T \beta) = \int_{-\infty}^\infty \psi(r_i) \frac{1}{\sqrt{v_i}} w(x_i) dt - \frac{1}{n} \sum_{j=1}^{n} \int_{-\infty}^\infty E\{\psi(r_j) \frac{1}{\sqrt{v_j}}\} w(x_j) dt
\]
with \( r_i = (y_i - \mu_i)/\sqrt{v_i}, \tilde{s} \) such that \( \psi((y_i - \tilde{s})/\sqrt{v(\tilde{s}))} = 0 \) and \( \tilde{t} \) such that \( E[\psi((y_i - \tilde{s})/\sqrt{v(\tilde{s}))} = 0. \psi(\cdot) \) is a bounded function that protects against large outlying points in the space of the response variables and \( w(\cdot) \) downweights leverage points. The estimator of \( \hat{\beta} \) of \( \beta_0 \), derived from the minimization of this loss function, is the solution of the estimating equations
\[
\Psi^{(n)}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \Psi(y_i, x_i^T \beta) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \psi(r_i) \frac{1}{\sqrt{v_i}} w(x_i) \frac{\partial \mu_i}{\partial \beta} - a(\beta) \right\} = 0,
\]
where \( a(\beta) = \frac{1}{n} \sum_{i=1}^{n} E\{\psi(r_i)\} \frac{1}{\sqrt{v_i}} w(x_i) \frac{\partial \mu_i}{\partial \beta} \) is a constant ensuring Fisher consistency.

### 3.3.2 Asymptotic analysis in high dimensions

We consider estimates which minimize (3.1) where \( \rho_n(\beta) \) is as in (3.3). Theorems 1 and 2 of Chapter 2 apply in this case. Here we want to go beyond these results and present an asymptotic analysis in a high dimensional set up where the number of parameters \( d \) can be larger than the sample size. For the choice of the penalty function \( p_{\lambda_n}(\cdot) \) we propose to use the \( \ell_1 \)-norm in a first stage. We then use the resulting lasso estimates for the construction of their corresponding adaptive lasso estimators. Note that this choice of initial estimators have been proven to yield variable selection consistency in the linear model in Fan et al. [2014a]. Given the initial lasso estimates \( \hat{\beta} \), we then define the weights of the adaptive lasso estimator as
\[
\hat{w}_j = \begin{cases} 
1/|\hat{\beta}_j|, & \text{if } |\hat{\beta}| > 0 \\
\infty, & \text{if } |\hat{\beta}| = 0
\end{cases}
\]
for \( j = 1, \ldots, d \), where we define \( 0 \cdot \infty = 0 \). Hence the variables that are shrunk to zero by the initial estimator are not included in the adaptive lasso minimization problem. The robust adaptive lasso minimizes the penalized robust quasilikelihood
\[
\rho_n(\beta) + \lambda_n \sum_{j=1}^{d} \hat{w}_j |\beta_j|.
\]
The following theorem shows that given an initial consistent estimate, the robust adaptive lasso enjoys the oracle properties when the non-sparsity is such that \( k \ll n, \log d = O(n^\alpha) \) for some \( \alpha \in (0, 1/2) \) and half of the minimum signal \( s_n = 2^{-1}\min\{|\beta_{0j}| : \beta_{0j} \neq 0\} \) is much larger than the tuning parameter \( \lambda_n \).

**Theorem 3.** Assume condition C and let \( \hat{\beta} \) be a consistent initial estimator with rate \( r_n \) in \( \ell_2 \)-norm defining weights (3.5) for \( j = 1, \ldots, d \). Let the dimensionality of the problem satisfy \( \log d = O(n^\alpha) \) for \( \alpha \in (0, 1/2) \). Further let the non-sparsity dimensionality be of order \( k = o(n^{1/3}) \) and assume the minimum signal is such that \( s_n \gg \lambda_n \) with \( \lambda_n \sqrt{n}k \to 0 \) and \( \lambda_n r_n \gg \max\{\sqrt{k}/n, n^{-\alpha}\sqrt{\log n}\} \). Finally, let \( a \) be a \( k \) dimensional vector with \( \|a\|_2 = 1 \). Then, with probability tending to 1 as \( n \to \infty \), the robust adaptive lasso estimator \( \hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^T \) satisfies the following oracle properties:

1. **Sparsity:** \( \hat{\beta}_2 = 0 \).
2. **Asymptotic normality:** \( \sqrt{n}a^TM_{11}Q_{11}^{-1/2}(\hat{\beta}_1 - \beta_1) \to_d N(0, 1) \).

It is clear that the conditions \( \lambda_n \gg \sqrt{k}/n \) and \( \lambda_n \sqrt{n}k \to 0 \) cannot be met simultaneously. This implies that the robust lasso cannot achieve the oracle properties. This is in line with well established results obtained already in the classical case with fixed \( d \) by Fan and Li [2001] and Zou [2006]. Fan and Lv [2011] pointed this out for the penalized likelihood estimator when \( \log d = O(n^\alpha) \) in the GLM framework.

The next theorem shows that the robust lasso is indeed consistent. Combined with the previous theorem, it implies that using the robust lasso as initial estimator for its adaptive counterpart yields an estimator that satisfies the oracle properties.

**Theorem 4.** Denote by \( \hat{\beta} \) the robust lasso obtained with \( \lambda = O(\sqrt{n^{-1}\log d}) \). Then under condition C and a sufficiently large \( n \) we have

\[
\|\hat{\beta} - \beta\|_2 = O\left(\sqrt{\frac{k\log d}{n}}\right)
\]

with probability at least \( 1 - 2e^{-\gamma\kappa_n} \), where \( \gamma > 0 \) and \( \kappa_n = \max\{n/k^2, \log d\} \).

Since the robust quasi-likelihood function is not convex, the proof of this proposition relies on recent developments for penalized estimators with nonconvexity given in Loh and Wainwright [2015]. Note that Theorem 4 still holds if we replace the lasso penalty by a decomposable penalty as defined in Loh and Wainwright [2015]. As shown next, we can also build robust estimators satisfying the oracle properties combining the robust quasi-likelihood loss with nonconvex penalty functions such as the scad (Fan and Li [2001]) and mc+ (Zhang [2010]).

**Theorem 5.** Assume conditions B and C. Let the dimensionality of the problem satisfy \( \log d = O(n^\alpha) \) for \( \alpha \in (0, 1/2) \). Further let the non-sparsity dimensionality be of order \( k = o(n^{1/3}) \) and assume the minimum signal is such that \( s_n \gg \lambda_n \) with \( \lambda_n \sqrt{n}k \to 0 \) and \( \lambda_n \gg \max\{\sqrt{k}/n, n^{-\alpha}\sqrt{\log n}\} \). Let \( a \) be a \( k \) dimensional vector with \( \|a\|_2 = 1 \). Then, with probability tending to 1 as \( n \to \infty \), the robust adaptive lasso satisfies the following oracle properties:

1. **Sparsity:** \( \hat{\beta}_2 = 0 \).
2. **Asymptotic normality:** \( \sqrt{n}a^TM_{11}Q_{11}^{-1/2}(\hat{\beta}_1 - \beta_1) \to_d N(0, 1) \).
Theorems 3, 4 and 5 state essentially that our estimator has the same nice properties that the penalized likelihood estimators for GLM (Fan and Lv [2011]) have at the model. Since the oracle estimator defined by our estimator is robust in a neighborhood of the model, we expect the latter to be so too.

It can be seen from the proofs that the oracle properties will hold under $F(\epsilon)$ as long as $E_F(\psi_i(\beta_0)) = 0$. This implies consistency of the robust estimator over a broader class of distributions than the ones covered by the usual likelihood based counterparts. This happens because the boundedness of $\psi$ can ensure consistency for heavy tailed distributions. We therefore require milder conditions on the distribution of the responses that the ones considered in Fan and Lv [2011] where the existence of a moment generating function is necessary. Our conditions are also milder than the fourth moment assumption of van de Geer and Müller [2012].

To illustrate this point consider a Poisson regression framework where there is a fraction $\epsilon$ of contaminated responses coming from a heavy tailed overdispersed Poisson mixture. Specifically, let the contamination distribution $G$ be such that for $U_i \sim G$, $U_i$ takes values in $\mathbb{Z}_{\geq 0}$ for all $i = 1, \ldots, n$ and we have $E[U_i] = \mu_i$ and $P(|U_i| > q) = c_\kappa q^{-\kappa}$, where $c_\kappa$ is a constant depending only on $\kappa \in (0, 3/4)$. In this case the distribution of the responses does not have a moment generating function and it has in fact an infinite variance. Still, in this scenario our estimator satisfies the oracle properties because $E_{F(\epsilon)}(\psi_i(\beta_0)) = 0$.

### 3.3.3 Weak oracle properties in a contamination neighborhood

The previous asymptotic analysis can be easily generalized to a shrinking contamination neighborhood where $\epsilon \to 0$ when $n \to \infty$ as in Theorem 2, if we let $\epsilon = o(\lambda_n)$ in Theorem 5. Here in addition we provide a different theoretical result where the contamination neighborhood does not shrink, but instead produces a small non-asymptotic bias on the estimated nonzero coefficients. If this bias is not too large and the minimum signal is large enough, we could expect to obtain a correct support recovery and bounded bias. In this sense we could expect a robust estimator that behaves as well as a robust oracle. We state formally this result next.

**Theorem 6.** Assume conditions B and C'. Let the dimensionality of the problem satisfy $\log d = O(n^{1-2\alpha})$ for $\alpha \in (0, 1/2)$, with $k = o(n)$ non-zero parameters and minimum signal such that $s_n \geq n^{-\zeta} \log n$. Let $\lambda_n$ satisfy $p'_s(s_n) = o(n^{-\zeta} \log n)$ and $\lambda_n \gg n^{-\alpha} \log^2 n$. Then for $n$ large enough, there is a contamination neighborhood where the robust penalized quasilikelihood estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^T$ with probability at least $1 - 2\{kn^{-1} + (d - k)e^{-n^{-2\alpha} \log n}\}$, satisfies the (weak) oracle properties:

1. **Sparsity:** $\hat{\beta}_2 = 0$.
2. **$\ell_\infty$-norm:** $\|\hat{\beta}_1 - \beta_1\|_\infty = O(n^{-\zeta} \log n + \epsilon)$.

Theorem 6 can be viewed as an extension of Theorem 2 of Fan and Lv [2011] to a robust penalized M-estimator in a contamination neighborhood. Moreover, it is a statement in the spirit of the infinitesimal robustness approach to robust statistics. In this line of thought the impact of moderate distributional deviations from ideal models on a statistical procedure is assessed by approximating and bounding the resulting bias; see Hampel et al. [1986]. It can be seen from the proof of Theorem 6 that as long as $E_{F(\epsilon)}(\psi_i(\beta_0)) = 0$, we obtain $\|\hat{\beta}_1 - \beta_1\|_\infty = O(n^{-\zeta} \log n)$. 
Chapter 3. Robust penalized M-estimators for GLM

3.4 Implementation

3.4.1 Fisher scoring coordinate descent

Estimating the robust lasso for generalized linear models when $d > n$ requires an appropriate algorithm. We propose a coordinate descent type algorithm based on successive \textit{expected} quadratic approximation of the quasilikelihood about the current estimates. Specifically for a given value of the tuning parameter we successively solve via coordinate descent the weighted least squares problem given by

$$\|\sqrt{W}(z - X\beta)\|_2^2 + \lambda \|\beta\|_1 \tag{3.7}$$

where $W = \text{diag}(W_1, \ldots, W_n)$ is a weight matrix and $z = (z_1, \ldots, z_n)^T$ a vector of pseudo data with components

$$W_i = E\{\psi(r_i)r_i\}V(\mu_i)^{-1}w(x_i)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

and

$$z_i = \eta_i + \frac{\psi(r_i) - E\{\psi(r_i)\}}{E\{\psi(r_i)r_i\}}V(\mu_i)^{-1/2}\frac{\partial \eta_i}{\partial \mu_i}.$$ 

These are the robust counterparts of the usual expressions appearing in the well known iterative reweighted least squares algorithm; cf. Heritier et al. [2009], Appendix E.3. Our coordinate descent algorithm is therefore a sequence of nested loops:

1. Outer loop: decrease $\lambda$.
2. Middle loop: Update $W$ and $z$ in (3.7) using the current parameters $\hat{\beta}_\lambda$.
3. Inner loop: run the coordinate descent algorithm on the weighted least squares problem (3.7).

This algorithm differs from the one of Friedman et al. [2010] in the quadratic approximation step, where the expectation of the usual approximation is computed. This step is crucial for the robust lasso because it guarantees that $W$ has only positive components in Poisson and Binomial regression when using Huberized residuals. In other words, this guarantees the convergence of the inner loop. Note also that for classical penalized log likelihood regression the algorithms coincide. The initial value of the tuning parameter in our algorithm is $\lambda_0 = n^{-1}\|\sqrt{WX}^Tz\|_\infty$ with $W$ and $z$ computed with $\beta = 0$. It is such that 0 is the maximizer of the robust lasso. We then follow the three steps mentioned above using warm starts and active set cycling to speed up computations as discussed in Friedman et al. [2010].

3.4.2 Tuning parameter selection

We choose the tuning parameter $\lambda_n$ based on a robust Schwarz information criterion. Specifically we select the parameter $\lambda_n$ that minimizes

$$\text{BIC}(\lambda_n) = \rho_n(\hat{\beta}_{\lambda_n}) + \frac{\log n}{n}|\text{supp}\hat{\beta}_{\lambda_n}|,$$
where $|\text{supp}\hat{\beta}_n|$ denotes the cardinality of the support of $\hat{\beta}_n$. Note that we write $\hat{\beta}_n$ to stress the dependence of the minimizer of (3.1) on the tuning parameter. In an unpenalized set up, an information criterion of this form was considered by Machado [1993] who provided solid theoretical justification for it by proving model selection consistency and robustness. In the penalized regression literature Lambert-Lacroix and Zwald [2011] and Li et al. [2011] have also used a robust BIC to select the tuning parameter. The soundness of this choice relies on the work, among others, of Wang et al. [2007b], Wang et al. [2009] and Fan and Tang [2013]. Other popular choices for the selection of the tuning parameter include generalized information criteria (see for instance in Zhang et al. [2010] and Flynn et al. [2013]) and $K$-fold cross-validation.

### 3.5 Numerical examples

#### 3.5.1 A simulation study: large outliers in the responses

<table>
<thead>
<tr>
<th>Method</th>
<th>L2-loss</th>
<th>size</th>
<th>#FN</th>
<th>L2-loss</th>
<th>size</th>
<th>#FN</th>
<th>L2-loss</th>
<th>size</th>
<th>#FN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\epsilon = 0, \nu = 0$</td>
<td></td>
<td></td>
<td>$\epsilon = 0.05, \nu = 5$</td>
<td></td>
<td></td>
<td>$\epsilon = 0.05, \nu = 105$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lasso</td>
<td>0.473</td>
<td>6</td>
<td>0</td>
<td>1.313</td>
<td>18</td>
<td>0</td>
<td>1.926</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.165)</td>
<td></td>
<td></td>
<td>(2.965)</td>
<td></td>
<td></td>
<td>(0.564)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r-lasso</td>
<td>0.481</td>
<td>6</td>
<td>0</td>
<td>0.585</td>
<td>7</td>
<td>0</td>
<td>0.629</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.187)</td>
<td></td>
<td></td>
<td>(1.482)</td>
<td></td>
<td></td>
<td>(0.377)</td>
<td>(2.965)</td>
<td></td>
</tr>
<tr>
<td>adalasso</td>
<td>0.277</td>
<td>3</td>
<td>0</td>
<td>1.576</td>
<td>10</td>
<td>0</td>
<td>2.327</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0.171)</td>
<td></td>
<td></td>
<td>(0.375)</td>
<td>(2.965)</td>
<td></td>
<td>(0.669)</td>
<td>(2.965)</td>
<td>(1.482)</td>
</tr>
<tr>
<td>r-adalasso</td>
<td>0.305</td>
<td>3</td>
<td>0</td>
<td>0.404</td>
<td>3</td>
<td>0</td>
<td>0.439</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.167)</td>
<td></td>
<td></td>
<td>(0.299)</td>
<td>(0)</td>
<td></td>
<td>(0.348)</td>
<td>(1.482)</td>
<td>(0)</td>
</tr>
<tr>
<td>lasso</td>
<td>0.306</td>
<td>6</td>
<td>0</td>
<td>1.179</td>
<td>30</td>
<td>0</td>
<td>1.754</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.121)</td>
<td></td>
<td></td>
<td>(2.965)</td>
<td></td>
<td></td>
<td>(0.304)</td>
<td>(1.482)</td>
<td></td>
</tr>
<tr>
<td>r-lasso</td>
<td>0.312</td>
<td>5</td>
<td>0</td>
<td>0.379</td>
<td>7</td>
<td>0</td>
<td>0.394</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.136)</td>
<td></td>
<td></td>
<td>(1.482)</td>
<td></td>
<td></td>
<td>(0.183)</td>
<td>(2.965)</td>
<td></td>
</tr>
<tr>
<td>adalasso</td>
<td>0.149</td>
<td>3</td>
<td>0</td>
<td>1.348</td>
<td>15</td>
<td>0</td>
<td>2.004</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.087)</td>
<td></td>
<td></td>
<td>(0.294)</td>
<td>(4.448)</td>
<td></td>
<td>(0.379)</td>
<td>(2.965)</td>
<td></td>
</tr>
<tr>
<td>r-adalasso</td>
<td>0.171</td>
<td>3</td>
<td>0</td>
<td>0.244</td>
<td>3</td>
<td>0</td>
<td>0.234</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.109)</td>
<td></td>
<td></td>
<td>(0.146)</td>
<td>(0)</td>
<td></td>
<td>(0.129)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>lasso</td>
<td>0.194</td>
<td>5</td>
<td>0</td>
<td>0.936</td>
<td>41</td>
<td>0</td>
<td>1.549</td>
<td>41</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.065)</td>
<td></td>
<td></td>
<td>(2.965)</td>
<td></td>
<td></td>
<td>(0.208)</td>
<td>(1.482)</td>
<td></td>
</tr>
<tr>
<td>r-lasso</td>
<td>0.198</td>
<td>5</td>
<td>0</td>
<td>0.218</td>
<td>7.5</td>
<td>0</td>
<td>0.221</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.066)</td>
<td></td>
<td></td>
<td>(1.482)</td>
<td></td>
<td></td>
<td>(0.073)</td>
<td>(2.965)</td>
<td></td>
</tr>
<tr>
<td>adalasso</td>
<td>0.101</td>
<td>3</td>
<td>0</td>
<td>1.085</td>
<td>22</td>
<td>0</td>
<td>1.812</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.063)</td>
<td></td>
<td></td>
<td>(0.183)</td>
<td>(7.413)</td>
<td></td>
<td>(0.194)</td>
<td>(2.965)</td>
<td></td>
</tr>
<tr>
<td>r-adalasso</td>
<td>0.107</td>
<td>3</td>
<td>0</td>
<td>0.248</td>
<td>3</td>
<td>0</td>
<td>0.147</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.068)</td>
<td></td>
<td></td>
<td>(0.269)</td>
<td>(0)</td>
<td></td>
<td>(0.036)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
</tbody>
</table>

**Table 3.1:** Summary of the performance of the penalized estimators computed under different high dimensional Poisson regression scenarios. The median of each measure over 100 simulations is given with its MAD in parenthesis.

We explore the behavior of classical and robust versions of both the lasso and adaptive lasso in a generalized linear model. For the robust estimators we use the loss function given by (5.5) with $\psi(r) = \psi_c(r)$, the Huber function defined by

$$
\psi_c(r) = \begin{cases} 
  r & \text{if } |r| \leq c \\
  c \cdot \text{sign}(r) & \text{otherwise.}
\end{cases}
$$

(3.8)
Figure 3.1: The plots show the $L2$ error of the classical and robust versions of the lasso, adaptive lasso and oracle. The severity of the contamination $\nu = 1, 5, 10$ increases from left to right and the sample size and dimensionality of the problems $(n, d) = \{(50, 100), (100, 400), (200, 1600)\}$ from top to bottom.

Figure 3.2: We show a zoom of the plots appearing in the bottom corners of Figure 3.1, that is the $L2$ error of the classical and robust versions of the lasso, adaptive lasso and oracle when $(n, d) = (200, 1600)$ and $\nu = 1, 10$. 
3.5. Numerical examples

We consider the Poisson regression model with canonical link \( g(\mu_i) = \log \mu_i = x_i^T \beta \), where \( \beta = (1, 0, 0, 1, 0, 1, 5, 0, \cdots, 0)^T \) and the covariates \( x_{ij} \) were generated from standard uniforms with correlation \( \text{cor}(x_{ij}, x_{ik}) = \rho^{|j-k|} \) and \( \rho = 0.5 \) for \( j, k = 1, \ldots, d \). This set up is reminiscent of example 1 in Tibshirani [1996]. The response variables \( Y_i \) were generated according to the Poisson distribution \( P(\mu_i) \) and a perturbed distribution of the form \((1 - b)P(\mu_i) + bP(\nu \mu_i)\), where \( b \sim \text{Bin}(1, \epsilon) \). The latter represents a situation where the distribution of the data lies in a small neighborhood of the model that can produce for instance overdispersion. We set \( \epsilon = 1.5, \nu = 5, 10 \) and \( \epsilon = 0, 0.05 \). The sample sizes and dimensionality were respectively \( n = 50, 100, 200 \) and \( d = 100, 400, 1600 \). We implemented the Fisher scoring coordinate descent algorithm as discussed in 3.4.1 with a grid of values \( \lambda \) of length 100 decreasing on the log scale. We stopped the algorithm when the tuning parameter gave models of size greater or equal than 20, 30 and 40 respectively when \( n \) was 50, 100 and 200. The tuning parameters were selected by BIC and the simulation size was 100. We show the performance of the classical lasso (L) and adaptive lasso (AL), the robust lasso (RL) and adaptive robust lasso (RAL) as well as the classical oracle (O) and the robust roracle (RO) estimators. Table 3.1 summarizes the overall performance of the different procedures in terms euclidean error of the estimated coefficients, the size of the selected model and the number of false negatives. Figures 3.1 and 3.2 illustrate in more detail the estimation error of the different estimators while 3.3 illustrate the model selection properties.

![Graphs showing performance of different estimators](image)

**Figure 3.3:** The plots show the cardinality of the support of the classical and robust versions of the lasso and adaptive lasso. The severity of the contamination \( \nu = 1, 5, 10 \) increases from left to right and the sample size and dimensionality of the problems \( (n, d) = \{(50, 100), (100, 400), (200, 1600)\} \) from top to bottom.
It is clear that without contamination the classical procedures and their robust counterparts have a very similar performance. As expected from our theoretical results the estimation error of all the penalized estimators seems to converge to zero as the sample size increases despite of the fact that the dimensionality increases at an even faster rate. The picture changes drastically under contamination for the classical estimators as they give poor estimation errors and keep too many variables. On the other hand, the robust estimators remain stable under contamination. In particular the robust adaptive lasso performs almost as well as the robust oracle in terms of $L_2$-loss and is successful in recovering the true support when $n = 100, 200$ even under contamination. In this example the robust adaptive lasso outperforms even the classical oracle estimator under contamination. The poor $L_2$ error of the classical penalized estimators under contamination is likely to stem from the large number of noise variables they tend to select in this situation.

3.5.2 An illustration

We illustrate the use of our estimator in a Poisson regression set up with a dataset issued from a study of the diversity of marsupials in the Montane ash forest (Australia). The number of different species of arboreal marsupials (possum) was observed on 151 different 3ha sites with uniform vegetation. For each site, the following measures were recorded: number of shrubs, number of cut stumps from past logging operations, number of stags (hollow-bearing trees), a bark index reflecting the decorticating bark, a habitat score indicating the suitability of nesting and foraging habitat for Leadbeater’s possum, the basal area of acacia species, the species of eucalypt with the greatest stand basal area (Eucalyptus regnans, Eucalyptus delegatensis, Eucalyptus nitens), and the orientation of the site. The problem is to model the relationship between diversity and the other variables. The study is fully described by Lindenmayer et al. [1990]. Weisberg and Welsh [1993] concluded that a canonical link fits the data well and Cantoni and Ronchetti [2001] suggested to further perform a stepwise model selection procedure based on a likelihood ratio type test. To assess the performance of our method with a large number of variables we add 100 $U(-1,1)$ noise covariates to the original data in order to obtain sample sizes $n = 151$ with $d = 75, 500$ predictors. An analysis in the same spirit can be found in Section 6 of Schelldorfer et al. [2014]. When $d = 75$ we are still in a low dimensional set up and we can fit unpenalized Poisson regression estimators. Because of the large number of variables the picture is blurred: the variables that were significant in the original data set can now have high p-values and many noise variables can be significant. We do not report the details. For the penalized estimators, as in the previous subsection, we implemented the Fisher scoring coordinate descent algorithm described in 3.4.1 with a grid of values $\lambda$ of length 100 decreasing on the log scale. Table 3.2 shows the fits obtained with the unregularized maximum likelihood and the robust quasilikelihood as well as their respective adaptive lassos with tuning parameter selected by AIC and BIC. The classical and robust unpenalized estimators give similar fits which suggest that there are no influencing outliers in these data. This is line with the results reported in Cantoni and Ronchetti [2001]. We ran the classical and robust adaptive lassos on the enlarged noisy data. Their respective tuning parameters were selected by AIC and BIC. The presence of large number of noise variables did not seem to have much effect of the adaptive lasso fits and as they selected essentially the same model they select based on the original data. For the sake of space we only reported the results obtained using the robust adaptive lasso with tuning parameter chosen with BIC. It can be seen that the model selected in
the presence of noise covariates coincides with the one chosen without them except for the variable SW-NW.

<table>
<thead>
<tr>
<th>Method</th>
<th>ML</th>
<th>RQ</th>
<th>AL_{AIC}</th>
<th>AL_{BIC}</th>
<th>RAL_{AIC}</th>
<th>RAL_{BIC}</th>
<th>RAL_{BIC}, d=75</th>
<th>RAL_{BIC}, d=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shrubs</td>
<td>0.012</td>
<td>0.011</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Stumps</td>
<td>-0.272</td>
<td>-0.231</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Stags</td>
<td>0.040**</td>
<td>0.041**</td>
<td>0.047</td>
<td>0.040</td>
<td>0.048</td>
<td>0.039</td>
<td>0.039 (0.004)</td>
<td>0.035 (0.000)</td>
</tr>
<tr>
<td>Bark</td>
<td>0.040**</td>
<td>0.039**</td>
<td>0.045</td>
<td>0.030</td>
<td>0.047</td>
<td>0.039</td>
<td>0.042 (0.006)</td>
<td>0.037 (0.001)</td>
</tr>
<tr>
<td>Habitat</td>
<td>0.072†</td>
<td>0.068†</td>
<td>0.098</td>
<td>0.103</td>
<td>0.094</td>
<td>0.086</td>
<td>0.140 (0.002)</td>
<td>0.141 (0.000)</td>
</tr>
<tr>
<td>Acacia</td>
<td>0.018†</td>
<td>0.018†</td>
<td>0.013</td>
<td>0</td>
<td>0.016</td>
<td>0</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>E. nitens</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>- (-)</td>
<td>- (-)</td>
</tr>
<tr>
<td>E. regnans</td>
<td>-0.015</td>
<td>-0.016</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>E. delegatensis</td>
<td>0.115</td>
<td>0.131</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>NW-NE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>- (-)</td>
<td>- (-)</td>
</tr>
<tr>
<td>NW-SE</td>
<td>0.066</td>
<td>0.067</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>SE-SW</td>
<td>0.117</td>
<td>0.083</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>SW-NW</td>
<td>-0.489*</td>
<td>-0.523*</td>
<td>-0.646</td>
<td>-0.283</td>
<td>-0.718</td>
<td>-0.569</td>
<td>-0.424 (0.394)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>#extra</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

Table 3.2: Performance of robust adaptive lasso with the Possum data. A double asterisk ** means that a variable is significant at the 1% level. An asterisk * means that a variable is significant at the 5% level. A dagger † means that a variable is significant at the 10% level. For the augmented data sets the median of each estimated coefficient over 100 simulations is given with its median absolute deviation in parenthesis.

3.6 Appendix

3.6.1 Regularity conditions

**Condition A**

\( E_F\{\Psi_i(\beta_0)\} = 0 \), where \( \Psi_i(\beta) = \Psi(\beta; z_i) \) is bounded and continuous in a neighborhood \( \mathcal{O} \) of \( \beta_0 \) uniformly in \( z_i \). Furthermore \( \Psi_i(\beta) \) has two continuous derivatives with respect to \( \beta \) in \( \mathcal{O} \).

\( M(\beta) = E_F\left\{\frac{\partial \Psi_i(\beta)}{\partial \beta^T}\right\} < \infty, \quad Q(\beta) = E_F\left\{\Psi_i(\beta)\Psi_i^T(\beta)\right\} < \infty \)

and \( \frac{\partial \Psi_{ij}(\beta)}{\partial \beta_k} < \infty \), for \( 1 \leq (j,k) \leq d \),

where \( M(\beta) \) is positive definite.

Conditions (A1)-(A2) are fairly standard regularity conditions in the robust statistics literature.

**Condition B**

Let \( P(t; \lambda) = \lambda^{-1}p_\lambda(t) \) be increasing and concave in \( t \in [0, \infty) \) and have a continuous derivative such that \( P'(0+) = P'(0+; \lambda) > 0 \) is independent of \( \lambda \). In addition, \( P'(t; \lambda) \) is decreasing in \( \lambda \in (0, \infty) \).
Condition C

Let \( \{(x_i, y_i)\}_{i=1}^{n} \) denote independent pairs, each having the same distribution. We assume the following set of conditions:

(C1) The matrices \( M(\beta) \) and \( Q(\beta) \) satisfy

\[
0 < c_1 < \lambda_{\min}\left\{ M(\beta) \right\} \leq \lambda_{\max}\left\{ M(\beta) \right\} < c_2 < \infty \quad \text{and} \quad Q(\beta) < \infty,
\]

where \( \lambda_{\min} \) and \( \lambda_{\max} \) denote the smallest and largest eigenvalues.

(C2) \( |\psi'(r)| \) and \( |\psi'(r)r| \) are bounded. Furthermore there is a sufficiently large open set \( \mathcal{O} \) that contains the true parameter \( \beta_0 \) such that for all \( \beta \in \mathcal{O} \)

\[
\frac{\partial Q_M(y, \mu)}{\partial \eta} \leq A_1(x) \quad \text{and} \quad \frac{\partial^2 Q_M(y, \mu)}{\partial \eta^2} \leq A_2(x)
\]

where \( \mu = g^{-1}(x^T \beta) \) and \( A_1(x), A_2(x) < \infty \). Furthermore for all \( 1 \leq j, k \leq d \) we have \( |A_2(x)|_{x_j x_k} < \infty \).

(C1) is fairly standard in the robustness literature. (C2) is effectively requiring both \( \Psi \) and \( \Psi' \) to be bounded. Boundedness of \( \Psi' \) is a standard requirement for robust estimators. Recently La Vecchia et al. [2012] introduced the concept of higher order infinitesimal robustness which requires \( \Psi' \) to be bounded too. We refer the reader to this paper for an extensive discussion and several examples illustrating the benefits of higher order robustness. Obviously the condition \( |A_2(x)|_{x_j x_k} < \infty \) is not satisfied for classical estimators if the covariates are unbounded. However, for robust estimates \( \Psi(\cdot) \) is bounded and \( w(x) \) typically goes to zero as \( ||x||^{-1} \). Our condition requires a slightly stronger version i.e. \( w(x) \) goes to zero as \( \left( \max |x_j|^2 \right)^{-1} \). It is easy to see that (C2) implies (C3) given next.

It is used when dealing with the high dimensional setting and is similar to Condition 4 in Fan and Lv [2011].

(C3) For all \( \beta \in \mathcal{O} \), the design matrix \( X \) satisfies

\[
\text{tr}\left[ X_i^T \left\{ \Upsilon(X(\beta)) \right\}^2 X_i \right] = O(kn) \quad \text{(3.9)}
\]

\[
\left\| X_i^T \Upsilon'(X(\beta)) X_i \right\|_{2, \infty} \leq O(n) \quad \text{(3.10)}
\]

where \( x^{(j)} \) denotes the \( j \)th column of \( X \), \( N_0 = \{ \beta' \in \mathbb{R}^k : \|\beta' - \beta_1\|_{\infty} \leq s_n \} \), \( \|A\|_{2, \infty} = \max_{\|v\|_2=1} \|Av\|_{\infty} \), \( \Upsilon(X(\beta)) = \frac{\partial}{\partial \eta} \left\{ \sum_{i=1}^{n} Q_M(y, \mu_i) \right\} \) and \( \Upsilon'(\cdot), \Upsilon''(\cdot) \) are to be understood as derivatives with respect to \( \eta_i \) of its corresponding diagonal element.

Condition C’

Assume (C1) and (C2). Furthermore

(C3’) For all \( \beta \in \mathcal{O} \)

\[
\left\| X_i^T \Upsilon'(X(\beta)) X_i \left\{ X_i^T \Upsilon'(X(\beta)) X_i \right\}^{-1} \right\|_{\infty} \leq \min \left\{ c_3 \frac{P'(0+)}{P(s_n; \lambda_n; \alpha_1)}, P(0+) \right\},
\]

where \( c_3 \in (0, 1) \), \( \alpha_1 \in (0, 1/2) \), \( \Upsilon(X(\beta)) = \frac{\partial}{\partial \eta} \left\{ \sum_{i=1}^{n} Q_M(y, \mu_i) \right\} \) and \( \Upsilon'(\cdot) \) is to be understood as the derivative with respect to \( \eta_i \) of its corresponding diagonal element.
(C3') is similar to equation (16) of Condition 2 in Fan and Lv [2011]. Our condition seems to be stronger because we require the bound to hold for all $\beta \in \mathcal{O}$. However, when a folded concave penalty function is used, the upper bound of (C3') can grow to infinity at rate $O(n^{\alpha_1})$. This remark and the boundedness condition of (C2) suffice to obtain (C3').

### 3.6.2 Proof of main results

#### Proof of Theorem 1

For a fixed $\lambda$, let $\beta^{(e)}$ and $\beta^*$ denote the asymptotic values of the penalized M-estimator in a $\epsilon$ contamination neighborhood and at the model $F$ respectively. Note that for a fixed tuning parameter $\lambda$, the Karush-Kuhn-Tucker (KKT) conditions of the penalized estimator are

\[
\Psi_1^{(n)}(\hat{\beta}) + \bar{p}_\lambda(\hat{\beta}_1) = 0 \quad \text{and} \quad \|\Psi_2^{(n)}(\hat{\beta})\|_\infty < p'_\lambda(0+),
\]

(3.11) (3.12)

where $\Psi^{(n)}(\beta) = \{\Psi_1^{(n)}(\beta)^T, \Psi_2^{(n)}(\beta)^T\}^T = \frac{1}{n} \sum_{i=1}^{n} \Psi(\beta; Z_i)$ and $\bar{p}_\lambda(\beta') = \{p'_\lambda(\beta'_1), \ldots, p'_\lambda(\beta'_d)\}^T$ for $\beta' \in \mathbb{R}^d$. These conditions define the functionals

\[
\int \Psi_1(T(F))dF + \bar{p}_\lambda(T_1(F)) = 0 \quad \text{and} \quad \int \Psi_2(T(F))dF \|_\infty < p'_\lambda(0+).
\]

(3.13) (3.14)

Replacing $F$ by the empirical distribution function in (3.13) and (3.14) we recover the KKT conditions for a penalized M-estimator obtained from a sample of size $n$. Since the supports of $\beta^{(e)}$ and $\beta^*$ do not have to be the same, consider

\[
\mathcal{J}_{11} = \{j : \beta^{(e)}_j \neq 0, \beta^*_j \neq 0\}, \quad \mathcal{J}_{10} = \{j : \beta^{(e)}_j \neq 0, \beta^*_j = 0\}, \\
\mathcal{J}_{01} = \{j : \beta^{(e)}_j = 0, \beta^*_j \neq 0\}, \quad \mathcal{J}_{00} = \{j : \beta^{(e)}_j = 0, \beta^*_j = 0\},
\]

and let $d_{ij} = |\mathcal{J}_{ij}|$ and $\mathcal{J}_1 = \mathcal{J}_{10} \cup \mathcal{J}_{11}$. Clearly

\[
\|\beta^{(e)}_{j_{00}} - \beta^*_{j_{00}}\|_2 = 0 \text{ and } \|\beta^{(e)}_{j_{01}} - \beta^*_{j_{01}}\|_2 = \|\beta^*_{j_{01}}\|_2 \leq c < \infty.
\]

(3.15)

Note also that for $\tilde{\beta}$ lying between $\beta^{(e)}$ and $(\beta^*_{\mathcal{J}_1}, 0)$, and $\tilde{\beta}_1$ lying between $\beta^{(e)}_{\mathcal{J}_1}$ and $\beta^*_{\mathcal{J}_1}$, we have

\[
\int \Psi_{\mathcal{J}_1}(T(F^{(e)}))dF^{(e)} = \int \Psi_{\mathcal{J}_1}(\beta^*)dF^{(e)} + \int \Psi_{\mathcal{J}_1}(\tilde{\beta})dF^{(e)}(\beta^{(e)}_{\mathcal{J}_1} - \beta^*_{\mathcal{J}_1}) \\
= \epsilon \int \Psi_{\mathcal{J}_1}(\beta^*)d(G - F) + \int \Psi_{\mathcal{J}_1}(\beta^*)dF + \{M_{11}(\tilde{\beta}) + \epsilon \int \Psi_{\mathcal{J}_1}(\tilde{\beta})d(G - F)\}(\beta^{(e)}_{\mathcal{J}_1} - \beta^*_{\mathcal{J}_1}),
\]

(3.16)

and

\[
\bar{p}_\lambda(T_1(F_i)) = \bar{p}_\lambda(\beta^*_{\mathcal{J}_1}) + \bar{p}_\lambda(\tilde{\beta}_1)(\beta^{(e)}_{\mathcal{J}_1} - \beta^*_{\mathcal{J}_1}).
\]

(3.17)

Combining (3.13), (3.15), (3.16) and (3.17) we have

\[
\beta^{(e)}_{\mathcal{J}_1} - \beta^*_{\mathcal{J}_1} = S^{-1}\{eb_1 + \int \psi_{\mathcal{J}_1}(\beta^*)dF + \bar{p}_\lambda(\beta^*_{\mathcal{J}_1})\},
\]

(3.18)
where \( b_1 = \int \Psi_{\beta_1}(\beta^*)d(G - F) \) and \( S = -\{M_{11}(\hat{\beta}) + \epsilon \int \hat{\psi}_{\beta_1}(\beta^*)d(G - F) + \bar{p}_\lambda(\hat{\beta}_1)\} \) is always invertible for an \( \epsilon \) small enough. Furthermore by (3.13) and Condition B, \( \int \hat{\psi}_{\beta_1}(\beta^*)dF \leq \sqrt{\alpha_0}\lambda P_\lambda(0+) \). Therefore (3.15) and (3.18) imply
\[
\|\beta^{(\epsilon)} - \beta^*\|_2 = O(\epsilon).
\]

\[\square\]

**Proof of Theorem 2**

The result can be obtained along the lines of Theorem 2 of Fan and Li [2001].\[\square\]

**Proof of Theorem 3**

We first prove part (1) by following essentially the arguments of the proof of Theorem 4 in Fan and Lv [2011]. Note that we do not require third moment conditions on the responses since the boundedness of the derivatives of our loss function allows us to use the Lindeberg-Feller conditions. Let \( \xi = (\xi_1, \ldots, \xi_d)^T = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Psi_i(\beta_0) = \{\Psi_1^{(n)}(\beta_0)^T, \Psi_2^{(n)}(\beta_0)^T\}^T \) and \( \tilde{\beta} = (\tilde{\beta}_1, 0^T)^T \). For part (1) we only need to check that \( \tilde{\beta} \) satisfies the KKT conditions. Given Lemma 2 we only need to verify that
\[
\left\| \hat{w}_2^{-1} \circ \Psi_2^{(n)}(\beta) \right\| \leq \lambda_n,
\]
where \( \hat{w}_2^{-1} \) is a \( d \) dimensional vector with components \( 1/\hat{w}_j \) for \( j \in \mathcal{A}^c = \{j : \beta_{0j} = 0\} \).

This inequality holds under the event \( \mathcal{E} = \left\{ \|\xi_{\mathcal{A}^c}\|_\infty \leq u_n/\sqrt{n} \right\} \) where \( u_n = c_0^{-1/2}n^{1/2} \sqrt{\log n} \).

To see this, first let \( \phi(\delta) = \{\phi_{k+1}(\delta), \ldots, \phi_p(\delta)\}^T = \Psi_2^{(n)}(\delta) \) where \( \delta \in \{\beta' \in \mathbb{R}^d : \beta_2 = 0\} \).

From an integral form Taylor expansion we have
\[
\Psi_2^{(n)}(\beta) = \Psi_2^{(n)}(\beta_0) + \tilde{M}_{21}^{(n)}(\hat{\beta}_1 - \beta_1),
\]
where \( \tilde{M}_{21}^{(n)} = \int_0^1 \dot{\Psi}_2^{(n)}(\beta_0 + t(\hat{\beta} - \beta_0)) \ dt \).

Combining this with the \( \sqrt{n/k} \) convergence of \( \hat{\beta}_1 \) and (3.10) yields
\[
\left\| \Psi_2^{(n)}(\beta) \right\|_\infty \leq \left\| \Psi_2^{(n)}(\beta_0) \right\|_\infty + \left\| \tilde{M}_{21}^{(n)}(\hat{\beta}_1 - \beta_1) \right\|_\infty \\
\leq n^{-\alpha} \sqrt{\log n} + O(1) \left\| \hat{\beta}_1 - \beta_1 \right\|_2 \\
= n^{-\alpha} \sqrt{\log n} + O_p(\sqrt{k/n})
\]

Hence
\[
\lambda_n^{-1} \left\| \hat{w}_2^{-1} \circ \Psi_2^{(n)}(\beta) \right\|_\infty \leq \lambda_n^{-1} \left\| \hat{w}_2^{-1} \right\|_\infty \left\{ n^{-\alpha} \sqrt{\log n} + O_p(\sqrt{k/n}) \right\} \\
= O_p \left( \lambda_n^{-1} r_n^{-1} n^{-\alpha} \sqrt{\log n} \right) + O_p \left( \lambda_n^{-1} r_n^{-1} \sqrt{k/n} \right) \\
= o_p(1).
\]

Furthermore, the event \( \mathcal{E} \) occurs with high probability. To see this notice that from Condition (A3) we have \( |\Psi_{ij}(\beta_0)| = |\frac{\partial Q_M(y_i, \mu_{0i})}{\partial \eta_i}| \leq |A_1(x_i)x_{ij}| \leq \tilde{c} \).

Hence from Bonferroni’s inequality and Hoeffding’s inequality we have
\[
P(\mathcal{E}) \geq 1 - \sum_{j \in \mathcal{A}^c} P(n|\xi_j| > u_n \sqrt{n}) \geq 1 - 2(d - k) \exp \left( - \frac{nu_n^2}{2n\tilde{c}^2} \right) \to 1
\]
since $\log d = O(n^a)$. This completes the proof of part (1).

We now prove part (2) of the theorem. From Lemma 2 we know that $\hat{\beta}_1$ is, for a sufficiently large $n$, the adaptive lasso solution of (3.6) restricted to the the subspace $\{\beta \in \mathbb{R}^d : \beta_{A^c} = 0\}$. By $\sqrt{n/k}$-consistency of $\hat{\beta}_1$, we have

$$0 = \Psi_1^{(n)}(\beta_0) + \int_0^1 \Psi_1^{(n)}\{\beta_0 + t(\hat{\beta} - \beta_0)\}dt(\hat{\beta}_1 - \beta_1) + \lambda_n \hat{w}_1 \circ \text{sgn}(\hat{\beta}_1)$$

$$= \Psi_1^{(n)}(\beta_0) + \hat{M}_1^{(n)}(\hat{\beta}_1 - \beta_1) + O_p(\lambda_n \sqrt{k})$$

Premultiplying the previous expression by $\sqrt{n}Q_{11}^{-1/2}$, we obtain

$$-Q_{11}^{1/2}\hat{M}_1^{(n)}(\hat{\beta}_1 - \beta_1) = Q_{11}^{1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^n \Psi_{11}(\beta_0) + o_p(1) = Q_{11}^{1/2} u_n + o_p(1).$$

By consistency of $\hat{\beta}_1$ and Lemma 1 we have $\hat{M}_1^{(n)} \to_p M_{11}$. Hence it suffices to prove that $a^TQ_{11}^{-1/2}u_n \to_d N(0, 1)$ and then invoke Slutsky’s lemma to complete the proof. Therefore we verify that $u_n$ satisfies the conditions of the Lindeberg-Feller Central Limit Theorem. Note that $E\{\Psi_{11}(\beta_0)\} = 0$ and

$$\text{Var}(u_n) = \text{Var}\left\{n^{-1/2} \sum_{i=1}^n \Psi_{11}(\beta_0)\right\} = Q_{11}.$$

By the Cauchy-Schwarz inequality for any $\epsilon > 0$ we have

$$\sum_{i=1}^n E\left[\left\|\Psi_{11}(\beta_0)\right\|_2^4 \mathbf{1}\{\left\|\Psi_{11}(\beta_0)\right\|_2 > \epsilon\right\}\right] = nE\left[\left\|\Psi_{11}(\beta_0)\right\|_2^4 \mathbf{1}\{\left\|\Psi_{11}(\beta_0)\right\|_2 > \epsilon\right\}\right]$$

$$\leq n\left\{E\left[\left\|\Psi_{11}(\beta_0)\right\|_2\right]^4\right\}^{1/2} \left\{P\left[\left\|\Psi_{11}(\beta_0)\right\|_2 > \epsilon\right]\right\}^{1/2}.$$

By Markov’s inequality and (C2)

$$P\left[\left\|n^{-1/2}\Psi_{11}(\beta_0)\right\|_2 > \epsilon\right] \leq E\left[\left\|\Psi_{11}(\beta_0)\right\|_2^2\right]/\epsilon^2 = O(k/n).$$

Furthermore

$$E\left[\left\|n^{-1/2}\Psi_{11}(\beta_0)\right\|_2^4\right] = E\left\{n^{-1}\Psi_{11}(\beta_0)^T\Psi_{11}(\beta_0)\right\}^2 = O(k^2/n^2).$$

From (3.19)-(3.21) we have

$$\sum_{i=1}^n E\left[\left\|\Psi_{11}(\beta_0)\right\|_2^2 \mathbf{1}\{\left\|\Psi_{11}(\beta_0)\right\|_2 > \epsilon\right\}\right] \leq O\left\{n(k/n)^{3/2}\right\} = o(1).$$

We can therefore invoke the Lindeberg-Feller Central Limit Theorem to show that $a^TQ_{11}^{-1}u_n \to_d N(0, 1)$. This completes the proof of part (2). □
Proof of Theorem 4

It suffices to show that the conditions of Theorem 1 in Loh and Wainwright [2015] hold with the claimed probability. Note that the set $\Omega$, the function $g : \mathbb{R}^p \to \mathbb{R}_+$ and the positive constant $R$ used in their theorem can be respectively replaced in our set up by $\mathbb{R}^p$, $\| \cdot \|_1$ and a positive term proportional to $1/\lambda$. In order to apply their result we show that with probability at least $1 - 2e^{-\gamma_1 n/k^2}$, the robust quasilikelihood satisfies the restricted strong convexity discussed in their article. This would complete the proof in view of Lemma 3.

The restricted strong convexity condition requires

$$
E_n(\Delta) \geq \begin{cases} 
\alpha_1 \| \Delta \|_2^2 - \frac{\log d}{n} \| \Delta \|_1^2 & \text{for all } \| \Delta \|_2^2 \leq 1 \\
\alpha_2 \| \Delta \|_2^2 - \frac{\log d}{n} \| \Delta \|_1^2 & \text{otherwise.}
\end{cases} \tag{3.22a}
$$

where the $\alpha_j$’s are strictly positive constants, the $\tau_j$’s are nonnegative constants and

$$
E_n(\Delta) = \frac{1}{n} \left\{ \sum_{i=1}^{n} \Psi_i(\beta + \Delta) - \sum_{i=1}^{n} \Psi_i(\beta) \right\}^T \Delta.
$$

Note that

$$
E_n(\Delta) = \Delta^T M_n(\beta') \Delta \geq \Delta^T M(\beta') \Delta - \left| \Delta^T \left\{ M(\beta') - M_n(\beta') \right\} \Delta \right|,
$$

where $M_n(\beta) = \frac{1}{n} \sum_i \hat{\Psi}_i(\beta)$ and $\beta'$ lies between $\beta$ and $\beta + \Delta$. Denote by $M^k$ and $M_n^k$ the $2k \times 2k$ submatrices of $M$ and $M_n$ formed from $2k$ variables. Then by the arguments given in Lemma 1, we have that

$$
u^T M_n^k(\beta) \nu \geq c_1 + u^T \left\{ M_n^k(\beta') - M^k(\beta') \right\} u
$$

and

$$
P\{ u^T (M_n^k - M^k) u > \delta \} \geq 1 - 2e^{-\delta^2 n/k^2},
$$

where $\delta$ and $c$ are some positive constants. In particular we can choose $\delta = c_1/54$ and apply Lemma 12 in Loh and Wainwright [2012] with $s = \frac{n}{\log d}$ to the second term in (3.23). From this we see that

$$
E_n(\Delta) \geq c_1 \| \Delta \|_2^2 - \left( \frac{c_1}{2} \| \Delta \|_2^2 + \frac{c_1 \log d}{n} \| \Delta \|_1^2 \right)
$$

holds for all $\Delta \in \mathbb{R}^d$. This shows that (3.22a) holds with probability at least $1 - 2e^{-\gamma_1 n/k^2}$ taking $\gamma_1 = cc_2^2/54^2$. By Lemma 10 in Loh and Wainwright [2015] taking a sufficiently large $n$ guarantees (3.22b). □

Proof of Theorem 5

The proof follows from arguments very similar to those provided for Theorem 3 and are omitted for the sake of space. □
Proof of Theorem 6

We follow an argument similar to the one given for the proof of Theorem 2 in Fan and Lv [2011].

Two high probability events: Remember that the KKT conditions for the penalized robust quasilikelihood estimator are given by (3.11) and (3.12). Consider the set \( A = \{ j : \beta_{0j} \neq 0 \} \) as well as its complement \( A^c \). Let \( \xi = (\xi_1, \ldots, \xi_d)^T = \frac{1}{n} \sum_{i=1}^{n} \Psi_i(\beta_{0i}), \) where \( \beta_{e,0} = \arg \min_{\delta \in \mathbb{R}^d} E_F, \rho(\delta ; Z_i) \) and \( R = \{ \delta : \delta \in \mathbb{R}^d, \delta = (\delta_{A^c}, 0^T) \} \). We will prove that the weak oracle properties hold under the events \( E_1 = \{ \| \xi_A \|_\infty \leq c_0^{-1/2} \sqrt{\log(n)/n} \} \) and \( E_2 = \{ \| \xi_{A^c} \|_\infty \leq u_n/\sqrt{n} + b_t \} \), where \( u_n = c_0^{-1/2} n^{-(1/2+\alpha)} \sqrt{\log n} \) and \( b_t = \| E_F, \Psi^{(n)}(A)(\beta_{c,0}) \|_\infty \).

This will be enough to prove the claimed result since from Bonferroni’s inequality we obtain

\[
P(E_1 \cap E_2) \geq 1 - \sum_{j \in A} P(|\xi_j| > c_0^{-1/2} \sqrt{\log(n)/n}) - \sum_{j \in A^c} P(|\xi_j| > u_n/\sqrt{n} + b_t) \\
\geq 1 - \sum_{j \in A} P(|\xi_j| > c_0^{-1/2} \sqrt{\log(n)/n}) - \sum_{j \in A^c} P(|\xi_j - E_\xi_j| > u_n \sqrt{n}) \\
\geq 1 - 2\{kn^{-1} + (d-k)e^{-n^{1-2\alpha}/2} \log n \},
\]

where the last inequality follows from Lemma 4.

Bounded \( \ell_\infty \)-norm property: Let us first focus on the non zero coefficients. We will show that for sufficiently large \( n \), (3.11) has a solution \((\hat{\beta}_1^T, 0^T)^T\), with \( \hat{\beta}_1 \) inside the hypercube \( \mathcal{C} = \{ \delta \in \mathbb{R}^k : \| \delta - \beta_{c,1} \|_\infty = n^{-\xi} \log n \} \) with high probability. With this result we obtain the desired property because the triangular inequality and a von Mises expansion yield

\[
\| \hat{\beta}_1 - \beta_1 \|_\infty \leq \| \hat{\beta}_1 - \beta_{c,1} \|_\infty + \| \beta_1 - \beta_{c,1} \|_\infty = O(n^{-\xi} \log n + \epsilon)
\]

To simplify notation, allow the slight abuse \( \Psi^{(n)}(\beta_1) = \Psi^{(n)}(\{\beta_1^T, 0^T\}^T) \). A Taylor expansion around \( \beta_1 \) gives

\[
\Psi^{(n)}(\delta) = \Psi^{(n)}(\beta_{c,1}) + \tilde{M}_{11}^{(n)}(\delta - \beta_{c,1}), \tag{3.24}
\]

where \( \tilde{M}_{11}^{(n)} = \int_0^1 \Psi^{(n)}(\beta_{c,1} + t(\delta - \beta_{c,1})) dt \) By Lemma 1, for a sufficiently large \( n \) and a small enough \( \epsilon \), \( \tilde{M}_{11}^{(n)} \) will be positive definite with bounded eigenvalues and probability tending to one. Therefore checking (3.11) is equivalent to verifying \( \bar{\Psi}(\delta) = 0 \), where

\[
\bar{\Psi}(\delta) = (\tilde{M}_{11}^{(n)})^{-1}\{\Psi^{(n)}(\beta_{c,1}) + \tilde{p}_\lambda(\delta)\} + \delta - \beta_{c,1} = u + \delta - \beta_{c,1} \tag{3.25}
\]

For any \( \delta = (\delta_1, \ldots, \delta_k)^T \in \mathcal{C} \), given that \( s_n \geq n^{-\xi} \log n \), we have \( \min_{1 \leq j \leq k} |\delta_j| \geq \min_{j \in A} |\beta_{0j}| = s_n \) and \( \text{sgn}(\delta_j) = \text{sgn}(\beta_{0j}) \). Using the monotonicity condition of \( P'(t, \lambda) \) we have \( \| \tilde{p}_\lambda(\delta) \|_\infty \leq p'_\lambda(s_n) \). Therefore under the event \( E_1 \) we have

\[
\| \Psi^{(n)}(\beta_{c,1}) + \tilde{p}_\lambda(\delta) \|_\infty \leq c_1^{-1/2} \sqrt{n \log n} + p'_\lambda(s_n). \tag{3.26}
\]

Combining (3.24)-(3.26), we have \( \| u \|_\infty = o(n^{-\xi} \log n) \). Hence for \( j = 1, \ldots, k \) the sign of \( (\delta - \beta_{c,1})_j \) will determine the sign of \( \Psi_j(\delta) \), where \( \Psi(\delta) = (\Psi_1(\delta), \ldots, \Psi_k(\delta))^T \) and \( \delta \in \mathcal{C} \). Hence by continuity of \( \bar{\Psi}(\delta) \), the equation \( \bar{\Psi}(\delta) = 0 \) has a solution \( \hat{\beta}_1 \in \mathcal{C} \).
**Sparsity property:** Let us now turn to the support recovery and verify that the solution \( \hat{\beta} = (\hat{\beta}_1^T, 0^T)^T \in \mathbb{R}^p \) satisfies (3.12), where \( \hat{\beta}_1 \in \mathcal{C} \). A Taylor expansion of \( \Psi^{(n)}(\hat{\beta}) \) around \( \beta_\epsilon \) yields

\[
\Psi^{(n)}_2(\hat{\beta}) = \Psi^{(n)}_2(\beta_{1,1}) + \bar{M}^{(n)}_{21}(\hat{\beta}_1 - \beta_{1,1}),
\]

where \( \bar{M}^{(n)}_{21} = \int_0^1 \Psi^{(n)}_2(\beta_{1,1} + t(\hat{\beta}_1 - \beta_{1,1}))dt \). It is clear that on the event \( E_2 \), the first term of (3.27) is bounded by \( n^{-1/2}u_n + \epsilon b_\epsilon \). Therefore given that \( \hat{\beta}_1 \) solves \( \bar{\Psi}(\delta) \) we have

\[
\| \Psi^{(n)}_2(\hat{\beta}) \|_\infty \leq \epsilon b_\epsilon + c_3\lambda K_{\delta} + o(1) + \| \bar{M}^{(n)}_{21} \|_{\infty} \| \Psi^{(n)}_2(\beta_{1,1}) \|_\infty + \| \bar{M}^{(n)}_{11} \|_{\infty} \{ o(1) + \lambda_n \epsilon b_\epsilon + \| \bar{M}^{(n)}_{11} \|_{\infty} \}. \tag{3.28}
\]

By (C3'), for sufficiently large \( n \) and \( \epsilon \) small enough, we further have that

\[
\| \Psi^{(n)}_2(\hat{\beta}) \|_\infty \leq \epsilon b_\epsilon + c_3\lambda_n P(0+) + o(1) < \lambda_n P(0+).
\]

This completes the proof. \( \square \)

### 3.6.3 Auxiliary Lemmas

The following lemmas are used to prove our main results.

**Hoeffding’s inequality.** Let \( Z_1, \ldots, Z_n \) be independent random variables in some space \( \mathcal{Z} \) and let \( \gamma \) be a real-valued function on \( \mathcal{Z} \) satisfying \( E[\gamma(Z_i)] = 0 \) and \( |\gamma(Z_i)| \leq c_i \) \( \forall i \). Then we have for all \( t \)

\[
P\left( \left| \sum_{i=1}^n \gamma(Z_i) \right| \geq t \right) \leq 2 \exp \left( -\frac{t^2}{2 \sum_{i=1}^n c_i^2} \right).
\]

**Proof:** See e.g. Lemma 14.11 in Bühlmann and Van De Geer [2011]. \( \square \)

**Lemma 1.** Under conditions (C1)-(C2) with probability at least \( 1 - 2e^{-cn/d} \) we have \( \lambda_{\min} \{ \sum_{i=1}^n \dot{\Psi}_i(\beta) \} \geq C n \) for all \( \beta \in \mathcal{O} \), where \( c \) and \( C \) are some positive constants. Furthermore if \( d/n \rightarrow 0 \) then \( \frac{1}{n} \sum_{i=1}^n \dot{\Psi}_i(\beta) \rightarrow_p M(\beta) \)

**Proof:** We need to show that \( \frac{1}{n} u^T \sum_{i=1}^n \dot{\Psi}_i(\beta) u > 0 \) for any \( d \) dimensional vector \( u \). Without loss of generality we assume \( ||u||_2 = 1 \). It is easily seen that

\[
u^T \sum_{i=1}^n \dot{\Psi}_i(\beta) u = u^T \left( \sum_{i=1}^n E\{ \dot{\Psi}_i(\beta) \} + \sum_{i=1}^n (\dot{\Psi}_i(\beta) - E\{ \dot{\Psi}_i(\beta) \}) \right) u \geq c_1 n + \sum_{i=1}^n \tilde{\gamma}_i
\]

where \( \tilde{\gamma}_i = (x_i^T u)^2 \gamma_i - E\{(x_i^T u)^2 \gamma_i \} \) and \( \gamma_i = \partial^2 Q_M \{ y_i, g^{-1}(x_i^T \beta) \} / \partial y_i^2 \). Note that the terms \( \gamma_i \) have mean zero, finite variance and by Condition (C2) are bounded by \( dK \), where \( K \) is some positive constant. Hence from Hoeffding’s inequality we have

\[
P\left( \left| \sum_{i=1}^n \tilde{\gamma}_i \right| \leq c_1 n \right) \geq 1 - 2 \exp \left( -\frac{c_1^2 n^2}{2nd^2 K^2} \right) = 1 - 2e^{-cn/d^2}.
\]

It is readily seen that the last inequality also implies \( \frac{1}{n} \sum_{i=1}^n \dot{\Psi}_i(\beta) \rightarrow_p M(\beta) \). \( \square \)
Lemma 2: Assume there are weights \( \hat{w}_j > 0 \) satisfying \( \hat{w}_j = O_p(1) \) for \( j \in A = \{ j : \beta_{0j} \neq 0 \} \). Let \( \lambda_n = O(n^{-1/2}) \) and \( s_n = 2^{-1} \min \{ \beta_{0j} : \beta_{0j} \neq 0 \} \) be half of the minimum signal with \( s_n > k/n \) and \( k = o(n) \). Further let \( \hat{\beta}_1 \) be the estimator of \( \beta_1 \) obtained by restricting the robust adaptive lasso solution to the subspace \( \{ \beta \in \mathbb{R}^d : \beta_A = 0 \} \). Then under the conditions (C1)-(C2) \( \hat{\beta}_1 \) is \( \sqrt{n/k} \)-consistent.

Proof: We follow closely the arguments in Step 1 of the proof of Theorem 3 in Fan and Lv [2011]. The oracle estimator of the robust adaptive lasso is defined as the minimizer of the constrained penalized robust quasilikelihood

\[
PQ_R(\beta_1') = \frac{1}{n} \sum_{i=1}^{n} Q_M(y, X_i^T \beta_1') + \lambda_n \sum_{j=1}^{k} \hat{w}_j |\beta'_j|.
\]

We define \( H_n = \{ PQ_R(\beta_1) < \min_{\delta} PQ_R(\delta) : \delta \in \partial N_\tau \} \), \( \partial N_\tau \) the boundary of the closed set \( N_\tau = \{ \delta \in \mathbb{R}^k : \| \delta - \beta_1 \|_2 \leq k/n \tau \} \) and \( \tau \in (0, \infty) \). In order to show \( \| \hat{\beta}_1 - \beta_1 \| = O(\sqrt{k/n}) \) we show that \( P(H_n) \rightarrow 1 \) as \( n \rightarrow \infty \) when \( \tau \) is large. This would complete the lemma because clearly on the event \( H_n \) there exists a minimizer \( \hat{\beta}_1 \) of \( PQ_R(\delta) \) in \( N_\tau \). Let \( n \) be sufficiently large such that \( \sqrt{k/n} \tau \leq s_n \). Then since \( \delta \in N_\tau \) we have \( sgn(\delta) = sgn(\beta_1) \), \( \| \delta - \beta_1 \|_\infty \leq s_n \) and \( \min_j |\delta_j| \geq s_n \). Therefore a Taylor expansion yields

\[
PQ_R(\beta_1) - PQ_R(\delta) = -\Psi_1^{(1)}(\beta_1)(\delta - \beta_1) - \frac{1}{2}(\delta - \beta_1)^T \hat{M}_{\tau}(\delta - \beta_1)
\]

\[
- \lambda_n \sum_{j=1}^{k} \hat{w}_j \text{sgn}(\delta_j)(\delta_j - \beta_{1j}),
\]

where \( \hat{M}_{\tau} = \int \Psi_1^{(1)}(\beta_1 + t(\delta - \beta_1)) dt \). Let \( A_n = \{ PQ_R \text{ is convex} \} \cap \{ \max_{j \in A} \hat{w}_j < K \} \) for some \( K < \infty \). From Lemma 1 and \( \hat{w}_j = O_p(1) \) we know that \( P(A_n) \rightarrow 1 \) as \( n \rightarrow \infty \). Furthermore under the event \( A_n \) there exists a constant \( l \) such that \( \lambda_{\min}(\hat{M}_{\tau}) \geq l > 0 \) and

\[
PQ_R(\beta_1) - \min_{\delta \in \partial N_\tau} PQ_R(\delta) \leq \sqrt{k/n} \tau \| v \|_2 (\| v \|_2 - l \sqrt{k/n} \tau / 2)
\]

where \( v = \Psi_1^{(1)}(\beta_1) + \lambda_n \hat{w}_1 \) and \( \hat{w}_1 = (\hat{w}_1, \ldots, \hat{w}_k)^T \). Combining this with Markov’s inequality yields

\[
P(H_n | A_n) \geq P \left( \| v \|_2^2 \leq \frac{l^2 k \tau^2}{4n} \right) \geq 1 - \frac{4nE \| v \|_2^2}{k l^2 \tau^2}.
\]

(3.29)

It follows from \( E[\Psi_1^{(1)}(\beta_1)] = 0 \) and (3.9) that

\[
E \| v \|_2^2 = E \| \Psi_1^{(n)}(\beta_1) \|_2^2 + \lambda_n^2 \| \hat{w}_1 \|_2^2
\]

\[
\leq n^{-2} E \text{tr} \left[ X_1^T \{ \Upsilon(X \beta_0) \}^2 X_1 \right] + \lambda_n^2 k K^2 = O(kn^{-1}).
\]

(3.30)

From (3.29) and (3.30) finally obtain

\[
P(H_n) \leq P(H_n | A_n) P(A_n) \rightarrow_n 1 - O(\tau^{-2}). \quad \square
\]
Lemma 3: Under conditions (C1)-(C3), for some positive constants $c$ and $K$ we have

$$
P\left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \Psi_i(\beta_0) \right\|_\infty < K \sqrt{\frac{\log d}{n}} \right\} \geq 1 - 2e^{-c \log d}.
$$

Proof: Note that from Condition (C2) we have $|\Psi_{ij}(\beta)| \leq A_1(x_i)|x_{ij}| \leq \tilde{c} < \infty$. Therefore letting $c = K^2/(2\tilde{c}^2)$ the result follows immediately from Hoeffding’s inequality.

Lemma 4: Let $a_n$ be a diverging sequence and $\xi = (\xi_1, \ldots, \xi_d)^T = \frac{1}{n} \sum_{i=1}^{n} \Psi_i(\beta_{\epsilon,0}) = \Psi^{(n)}(\beta_{\epsilon,0})$, where $\beta_{\epsilon,0} = \arg \min_{\delta \in \mathcal{R}} E_F \rho(\delta; z_i)$, $\mathcal{R} = \{ \delta : \delta \in \mathbb{R}^d, \delta = (\delta_1^T, 0^T)^T \}$. Then under condition C we have

1. $P(|\xi_j - E\xi_j| \geq a_n/\sqrt{n}) \leq 2e^{-c_0 a_n^2}$
2. $\max_{k+1 \leq j \leq d} P(|\xi_j| \geq a_n/\sqrt{n} + b_\epsilon) \leq 2e^{-c_0 a_n^2}$, where $b_\epsilon = \max_{k+1 \leq j \leq d} |\xi_j|$.

Proof: Note that condition (C2) implies $\tilde{c} \geq |\psi_{ij}(\beta_0)|$. Thus Hoeffding’s inequality gives $P(n|\xi_j - E\xi_j| \geq a_n/\sqrt{n}) \leq e^{-a_n^2/(2n^2)}$ for $j = 1, \ldots, d$. Hence (i) is proven by letting $c_0 = 1/2\tilde{c}^{-2}$. Since $|\xi_j| - b_\epsilon \leq |\xi_j| - E|\xi_j| \leq |\xi_j| - |E\xi_j| \leq |\xi_j - E\xi_j|$, we can apply (i) to obtain (ii). □
Chapter 4

The influence function of penalized M-estimators

General remarks

We study the local robustness properties of general nondifferentiable penalized M-estimators via the influence function. More precisely, we propose a framework that allows us to define rigorously the influence function as the limiting influence function of a sequence of approximating estimators. We show that it can be used to characterize the robustness properties of a wide range of sparse estimators and we derive its form for general penalized M-estimators including lasso and adaptive lasso type estimators. We prove that our influence function is equivalent to a derivative in the sense of distribution theory.

4.1 Introduction

Sparse models have become very popular in recent years. Since the introduction of penalized methods to study then in the linear model (Breiman [1995], Tibshirani [1996]), many extensions and algorithms have been proposed. Fan and Lv [2010] and Tibshirani [2011] provide good reviews. Asymptotic properties of lasso-type estimators have been studied in the fixed dimensional parameter case (Knight and Fu [2000], Fan and Li [2001], Zou [2006]), as well as in the high dimensional set up where the number of parameters is allowed to grow at an even faster rate than the sample size (Bühlmann and Van De Geer [2011]).

Given the increasing importance that sparsity inducing penalties play in modern statistics, the need for a clear understanding of the robustness properties of these type of procedures is evident. Robust statistics develops a theoretical framework that allows us to take into account that the models used for fitting the data are only idealized approximations of reality. It provides methods that still give reliable results when slight deviations from the stochastic assumptions on the model occur. Book-length expositions can be found in Huber [1981] and 2nd edition by Huber and Ronchetti [2009], Hampel et al. [1986] and Maronna et al. [2006].

Some authors have suggested sparse estimators that limit the impact of contamination in the data (e.g. Sardy et al. [2001], Wang et al. [2007a], Li et al. [2011], Lozano and Meinshausen [2013] and Fan et al. [2014a] among many others). These procedures rely on the intuition that a loss function that defines robust estimators in the well understood
unpenalized fixed dimensional M-estimation set up, should also define robust estimators when it is penalized by a deterministic function. In the linear model for instance, Fan et al. [2014a] show that under very mild conditions on the error term their estimator satisfies the oracle properties.

One of the main lines of research in the robustness literature was opened by Hampel [1974] who considered local robustness, i.e. the impact of moderate distributional deviations from ideal models on a statistical procedure. In this setting the quantities of interest are viewed as functionals of the underlying distribution. Typically their linear approximation is studied to assess the behavior of the estimators in a neighborhood of the model. In this approach the influence function plays a crucial role in describing the local stability of the functional analyzed. It allows for an easy assessment of the relative influence of individual observations on the value of an estimate. If it is unbounded, a single outlier could cause trouble. If a statistical functional $T(F)$ is sufficiently regular, a von Mises expansion (von Mises [1947]) yields

$$ T(G) \approx T(F) + \int IF(z; F, T)d(G - F)(z), $$

(4.1)

where $IF(z; F, T)$ denotes the influence function of the functional $T$ at the distribution $F$. Considering the approximation (4.1) over an $\epsilon$ neighborhood of the model $F_\epsilon = \{G | G = (1 - \epsilon)F + \epsilon H, H$ arbitrary $\}$, we see that the influence function can be used to linearize the asymptotic bias in a neighborhood of the ideal model. Therefore a bounded influence function implies a bounded approximate bias.

The goal of this article is to give a formal definition of the influence function for a wide class of penalized M-estimators, that covers most of the existing proposals. This requires developing a new framework. Indeed, the typical tools used to derive the influence function of M-estimators suffer from a problems when considering penalized M-estimators: they cannot handle nondifferentiable penalty functions which are necessary for achieving sparsity (Fan and Li [2001]). Note that in a recent paper Wang et al. [2013] give the form of the influence function for their penalized M-estimator. However it is derived in a limited set up without developping an appropriate rigorous framework. Further details can be found in Section 5.

Our work provides a number of contributions to the existing literature. First, we introduce an influence function defined through a sequence of approximating estimators and show that it is uniquely defined for penalized M-estimators and two-stage penalized M-estimators. The former class covers lasso and group lasso type estimators. The latter includes adaptive lasso type estimators. We compute the influence function of all these important examples. Second, we show that the two main features of the influence functions for M-estimators can also be valid for the influence function of penalized M-estimators, i.e. (a) it allows to assess the relative influence of individual observations towards the value of an estimate; (b) it allows an immediate and simple, informal assessment of the asymptotic properties of an estimate (Huber and Ronchetti [2009], p. 14-15). Third, we show that our limiting influence function can be viewed as a distributional derivative in the sense of Schwartz [1959]. This opens the door for further research exploiting the tools of distribution theory, which have essentially not been used in the statistical literature previously. Finally, a key step in our theoretical argument is the innovative use of Berge’s maximum theorem. This is a powerful tool that could have more applications in statistics.

The rest of the article is organized as follows. Section 2 introduces the general framework and provides the main results regarding our influence function as well as some examples. Section 3 extends the results to two step M-functionals. Section 4 provides some
numerical illustrations of the local robustness of penalized estimators and its relation to our influence function. Section 5 establishes the connection between our influence function and distributional derivatives. Finally Section 6 concludes with a discussion around the empirical influence function of penalized M-estimators. A selection of the proofs of our main results is given at the end of the article. Additional proofs and auxiliary results are provided in Appendix 4.8.

4.2 Penalized M-estimators and Influence Functions

4.2.1 Background

Consider a collection of $n$ observations $\{z_1, \ldots, z_n\}$ drawn from a common distribution $F$ over the space $Z$, a parameter space $\Theta$ which is an open subset of $\mathbb{R}^d$ and a loss function $L : Z \times \mathbb{R}^d \mapsto \mathbb{R}$. Let $\hat{F} = \frac{1}{n} \sum_{i=1}^{n} \Delta_{z_i}$ be the empirical distribution of $F$ corresponding to the observed sample, where $\Delta_{z}$ is the distribution probability that assigns mass 1 at the point $z$ and 0 elsewhere. Then $E_{\hat{F}}[L(Z, \theta)] = \frac{1}{n} \sum_{i=1}^{n} L(z_i, \theta)$ serves as a measure of fit between the parameter vector $\theta \in \mathbb{R}^p$ and the observed data. This empirical loss function can be seen as an estimator of the unknown population risk function $E_F[L(Z, \theta)]$. We study the functionals resulting from the minimization of the regularized risk

$$\Lambda_{\lambda}(\theta; F) = E_F[L(Z, \theta)] + p(\theta; \lambda)$$

with respect to $\theta$, where $p(\cdot; \lambda)$ is a continuous penalty function with regularization parameter $\lambda$. Important illustrations of such functionals are considered in Examples 1, 2 and 3 of this paper.

Let $T(F) = \theta_{F,\lambda} = \theta^* = \text{argmin}_\theta \Lambda_{\lambda}(\theta; F)$. In robust statistics we are interested in constructing and studying smooth and bounded functionals $T$, because they yield stable minimizers within small neighborhoods of $F$. A bounded derivative $\nabla T(F)$ of $T(F)$ implies that the functional $T(F)$ cannot increase or decrease arbitrarily in small neighborhoods of $F$. One general approach to robustness is the one based on the influence function (Hampel et al. [1986]). Given a distribution space $\mathcal{F}$, a parameter space $\Theta$ and a functional $T : \mathcal{F} \mapsto \Theta$, the influence function of $T$ at a point $z \in Z$ for a distribution $F$ is defined as

$$\text{IF}(z; F, T) = \lim_{\epsilon \to 0^+} \frac{T(F_\epsilon) - T(F)}{\epsilon},$$

where $F_\epsilon = (1-\epsilon)F + \epsilon \Delta_{z}$. It has the heuristic interpretation of describing the effect of an infinitesimal contamination at the point $z$ on the estimate, standardized by the mass of contamination. The well studied M-functionals are defined implicitly as a root of

$$\int \Psi(z, T(F))dF = 0.$$  

They are a special case of the minimizers of (4.2) when the penalty function is set to be zero and the risk function is differentiable with respect to the parameter. The standard argument for showing the existence and deriving the form of the influence function of M-functionals, is to use an appropriate implicit function theorem. In our set up this would require that $\Lambda_{\lambda}$ has two derivatives respect to $\theta$. However it is well known that a penalty function has to be singular at the origin in order to achieve sparsity (Fan and Li [2001]). Therefore new tools that can deal with nondifferentiable penalty functions are required to derive influence functions of many modern penalized estimators. We propose to define
the influence function as the limit of the influence functions of a sequence of differentiable penalized M-estimators that converge to the penalized M-estimator of interest.

### 4.2.2 Limiting Influence Function

We will require the following set of assumptions for the derivation of our theoretical results:

(A1) \( E_F[L(Z, \theta)] \) is continuous in \( \Theta \) and has two bounded derivatives with respect to \( \theta \) denoted by \( E_F[\psi(Z, \theta)] \) and \( E_F[\dot{\psi}(Z, \theta)] \) respectively, with \( E_F[\dot{\psi}(Z, \theta)] \) nonsingular.

(A2) There is a unique point \( \theta_0 \in \Theta \) such that \( E_F[L(Z, \theta_0)] < \inf_{\theta \in \Theta \setminus U} E_F[L(Z, \theta)] \) for every open set \( U \) containing \( \theta_0 \).

(A3) There is a unique \( \theta^* \) in an open ball \( O \) containing \( \theta_0 \). The functions \( \psi(z, \theta) \) and \( E_F[\dot{\psi}(Z, \theta)] \) are continuous at \( \theta^* \) for all \( z \in \mathbb{R}^d \).

Assumptions (A1) and (A2) are regularity conditions on the population risk. Namely, they require some smoothness and a well separated maximum. Assumption (A3) imposes regularity conditions on the score function \( \psi \) and a unique regularized risk minimum on a neighborhood of the global risk minimum. Obviously when both the population risk and the penalty function are convex we can take \( O = \Theta \). Nonconvex loss functions and penalty functions restrain the size of the ball \( O \).

As discussed in the previous subsection, when the penalty function is sufficiently smooth, a simple application of the implicit function theorem establishes the existence and the form of the influence function of the minimizer of (4.2). This result is stated in the following lemma.

**Lemma 1:** Assume (A1)-(A3). Let \( p : \Theta \rightarrow \mathbb{R} \) be twice differentiable and \( S := E_F[\dot{\psi}(Z, \theta^*)] + \nabla^2 p(\theta^*; \lambda) \) be invertible. Then the influence function of \( T(F) \) exists for all \( z \in \mathbb{R}^d \) and we have

\[
IF(z; F, T) = -S^{-1}\left(\psi(z, \theta^*) + \nabla p(\theta^*; \lambda)\right).
\]

Note that \( S \) is a fixed matrix that depends on the distribution \( F \) as well as on the loss and penalty functions. From Lemma 1 we conclude that just as for a M-estimators, a bounded derivative for the loss function is required in order to obtain bounded influence estimators in the penalized setting. When the penalty function in (4.2) is not differentiable the conditions of Lemma 1 do not hold. We therefore propose to study the limiting form of the influence function of penalized M-estimators obtained using smooth penalty functions \( p_m \) such that \( \lim_{m \rightarrow \infty} p_m = p \). Such penalized M-estimators, denoted by \( T(F; p_m) \), are defined by the minimization problem

\[
\min_{\theta} \Lambda_\lambda(\theta; F, p_m) = \min_{\theta} \left\{ E_F[L(Z, \theta)] + p_m(\theta; \lambda) \right\}.
\]

We let \( IF_{p_m}(z; T, F) \) be the influence function of \( T(F; p_m) \) and define the influence function of \( T(F) \) as

\[
IF(z; F, T) := \lim_{m \rightarrow \infty} IF_{p_m}(z; F, T).
\]

A natural question that arises from this definition is whether the limit depends on the limiting sequence \( \{p_m\} \) chosen. In order to answer this question we first show that the
limiting functional $T(F)$ is unique. The use of Berge’s maximum theorem (Berge [1997]) is a key step in our proof. It is an innovative tool in the statistical literature. The lemma is crucial for the uniqueness argument of the limiting influence function, stated in Proposition 1. While completing this article, the author noticed that Machado [1993] has also used Berge’s maximum theorem for the derivation of qualitative robustness for model selection criteria based on M-estimators. We use the notation $C^2_b(\Theta)$ for the set of continuous and infinitely differentiable functions $p : \Theta \to \mathbb{R}$ and $W^{2,2}(\Theta)$ for the Sobolev space given by the subset of functions in $L^2(\Theta)$ whose second weak derivative is integrable in $L^2$ norm.

**Lemma 2** : Consider a sequence $\{p_m\}_{m \geq 1}$ in $C^\infty(\Theta)$ converging to $p$ in the Sobolev space $W^{2,2}(\Theta)$ when $m \to \infty$. Suppose that for the problem (4.2), each of the approximating problems (4.3) resulting from $\{p_m\}_{m \geq 1}$ satisfy (A1)-(A3). Then we have $\lim_{m \to \infty} T(F; p_m) = T(F)$.

**Proposition 1** : Under the conditions of Lemma 2, the limiting influence function defined in (4.4) does not depend on the choice of $p_m$.

**Remark A** : The uniqueness of (4.4) holds for any local minimum as long as they are well separated. Indeed (A3) can be relaxed by simply considering a unique minimizer contained in an open ball $O$ that does not necessarily contain $\theta_0$. We can see from the proofs of Lemma 2 and Proposition 1 that this is enough for them to hold.

**Example 1** : Lasso type penalties

Without loss of generality we will assume that the tuning parameter $\lambda$ is such that the resulting estimators are sparse. More specifically, we consider $\theta_{F,\lambda} = \theta^* = (\theta_1^*, \theta_2^T)^T$ with $\theta_1^* \in \mathbb{R}^s$, $s < d$ and $\theta_2^2 = 0$. The following proposition gives the form of the influence function of estimators that arise when considering a general class of penalty functions. It covers as special cases convex penalties such as the lasso (Tibshirani [1996]) and nonconvex penalties such as the scad (Fan and Li [2001]).

**Proposition 2** : Denote by $\theta^* = T(F)$ the penalized M-functional obtained as the minimizer of (4.2) with penalty functions of the form $p_\lambda(\theta) = \sum_{j=1}^p p_{\lambda,j}(\theta_j)$, where $p_{\lambda,j}(\cdot)$ are differentiable functions. Then under (A1)-(A3) the influence function (4.4) of $T(F)$ has the form

$$IF(z; F, T) = -S^{-1}\left(\psi(z, \theta^*) + \phi_\lambda(\theta^*)\right),$$

where $S^{-1} = \text{blockdiag}\{(M_{11} + P_\lambda)^{-1}, 0\}$, $M_{11} = E_F[\psi_{11}(Z, T(F))]$, $P_\lambda$ is a diagonal matrix with diagonal elements $p_{\lambda,j}(\theta_j)\left|\psi(\theta_j)\right|$ for $j = 1, \ldots, s$, and $\phi_\lambda(\theta^*)$ is a $d$ dimensional vector with components $p_{\lambda,j}(\theta_j)\text{sgn}(\theta_j)$ for $j = 1, \ldots, s$ and 0 elsewhere.

**Example 2** : Group lasso type penalties

We now give the form of the influence function of penalized M-estimators achieving sparsity for grouped variables via group lasso type penalties (e.g. Yuan and Lin [2006], Wang et al. [2007c], Huang et al. [2009]). We suppose, as in the previous example, that the regularization parameter $\lambda$ is such that the resulting regularized $\theta^*$ is sparse.

**Proposition 3** : Denote by $\theta^* = T(F)$ the penalized M-functional obtained as the minimizer of (4.2) with group penalty functions of the form $p_\lambda(\theta) = \sum_{g=1}^G p_{\lambda,g}(\|\theta_g\|_2)$,
where \( p_{\lambda,g}(\cdot) \) are differentiable functions, \( \theta = (\theta_1, \ldots, \theta_{(G)}) \) and each \( \theta_{(g)} \) is a subvector of \( \theta \) corresponding to the \( g \)th group of variables. Then under (A1)-(A3) the influence function (4.4) of \( T(F) \) has the form

\[
\text{IF}(z; F, T) = -S^{-1}\left( \psi(z, \theta^*) + \phi_\lambda(\theta^*) \right),
\]

where \( S^{-1} = \text{blockdiag}\{ (M_{11} + P_\lambda)^{-1}, 0 \} \), \( M_{11} = E_F[\psi_{11}(Z, T(F))] \) and \( P_\lambda \) is a block diagonal matrix with blocks \( p'_{\lambda,g}(\|\theta_{(g)}\|_2)(\theta_{(g)}'\theta_{(g)} - \|\theta_{(g)}\|_2 I_{|g|})/\|\theta_{(g)}\|_2^2 \) where

\[
\phi_\lambda(\theta^*) = \begin{cases} 
p'_{\lambda,g}(\|\theta_{(g)}\|)\theta^*_j/\|\theta_{(g)}\|_2 & \text{if } \theta^*_{(g)} \neq 0 \\
0 & \text{if } \theta^*_{(g)} = 0,
\end{cases}
\]

\( I_{|g|} \) is a diagonal matrix of size \( |g| \), i.e. the cardinality of group \( g \), and \( p'_{\lambda,g}(t) \) and \( p''_{\lambda,g}(t) \) are the first two derivatives of \( p_{\lambda,g}(t) \) for \( t > 0 \).

It is clear from Proposition 2 and Proposition 3 that a bounded \( \psi \) function is necessary for a penalized M-estimator to have bounded limiting influence function. The fact that the influence function has some zero components is rather surprising. Further discussion on this feature can be found in the last two sections.

### 4.3 Two-stage penalized M-estimators

We can extend the results of the previous section to a class of two-stage penalized M-estimators. Important examples of such estimators are the adaptive lasso estimators discussed below. The following set up can be viewed as a direct extension of the framework provided by Zhelonkin et al. [2012]. Let \( F \) be the distribution function of \( Z = (Z^{(1)}, Z^{(2)}) \) and let \( \theta = (\theta_1, \theta_2) \) be a vector defining the arguments of the first and second stages, with \( \theta_1 \in \Theta_1 \subset \mathbb{R}^{d_1}, \theta_2 \in \Theta_2 \subset \mathbb{R}^{d_2} \). We consider penalized M-estimators \( (\theta_1^*, \theta_2^*) = (S(F), T(F)) \) defined by

\[
\theta_1^* = \arg\min_{\theta_1} \left\{ E_F[L^{(1)}(Z^{(1)}, \theta_1)] + p^{(1)}(\theta_1; \gamma) \right\} \quad (4.5)
\]

\[
\theta_2^* = \arg\min_{\theta_2} \left\{ E_F[L^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1^*)] + p^{(2)}(\theta_2, \theta_1^*; \lambda) \right\} \quad (4.6)
\]

where \( L^{(i)} \) and \( p^{(i)} \) denote respectively the loss and penalty functions in the \( i \)th stage. For the theoretical argument we adapt assumptions (A1)-(A3) to this set up in the following straightforward way:

**(A1')** The loss function \( L^{(1)} \) and \( L^{(2)} \) are such that:

(i) \( E_F[L^{(1)}(Z^{(1)}, \theta_1)] \) is continuous in \( \Theta_1 \) and has two derivatives with respect to \( \theta_1 \) denoted by \( E_F[\psi^{(1)}(Z^{(1)}, \theta_1)] \) and \( E_F[\hat{\psi}^{(1)}(Z^{(1)}, \theta_1)] \) respectively. Furthermore \( E_F[\hat{\psi}^{(1)}(Z^{(1)}, \theta_1)] \) is nonsingular.

(ii) \( E_F[L^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)] \) is continuous in \( \Theta_1 \times \Theta_2 \) and has two derivatives with respect to \( \theta_2 \) denoted by \( E_F[\psi^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)] \) and \( E_F[\hat{\psi}^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)] \) respectively, with \( E_F[\hat{\psi}^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)] \) nonsingular. \( E_F[\psi^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)] \) has a derivative with respect to \( \theta_1 \) denoted by \( E_F[\hat{\psi}^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)] \).
4.3. Two-stage penalized M-estimators

We first provide the influence function of (4.6) for sufficiently smooth penalty functions. The adaptive lasso of Zou [2006] is a popular two stage procedure that improves on the results of the lasso by ensuring the oracle properties of Fan and Li [2001] under milder conditions. The tools developed in 4.2.2 allow us to derive the influence function of adaptive lasso type estimators.

(A2') There are unique points \( \theta_{01} \in \Theta_1 \) and \( \theta_{02} \in \Theta_2 \) such that:

(i) \( E_F[L^{(1)}(Z^{(1)}, \theta_{01})] < \inf_{\theta_1 \in \Theta_1 \setminus U_1} E_F[L^{(1)}(Z^{(1)}, \theta_1)] \) for every open set \( U_1 \) containing \( \theta_{01} \).

(ii) \( E_F[L^{(2)}(Z^{(2)}, \theta_{02}, Z^{(1)}, \theta_{01})] < \inf_{\theta_2 \in \Theta_2 \setminus U_2} E_F[L^{(2)}(Z^{(2)}, \theta_{02} Z^{(1)}, \theta_{01})] \) for every open set \( U_2 \) containing \( \theta_{02} \).

(A3') There are open subsets of \( \Theta_1 \) and \( \Theta_2 \) denoted by \( \mathcal{O}_1 \) and \( \mathcal{O}_2 \), and unique solutions \( \theta_1^* \in \mathcal{O}_1 \) and \( \theta_2^* \in \mathcal{O}_2 \) such that

(i) \( \theta_{01} \in \mathcal{O}_1 \), and the functions \( \psi(1)(z^{(1)}, \theta_1) \) and \( E_F[\psi(1)(Z^{(1)}, \theta_1)] \) are continuous at \( \theta_1^* \) for all \( z^{(1)} \in \mathbb{R}^{d_1} \).

(ii) \( \theta_{02} \in \mathcal{O}_2 \), and the functions \( \psi(2)(z^{(2)}, \theta_2, z^{(1)}, \theta_1) \) and \( E_F[\psi(2)(Z^{(2)}, \theta_2; Z^{(1)}, \theta_1)] \) are continuous at \( \theta^* \) for all \( z \).

We first provide the influence function of (4.6) for sufficiently smooth penalty functions.

**Lemma 3:** Denote by \( \theta^* = (S(F), T(F)) \) the estimators defined by (4.5)-(4.6), and assume (A1')-(A3'). Let \( p^{(i)} : \Theta_i \rightarrow \mathbb{R} \) be twice differentiable with respect to \( \theta_i \) and \( \nabla_{\theta_i} p^{(2)} \) be differentiable with respect to \( \theta_1 \). Let also \( S := E_F[\psi(2)(Z^{(2)}, \theta_2^*, Z^{(1)}, \theta_1^*); \lambda] + \nabla_{\theta_2, \theta_1} p^{(2)}(\theta_2^*, \theta_1^*; \lambda) \) be invertible. Then the influence function of \( T(F) \) exists for all \( z \in \mathbb{R}^{d} \) and we have:

\[
IF(z; F, T) = -S^{-1} \left( \psi(2)(Z^{(2)}, \theta_2^*, Z^{(1)}, \theta_1^*) + \nabla_{\theta_2} p^{(2)}(\theta_2^*, \theta_1^*; \lambda) \right)
\]

\[
+ \left( E_F[\psi_1^{(2)}(Z^{(2)}, \theta_2^*, Z^{(1)}, \theta_1^*); \lambda] + \nabla_{\theta_2, \theta_1} p^{(2)}(\theta_2^*, \theta_1^*; \lambda) \right) \right) IF(z^{(1)}; F, S) \right),
\]

where \( IF(z^{(1)}; S, F) \) has the form given in Lemma 1.

Unsurprisingly, bounded-influence estimators are obtained only by taking loss functions with bounded derivatives in both stages. The expression obtained is very similar to the one derived in the unpenalized set up by Zhelonkin et al. [2012]. We are now ready to state the uniqueness of the limiting two-stage estimator (4.6) and its influence function.

**Lemma 4:** For \( i = 1, 2 \), let \( \{p_{m_i}^{(i)}\} \) be a sequence in \( C^\infty(\Theta_i) \) converging to \( p^{(i)} \) in the Sobolev space \( W^{2,2}(\Theta_i) \) when \( m_i \rightarrow \infty \) and assume \( \{\nabla_{\theta_i} p_{m_i}^{(2)}\} \) is differentiable with respect to \( \theta_1 \). Consider the sequence of problems of the (4.2) implied by (4.5)-(4.6) and \( \{p_{m_i}^{(i)}\} \) for \( i = 1, 2 \). Assume that each of them satisfy (A1')-(A3'). Then we have \( \lim_m T(F; p_{m}^{(2)}) = T(F) \).

**Proposition 4:** Under the conditions of Lemma 4, the limiting influence function (4.4) of (4.6) does not depend on the choice of \( p_m = (p_{m_1}^{(1)}, p_{m_2}^{(2)}) \).

**Example 3:** Adaptive lasso type penalties

The adaptive lasso of Zou [2006] is a popular two stage procedure that improves on the results of the lasso by ensuring the oracle properties of Fan and Li [2001] under milder conditions. The tools developed in 4.2.2 allow us to derive the influence function of adaptive lasso type estimators.
Proposition 5: Let \( \theta \) be the \( d \) dimensional parameter of interest and let \( \theta^{(0)} = S(F) \) be an initial estimate of \( \theta \), with \( \theta^{(0)} \) non-zero components, defined by (4.5). For \( j = 1, \ldots, d \) and some nonnegative function \( w \), define the weights \( w_j = w(|\theta_j^{(0)}|) \). Denote by \( \theta^* = T(F) \) the penalized M-estimators obtained as the minimizer of (4.6) with a penalty function of the form \( p(\theta, \theta^{(0)}; \lambda) = \lambda \sum_{j=1}^d w_j |\theta_j| \) and loss function \( L(Z, \theta, \theta^{(0)}) \). Then under (A1')-(A3') the influence function (4.4) of \( T(F) \) has the form

\[
IF(z; F, T) = -S^{-1} \left( \psi(z, \theta^*) + \phi_\lambda(\theta^*, \theta^{(0)}) + \varphi_\lambda(\theta^*, \theta^{(0)}) \right) IF(z; F, S),
\]

where \( S^{-1} = \text{blockdiag}\{M_{11}^{-1}, 0\}, M_{11} = E_F[\dot{\psi}^{(1)}(Z, T(F))] \), \( \phi_\lambda(\theta^*, \theta^{(0)}) \) is a \( d \) dimensional vector with components \( \lambda w'(|\theta_j^{(0)}|) \text{sgn}(\theta_j^{(0)}) \text{sgn}(\theta_j^*) \) for \( j = 1, \ldots, d \) with \( w' \) denoting the derivative of \( w \), and \( \varphi_\lambda(\theta^*, \theta^{(0)}) = \text{blockdiag}\{H_\gamma, 0\} \) where \( H_\gamma = \{\nabla_{\theta_j^{(0)}} \phi_\lambda(\theta^*, \theta^{(0)})\}_{j,k=1}^{d} \).

The form of IF\((z; F, T)\) depends on the choice of the initial estimator. A bounded influence estimator \( T(F) \) can only be obtained by taking a bounded influence initial estimator \( S(F) \) and choosing a loss function defining a bounded \( \psi \) function in the second stage. Among the non-sparse initial estimates proposed in the literature, Zou [2006] proposed to use maximum likelihood estimates for the fixed parameter case where \( d \) does not vary with \( n \). In the high dimensional set up where \( d > n \), Huang et al. [2008] proposed to use an initial zero consistent estimate, e.g. marginal least squares. For those cases the influence function of \( S(F) \) is simply proportional to \( \psi^{(1)} \) as they are well known M-estimators. Lasso estimators have also been proposed as initial estimates. See for instance Fan et al. [2014a] and Chapter 3 of this thesis, in the context of high dimensional penalized likelihood and robust quasilikelihood estimation respectively. Note that for \( w_j = 1/|\theta_j^{(0)}| \) the usual convention is that for \( \theta_j^{(0)} = 0 \) we define \( w_j = \infty \) and \( \infty \cdot 0 = 0 \). Hence for this choice of \( w_j \), a coefficient set to zero in the first step will never appear in \( T(F) \).

4.4 Numerical illustrations

4.4.1 Orthogonal design

In the special case of the linear model with orthogonal design the asymptotic bias of the lasso, group lasso and adaptive lasso estimators under a contamination neighborhood \((1 - \epsilon)F + \epsilon G\) can be computed explicitly. In particular, it can be seen that if \( h = \int X_j Y d(G - F) \) is bounded and \( \epsilon \) is sufficiently small, the bias of \( j \)th component of the lasso estimator \( \hat{\theta}_j \) is very well approximated by the limiting influence function. A similar conclusion can be reached for the bias of the group and adaptive lassos.

We illustrate this point in an orthogonal design linear model simulation study. Specifically, we simulated a linear model with 9 covariates of the form \( y_i = x_{i1} + x_{i2} + x_{i3} + u_i \), where \( u_i \sim (1 - b) \cdot \mathcal{N}(0, 1) + b \cdot \mathcal{N}(0, 10), b \sim \text{Bin}(1, \epsilon) \) and \( i = 1, \ldots, 200 \). Figure 4.1 illustrates how the influence function approximates the asymptotic bias of \( \hat{\beta}_1 \) for \( \epsilon \in (0, 0.2) \) and tuning parameter \( \lambda = \{0.2, 1.2, 1.6, 2\} \). For the group lasso, three groups of variables of dimension 3 were considered. The least squares estimator was used to construct the weights of the adaptive lasso.

We see that the linearized asymptotic bias obtained from the influence function is quite accurate for small values of \( \epsilon \), especially for small and large values of \( \lambda \). Given that
the influence function gives an insensitive approximation to the asymptotic bias for the coefficients estimated as 0, the accuracy of this approximation will depend on how much contamination the estimator can absorb before the estimated null coefficients become non null. If a penalized estimator is insensitive to contamination in terms of the model selected, then the linearized asymptotic bias will be exact. This question seems to be related to the definition of qualitative model selection robustness proposed by Machado [1993] and is beyond the scope of this work.

Note that in theory the $\sqrt{n/(k \log p)}$ rates of convergence of penalized estimators are obtained for tuning parameters of order $O(\sqrt{n^{-1} \log p})$ (Loh and Wainwright [2015]). Therefore in practice we expect to choose small values of $\lambda$. Figure 4.2 further stresses this point. It provides another picture of plots (a) and (b) showed in Figure 4.1 by taking into account the distance of the lasso functional and the true value of the parameter $\beta_1$. It is clear from (b) that $\lambda = 1.2$ is already too large a tuning parameter.
4.4.2 High dimensional Poisson regression

We consider now a more sophisticated example where an analytic expression of the asymptotic bias cannot be computed. We show instead how the estimators behave in terms of $L^2$-loss, i.e. $\|\hat{\beta} - \beta\|_2$ when the contamination increases. We simulated a Poisson regression model with canonical link $g(\mu_i) = \log \mu_i = 1.8x_{i1} + x_{i2} + 1.5x_{i5}$, $n = 1, \ldots, 100$. The covariates $x_{ij}$ were generated from standard uniforms with correlation $\text{cor}(x_{ij}, x_{ik}) = \rho^{|j-k|}$ and $\rho = 0.5$ for $j, k = 1, \ldots, 250$. The response variables $Y_i$ were generated according to a perturbed Poisson distribution of the form $(1-b) \cdot \mathcal{P}(\mu_i) + b \cdot \mathcal{P}(\nu \mu_i)$ where $b \sim \text{Bin}(1, \epsilon)$. We set $\nu = 5, 10$ and $\epsilon = \{0, 0.025, 0.05, 0.075, 0.1\}$. For each combination of $\epsilon$ and $\nu$ we generated 100 data sets. The tuning parameter was chosen by 5-fold cross-validation. We consider classical and robust counterparts of the lasso and the adaptive lasso based on the robust quasilikelihood of Cantoni and Ronchetti [2001] as in Chapter 3. For the robust losses we use Huber’s $\psi_c$ function with $c = 1.5$. The estimators where computed with the coordinate descent algorithm described in Chapter 3.

Figures 4.3 and 4.4 show that the robust penalized estimators are stable under moderate contamination whereas their classical counterparts are not. When $\nu = 10$ even very small contamination completely ruin the performance of the classical procedures. It seems that the breakdown point of this particular robust estimator is somewhere around $\epsilon = 0.075$.

4.5 Connections to distribution theory

Wang et al. [2013] studied the robustness properties of their proposed robust penalized estimator by calculating its finite sample breakdown point and influence function. It can be seen that their derivation of the influence function (Theorem 3) could easily be extended to more general loss functions. In their proof they implicitly use distributions in the sense of Schwartz [1959], since they require the first two derivatives of the absolute
value function. They use \( \text{sgn}(x) \) as first derivative of \(|x|\) and the Dirac delta function \( \delta(x) \) as its second derivative. These derivatives are justified by the theory of distributions. However, working explicitly with the Dirac delta function understood informally as

\[
\delta(x) = \begin{cases} 
+\infty & \text{if } x = 0 \\
0 & \text{otherwise}
\end{cases}
\]

and inverting a matrix containing such an expression, is not fully satisfactory from a formal mathematical standpoint. Interestingly, the expression obtained in Theorem 3 in Wang et al. [2013] is the same as the one we give in Proposition 3. This suggests that a more careful treatment of the problem with a rigorous use of differentiation in the sense of distribution theory will yield the same influence function.

At first glance the theory of distributions seems to provide a natural and rigorous way of tackling nondifferentiable penalties. However, the theory suffers from at least two major drawbacks for the purposes of deriving influence functions with direct computations. The product of two distributions cannot be consistently defined in general. This makes the manipulation of distributions delicate. Furthermore, to the best of our knowledge there is no implicit function theorem for distributions. This closes the door to the derivation of the influence function as the derivative of an implicit function as in Lemma 1. We relegate to the Appendix 4.8.4 an example where we explicitly compute the distributional derivative.

Figure 4.3: The boxplots show the performance measured by the \( L2 \)-loss of classical and robust counterparts of the lasso and adaptive lasso as the contamination increases. In all the simulations we set \( \nu = 5 \).
Figure 4.4: The boxplots show the performance measured by the $L_2$-loss of classical and robust counterparts of the lasso and adaptive lasso as the contamination increases. In all the simulations we set $\nu = 10$.

of the lasso and scad functionals. Extending this approach to more general problems does not look obvious. We can however show that the influence functions derived in Section 2 using the limiting influence function can be viewed as distributional derivatives. Before giving this result in Proposition 6, we need an intermediate result that is interesting on its own and concerns the continuity of $T(F_\epsilon)$ with respect to $\epsilon$. Its proof uses Berge’s maximum theorem and is similar to the proof of Lemma 2.

**Lemma 5**: Under (A1)-(A3), the penalized M-estimator $T(F_\epsilon)$ resulting from the minimization of (4.2) is continuous with respect to $\epsilon \in (-\varepsilon, \varepsilon)$, $\varepsilon > 0$.

**Proposition 6**: Under the assumptions of Proposition 1, the influence function (4.4) of the minimizer $T(F)$ of (4.2) is the distributional derivative of $T(F_\epsilon)$ with respect to $\epsilon$ evaluated at 0.

### 4.6 Discussion

We introduced the idea of calculating the influence function of penalized M-estimators with the help of a sequence of approximating estimators. In Sections 2 and 3 we justified the validity of such an approach and derived the limiting influence functions of general...
penalized M-functionals. In particular, these computations show that the local robustness properties are a direct result of the form and boundedness of the derivative of the loss function. This reflects the intuition that the sources of local instability for M-estimators and their penalized counterparts should be the same. In Section 4 we illustrated via numerical examples the bias problem arising in a contamination neighborhood and its relation to the influence function. Finally, in Section 5 we showed that the influence function can be viewed as a derivative in the sense of distribution theory.

Let us now discuss about the uniqueness of the empirical influence function in high dimensions. It is fairly intuitive to think of an empirical influence function by taking the limiting expression in (4.4) and plugging in \( \hat{F} \) instead of \( F \). Remark A suggests that this approach is valid, in the sense that the obtained expression is unique, even in the presence of multiple minima as long as they are well separated. It is clear that a unique estimator implies a unique influence function. Special care has to be taken when considering high dimensional settings where \( d > n \). In order to illustrate this point, we will first focus on the lasso estimator by reviewing some results summarized in Tibshirani [2013]. We consider more general problems subsequently.

For a given an \( n \) dimensional response vector \( y = (y_1, \ldots, y_n)^T \), a \( n \times d \) design matrix \( X = (x_1, \ldots, x_n)^T \) and a tuning parameter \( \lambda > 0 \), the lasso estimate is defined as

\[
\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda \|\theta\|_1.
\]

It is clear that the solution is unique when \( \text{rank}(X) = p \) since both the loss function and the penalty function are convex. However when \( d > n \) the picture changes drastically. Indeed, the lasso problem admits at least one solution and given that the problem is convex it will have either a unique solution or an uncountable infinite number of solutions. Despite this unpleasant property, the lasso is very useful. Perhaps the best theoretical justification for this is that any lasso solution \( \hat{\theta} \) achieves fast rates of convergence to \( \theta_0 \) in euclidean norm; cf. Negahban et al. [2012] and Loh and Wainwright [2015]. Furthermore any lasso solution gives the same fit \( X\hat{\theta} \). Regarding the sparsity of the solutions, it is worth mentioning that it can be shown that there exists a lasso solution whose active set \( \mathcal{A} \) has size \( |\mathcal{A}| \leq \min\{n, d\} \) and that any two lasso solutions must have the same signs over their common support. Regarding the uniqueness of the solution, a sufficient condition is the continuity of the covariates. For necessary and sufficient conditions see Tibshirani [2013] and Zhang et al. [2014b]. Tibshirani [2013] also considered the more general minimization problem

\[
\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} f(X\theta) + \lambda \|\theta\|_1
\]

and shows that as long as \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is differentiable and strictly convex the properties discussed for the lasso above hold. In particular, when the entries of \( X \) are drawn from a continuous probability distribution on \( \mathbb{R}^{nd} \) the lasso solution is unique.

Many authors have advocated the use of nonconvex sparsity inducing penalty functions. Perhaps the most studied penalty function of this class is the scad proposed by Fan and Li [2001]. Evidently for these type of penalties the uniqueness discussion considered above is not satisfactory.

Fan et al. [2014c] provided a general framework to analyze nonconvex penalized estimators of the form

\[
\hat{\theta} \in \arg\min_{\theta} \ell_n(\theta) + \sum_{j=1}^{d} p_{\lambda}(|\theta_j|),
\]
where $\ell_n(\theta)$ is a convex loss function and $p_\lambda(\cdot)$ a non convex penalty function as those mentioned above. Consider a sparse vector $\theta_0$ i.e. $|I| = k \ll p$ for $I = \{j : \theta_{0j} \neq 0\}$. The oracle estimator is the estimator that minimizes $\ell_n$ knowing in advance the true support of $\theta_0$. It is defined as

$$\hat{\theta}_{\text{oracle}} = (\hat{\theta}_{\text{oracle}}, 0) = \arg\min_{\theta : \theta_{Ic} = 0} \ell_n(\theta) \quad (4.7)$$

and assumed to be unique. Fan et al. [2014c] show that under some conditions the oracle estimator is attainable with high probability, in one iteration step of the local linear algorithm proposed by with Zou and Li [2008] when the lasso is used as the initial estimator. Since the local linear algorithm iteratively solves weighted lasso problems at each step to solve (4.7), the uniqueness related properties discussed above carry over to this class of penalized estimators. Note also that Loh and Wainwright [2015] showed that under some regularity conditions, any solution of nonconvex penalized problem achieves fast rates of convergence to $\theta_0$ in euclidean norm.

Although in theory the convergence to the oracle avoids the disturbing problem of potentially obtaining multiple solutions with different supports, in practice the tuning parameter $\lambda$ has to be selected appropriately to reach the oracle. This can turn out to be a tricky task and there are many articles in the literature specifically devoted to this issue. See for instance the work of Fan and Tang [2013], Flynn et al. [2013], Hui et al. [2015] and references thereof.

At a somehow more basic level there is another major issue regarding the empirical influence function when $d > n$. Namely, in such a set up (A1)-(A2) do not hold. Hence the intuitive empirical influence function cannot be viewed as a limiting derivative at $\hat{F}$, but simply as a plug-in expression. Still, it could be useful for asymptotic considerations. Indeed in the case of Fisher consistent M-estimators, substituting $G$ by the empirical distribution $\hat{F}$ in (4.1), we have

$$\sqrt{n}\left(T(\hat{F}) - T(F)\right) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF(z_i; F, T)$$

because $\int IF(z; F, T)dF(z) = 0$. Then by the central limit theorem $\sqrt{n}\left(T(\hat{F}) - T(F)\right)$ is asymptotically $N(0, V(F, T))$ where $V(F, T) = \int IF(z; F, T)IF^T(z; F, T)dF(z)$. A rigorous general argument can be found in Huber [1967]. For M-estimators, the conditions for Fréchet differentiability of Clarke [1986] guarantee the validity of the von Mises expansion and imply good robustness properties as discussed in Bednarski [1993]. Using these formulas and the influence functions derived in Sections 2 and 3, we see that for penalized M-estimators, the approximation (4.1) leads to the expression

$$\sqrt{n}(\hat{\theta} - \theta^\ast) \rightarrow_d \mathcal{N}(0, V(F, T)),$$

where $V(F, T)$ is block diagonal with 0 entries for the elements corresponding to $\theta_2$. All the analysis in Sections 2 and 3 was developed for a fixed $\lambda$. If we assume that there is a true underlying set of parameters $\theta_0$, this implies in general that $\theta^\ast \neq \theta_0$. Note however that if we work instead with a $\lambda_n$ tending to zero at an appropriate rate, the heuristic asymptotic distribution would match the oracle properties (Fan and Li [2001]) for the class of lasso type estimators satisfying such properties. Examples of this type of estimators were derived for instance in Fan and Lv [2011] and in our Chapter 3. The von Mises expansion can be therefore used to assess informally the asymptotic properties of an important class of penalized M-estimators. A more careful study of this phenomenon
is required for a better understanding of the conditions under which it holds and is left for future research. Still, as pointed out in Hampel et al. [1986], p.85: “...it is usually easier to verify the asymptotic normality in another way instead of trying to assess the necessary regularity conditions to make this approach rigorous”.

4.7 Proofs

Proof of Lemma 2

Let $b(p) = \inf_{\theta \in \Theta} \Lambda_\theta(\theta; F, p)$ and $f(\theta, p) = b(p) - \Lambda_\theta(\theta; F, p)$. The mapping $\Gamma : W^{2,2} \mapsto \Theta$, $\Gamma p = \{\theta | \theta \in \Theta, f(\theta, p) \leq 0\}$ is closed by construction (Berge [1997], Example, p.111). Therefore $\Gamma p$ is compact for any $p$, which implies that $\Gamma$ is continuous (Berge [1997], Example p.109). From Proposition 7 in the Appendix, $M(p) = \max\{-\Lambda_\theta(\theta; F, p) | \theta \in \Gamma p\}$ is continuous in $W^{2,2}$ and the mapping $\Phi p = \{\theta | \theta \in \Gamma p, -\Lambda_\theta(\theta; F, p) = M(p)\}$ is upper hemicontinuous from $W^{2,2}$ to $\Theta$. Since $\theta_m = \phi p_m = \{\theta | \theta \in \Phi p_m, -\Lambda_\theta(\theta; F, p) = \sup M(p_m)\}$ is single valued and upper hemicontinuous, it is continuous. Therefore $\lim_m \theta_m = \theta^* = \phi p = \lim_m \phi p_m$.  

Proof of Proposition 1

For ease of notation we will write $\psi = \psi(Z; T(F; p_m))$, $S_p = E_F[\psi(Z; T(F; p_m))] + \nabla^2 p_m(T(F; p_m))$, $D = F - \Delta_{\bar{z}}$, and $IF(p_m) = IF(p_m(z; F,T))$. Further let $\{p_m\}_{m \geq 1}$ and $\{p'_m\}_{m \geq 1}$ be to sequences in $C^\infty(\Theta)$ converging to $p$ in $W^{2,2}(\Theta)$. Then Lemma 1 and (A3) guarantee that

$$\text{IF}(p'_m) - \text{IF}(p_m) = S_p^{-1}E_D[\psi_p] - S_{p'}^{-1}E_D[\psi_{p'}] = S_{p'}^{-1}E_D[\psi_p - \psi_{p'}] + S_p^{-1}(S_p - S_{p'})S_p^{-1}E_D[\psi_p] + o(||S_p - S_{p'}||).$$

By Lemma 2 we have $\lim_m T(F; p'_m) = \lim_m T(F; p_m) = T(F)$. Therefore $E_D[\psi_p - \psi_{p'}] \rightarrow 0$ and $||S_p - S_{p'}|| \rightarrow 0$. Hence (A3) and $p_m = \lim_p p'_m = p$ give $\lim_m \left[\text{IF}(p_m) - \text{IF}(p'_m)\right] = 0$.

Proof of Proposition 2

From Proposition 1 it suffices to show that the limiting influence function of a smooth approximation of the problem has the desired form. One possible infinitely differentiable approximation for the absolute value is

$$s_m(t) = \frac{2}{m} \log(e^{tm} + 1) - t \xrightarrow{m \rightarrow \infty} |t|.$$

Its first two derivatives have the form

$$s'_m(t) = \frac{2e^{tm}}{e^{tm} + 1} - 1 \xrightarrow{m \rightarrow \infty} \text{sgn}(t) = \begin{cases} -1 & \text{if } t < 0, \\ +1 & \text{if } t > 0, \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

and

$$s''_m(t) = \frac{2me^{tm}}{(e^{tm} + 1)^2} \xrightarrow{m \rightarrow \infty} \begin{cases} 0 & \text{if } t \neq 0, \\ +\infty & \text{otherwise}. \end{cases} \tag{4.9}$$
Defining $p_m(\theta) = \sum_{j=1}^{p} p_{\lambda,j}(s_m(\theta_j))$, from Lemma 1 and the notation of Proposition 1 we have

$$\text{IF}_{p_m}(z; T, F) = -S_{p_m}^{-1}\left(\psi(z; T(F; p_m)) + \nabla p_m(T(F; p_m))\right)$$

Remember that for a partitioned matrix $A$, if all the necessary inverses exist, the elements of $A^{-1}$ are

$$A^{11} = (A_{11} - A_{12}A^{-1}_{22}A_{21})^{-1}, \quad A^{22} = (A_{22} - A_{21}A^{-1}_{11}A_{12})^{-1},$$
$$A^{12} = -A^{11}A^{22}, \quad A^{21} = -A^{22}A^{11}.$$  \hspace{1cm} (4.10)

Since (4.9) implies $\nabla^2 p_m(T(F))_{jj} \rightarrow 0$ for $j = 1, \ldots, s$ and $\nabla^2 p_m(T(F))_{jj} \rightarrow \infty$ for $j = s+1, \ldots, d$, (4.10) yields

$$S_{p_m}^{-1} \overset{m \to \infty}{\longrightarrow} \begin{bmatrix} E_F[\psi_{11}(Z, T(F))] + \nabla^2 p_m(T(F)) & E_F[\psi_{12}(Z, T(F))] \\ E_F[\psi_{21}(Z, T(F))] & E_F[\psi_{22}(Z, T(F))] + \nabla^2 p_m(T(F)) \end{bmatrix}^{-1} \begin{bmatrix} (M_{11} + P_\lambda)^{-1} & 0 \\ 0 & 0 \end{bmatrix} = S^{-1}.  \hspace{1cm} (4.11)$$

Hence from (4.8), (4.9) and (4.11), it is easily seen that $\lim_m \text{IF}_{p_m}(z; F, T) = \text{IF}(z; F, T)$ has the claimed form. \hspace{1cm} \square

**Proof of Proposition 3**

The difficulty of deriving the influence function for group lasso type penalties comes from the fact that $\sqrt{t}$ is not differentiable. Following the proof of Proposition 2, we will approximate $\sqrt{t}$ by $\sqrt{s_m(t)}$, $t \geq 0$. This yields

$$\left(\sqrt{s_m(t)}\right)' = \frac{e^{tm} - 1}{(e^{tm} + 1)s_m^2(t)} \overset{m \to \infty}{\longrightarrow} \begin{cases} t^{-1/2} & \text{if } t > 0 \\ \infty & \text{for } t \downarrow 0 \end{cases}  \hspace{1cm} (4.12)$$

and

$$\left(\sqrt{s_m(t)}\right)'' = \frac{2me^{tm}}{(e^{tm} + 1)^2s_m^2(t)} - \frac{(e^{tm} - 1)^2}{2(e^{tm} + 1)^2s_m^2(t)} \overset{m \to \infty}{\longrightarrow} \begin{cases} -t^{-3/2} & \text{if } t > 0 \\ -\infty & \text{for } t \downarrow 0 \end{cases}.  \hspace{1cm} (4.13)$$

Note that after some simplifications

$$\left(\left(\sqrt{s_m(t)}\right)''\right)^{-1}\left(\sqrt{s_m(t)}\right)' = \frac{s_m(t)(e^{tm} + 1)(e^{tm} - 1)}{4me^{tm}s_m(t) - (e^{tm} - 1)^2} \overset{m \to \infty}{\longrightarrow} -t, \ t \geq 0.  \hspace{1cm} (4.13)$$

Therefore for the penalty $p_m(\theta) = \sum_{g=1}^{G} p_{\lambda,g}(\sqrt{s_m(\|\theta_{(g)}\|_2^2)})$ we have $\nabla^2 p_m, g(T(F)_{(g)})_{jj} \rightarrow \infty$ for $j = 1, \ldots, |g|$ if $T(F)_{(g)} = 0$ and

$$\left(E_F[\psi_{22}(Z, T(F))] + \nabla^2 p_m(T(F))\right)^{-1}\nabla^2 p_m(T(F)) \overset{m \to \infty}{\longrightarrow} 0.  \hspace{1cm} (4.14)$$

The same argument used in (4.11) gives

$$S_{p_m}^{-1} \overset{m \to \infty}{\longrightarrow} \begin{bmatrix} (M_{11} + P_\lambda)^{-1} & 0 \\ 0 & 0 \end{bmatrix} = S^{-1}.  \hspace{1cm} (4.15)$$

Finally, from Lemma 1 we have

$$\text{IF}_{p_m}(z; T, F) = \left(E_F[\psi(Z; T(F; p_m))] + \nabla^2 p_m(T(F; p_m))\right)^{-1} \cdot \left(\psi(z; T(F; p_m)) + \nabla p_m(T(F; p_m))\right).  \hspace{1cm} (4.16)$$

Thus, the claimed results follows using (4.12)-(4.15) when taking the limit of (4.16) as $m \to \infty$. \hspace{1cm} \square
Proof of Proposition 5

From Lemma 4 and Proposition 4 we know that we only need to take a sequence of approximating influence functions. First note that $\nabla_{\theta(0)} E_F[\psi(Z, T(F))] = 0$ and $\nabla_{\theta_j, \theta} w_j | \theta_j = 0$ for $\theta_j \neq 0$. Then an argument similar to the one given in Proposition 2 completes the proof. □

Proof of Proposition 6

By Lemma 5, $T(F_\epsilon)$ is a continuous function of $\epsilon$ in a neighborhood of 0. Therefore $T(F_\epsilon)$ is locally integrable. Lemma 7 in the Appendix tells us that $T(F_\epsilon)$ has a distributional derivative given by the limiting form of the derivatives of $T(F_\epsilon; p_m)$. □

4.8 Appendix

4.8.1 Auxiliary results for Berge’s maximum theorem

Let $\Gamma$ be a mapping (possibly multivalued) of the topological space $X$ to the topological space $Y$. We say that $\Gamma$ is lower hemicontinuous at $x_0 \in X$ if for each open set $O$ meeting $\Gamma x_0$ there is a neighborhood $N(x_0)$ such that $x \in N(x_0) \Rightarrow \Gamma x_0 \cap O \neq \emptyset$. We say that $\Gamma$ is upper hemicontinuous at $x_0 \in X$ if for each open set $O$ containing $\Gamma x_0$ there is a neighborhood $N(x_0)$ such that $x \in N(x_0) \Rightarrow \Gamma x_0 \subset O$. If $\Gamma$ is both lower and upper semi-continuous at $x_0$ we say that $\Gamma$ is continuous at $x_0$.

If $\Gamma$ is lower semi-continuous at each point of $X$ it is called lower hemicontinuous in $X$. We say that $\Gamma$ is upper hemicontinuous in $X$ if it is upper semi-continuous at each point of $X$ and if also $\Gamma x$ is compact for each $x \in X$. If $\Gamma$ is both lower and upper hemicontinuous in $X$, then it is called continuous in $X$.

Note the above definitions of upper and lower-hemicontinuous mappings are called upper and lower-semicontinuous respectively in Berge [1997]. We opted for our terminology as it is frequently used in mathematical economics and it avoids confusion with the usual definition of continuity. Note also that if a function is single valued and upper-hemicontinuous, it is continuous.

**Berge’s maximum theorem** : (Berge [1997], p. 116) If $h$ is a continuous function in $Y$ and $\Gamma$ is a continuous mapping of $X$ to $Y$ such that, for each $x$, $\Gamma x \neq \emptyset$, then the function $M$ defined by $M(x) = \max \{ h(y) | y \in \Gamma x \}$ is continuous in $X$ and the mapping $\phi$ defined by $\phi x = \{ y | y \in \Gamma x, h(x, y) = M(x) \}$ is a upper hemicontinuous mapping of $X$ into $Y$.

**Lemma 6** : (Berge [1997], Theorem 7, p.112 ) If $\Gamma_1$ is a closed mapping of $X$ into $Y$ and $\Gamma_2$ is an upper hemicontinuous mapping of $X$ into $Y$, then the mapping $\Gamma = \Gamma_1 \cap \Gamma_2$ is upper hemicontinuous.

We provide a variant of Berge’s maximum theorem. Its proof follows closely the one given in Berge [1997], p.116 for the original theorem and is given for completeness.

**Proposition 7** : If $h$ is a continuous function in $X \times Y$ and $\Gamma$ is a continuous mapping of $X$ to $Y$ such that, for each $x$, $\Gamma x \neq \emptyset$, then the function $M$ defined by $M(x) = \max \{ h(x, y) | y \in \Gamma x \}$ is continuous in $X$ and the mapping $\phi$ defined by $\phi x = \{ y | y \in \Gamma x, h(x, y) = M(x) \}$ is a upper semi-continuous mapping of $X$ into $Y$. 
The function $h$ is continuous in $\mathcal{X} \times \mathcal{Y}$ and so $M$ is a continuous function. Furthermore the mapping $\Delta$ given by

$$\Delta x = \{ y \mid M(x) - h(x, y) \leq 0 \}$$

is closed (Berge [1997], Example, p.111). Hence by Lemma 6, $\phi = \Gamma \cap \Delta$ is upper hemicontinuous. □

### 4.8.2 Auxiliary results from distribution theory

Define $C_c^\infty(\Omega) = \{ f : \Omega \to \mathbb{R} \mid f \in C^\infty(\Omega), \text{ with compact support} \}$ where $\Omega$ is an open set in $\mathbb{R}^d$.

**Definition:** A sequence $\{\phi_n\}_{n \geq 1}$ of functions $\phi_n \in C_c^\infty(\Omega)$ converges to $\phi \in C_c^\infty(\Omega)$ in the sense of test functions if:

(a) there exists $\Omega' \subseteq \Omega$ such that $\text{supp} \phi_n \subseteq \Omega'$ for every $n \in \mathbb{N}$;

(b) $\partial^n \phi_n \to \partial^n \phi$ as $n \to \infty$ uniformly on $\Omega$ for every multi-index $\alpha \in \mathbb{N}^d$.

The **topological vector space** $\mathcal{D}(\Omega)$ consists of $C_c^\infty(\Omega)$ equipped with the topology that corresponds to convergence in the sense of test functions. A **linear functional on** $\mathcal{D}(\Omega)$ is a linear map $T : \mathcal{D}(\Omega) \to \mathbb{R}$. The value of $T$ acting on a test function $\phi$ is denoted by $\langle T, \phi \rangle$. Therefore if $T$ is linear we have

$$\langle T, \lambda \phi + \mu \psi \rangle = \lambda \langle T, \phi \rangle + \mu \langle T, \psi \rangle, \quad \text{for all } \lambda, \mu \in \mathbb{R} \text{ and } \phi, \psi \in \mathcal{D}(\Omega).$$

A functional $T$ is **continuous** if $\phi_n \to \phi$ in the sense of test functions implies that $\langle T, \phi_n \rangle \to \langle T, \phi \rangle$ in $\mathbb{R}$.

**Definition:** A distribution on $\Omega$ is a continuous linear functional $T : \mathcal{D}(\Omega) \to \mathbb{R}$.

A sequence of distributions $\{T_n\}$ converges to $T$ if $\langle T_n, \phi \rangle \to \langle T, \phi \rangle$ for every $\phi \in \mathcal{D}(\Omega)$. The **topological vector space** $\mathcal{D}'(\Omega)$ consists of the distributions on $\Omega$ equipped with the topology corresponding to this notion of convergence. An important example of distribution is the **Dirac delta function supported at** $x \in \Omega$, which is the distribution $\delta_x : \mathcal{D}(\Omega) \to \mathbb{R}$ defined by $\langle \delta_x, \phi \rangle = \phi(x)$.

**Definition:** For $1 \leq i \leq d$, the $i$th partial derivative of a distribution $T \in \mathcal{D}'(\Omega)$ is the distribution $\partial_i T \in \mathcal{D}'(\Omega)$ defined by $\langle \partial_i T, \phi \rangle = -\langle T, \partial_i \phi \rangle$ for all $\phi \in \mathcal{D}(\Omega)$. For $\alpha \in \mathbb{N}^d$, the derivative $\partial^\alpha T \in \mathcal{D}'(\Omega)$ of order $|\alpha|$ is defined by $\langle \partial^\alpha T, \phi \rangle = (-1)^{|\alpha|} \langle T, \partial^\alpha \phi \rangle$ for all $\phi \in \mathcal{D}(\Omega)$.

We recall that for $1 \leq p < \infty$, the space $L^p(\Omega)$ consists of all integrable functions, i.e. the Lebesgue measurable functions $f : \Omega \to \mathbb{R}$ such that $\int_\Omega |f|^p dx < \infty$. Equipped with the norm norm $\|f\|_p = (\int_\Omega |f|^p dx)^{1/p}$ they constitute a Banach space. Note that any (locally) integrable function defines a regular distribution $T_f \in \mathcal{D}'(\Omega)$ by $\langle T_f, \phi \rangle = \int_\Omega f \phi dx$. Typically the function $f$ and the distribution $T_f$ are regarded as equivalent.

**Lemma 7:** Let $f \in L^1(\Omega)$ and $\{f_n\}$ be a sequence of functions in $C^\infty(\Omega)$ such that $f_n \to f$ and $\partial^\alpha f_n \to g$ in $L^1(\Omega)$. Then $f$ has distributional derivative given by $g = \partial^\alpha f \in L^1(\Omega)$.

**Proof:** Note that since $f_n \to f$ in $L^1(\Omega)$ and $\phi \in C^\infty_c(\Omega)$, we have

$$\int_\Omega f_n \phi dx \xrightarrow{n \to \infty} \int_\Omega f \phi dx.$$
because for $K = \text{supp} \phi$
\[
\left| \int_{\Omega} f_n \phi dx - \int_{\Omega} f \phi dx \right| = \left| \int_{\Omega} (f_n - f) \phi dx \right| \leq \sup_{K} |\phi| \int_{K} |f_n - f| dx \rightarrow 0.
\]
Hence, for every $\phi \in C^\infty_c(\Omega)$, the convergence of $f_n$ and $\partial^\alpha f_n$ implies that
\[
\int_{\Omega} f \partial^\alpha \phi dx = \lim_{n \to \infty} \int_{\Omega} f_n \partial^\alpha \phi dx = (-1)^{|\alpha|} \lim_{n \to \infty} \int_{\Omega} \partial^\alpha f_n \phi dx = (-1)^{|\alpha|} \int_{\Omega} g \phi dx.
\]
Therefore the distributional derivative of $f$ is $\partial^\alpha f = g$. □

### 4.8.3 Additional Proofs

**Proof of Lemma 1**

The penalized M-functional $T(F_c) = \theta_{F_c,\lambda}$ is defined as a root of
\[
H(\epsilon, \theta) := E_F[\psi(Z, \theta)] + \nabla p(\theta; \lambda) = 0.
\]
Note that $H$ has partial derivatives with respect to $\epsilon$ and $\theta$ at $(0, \theta_{F_c,\lambda})$. Therefore the claimed result follows immediately from the implicit function theorem. □

**Proof of Lemma 3**

The existence of $\text{IF}(z^{(1)}; F, S)$ follows from Lemma 1. The penalized M-functional $T(F_c) = \theta_2^*$ is obtained as a root of
\[
H(\epsilon, \theta_2) := E_F[\psi^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1^*)] + \nabla_{\theta_2} p^{(2)}(\theta_2, \theta_1^*; \lambda) = 0.
\]
Since $H$ has partial derivatives with respect to $\epsilon$ and $\theta$ at $(0, \theta_2^*)$, the existence of $\text{IF}(z; F, T)$ follows from the implicit function theorem. The claimed expression is obtained by slightly modifying the computations provided in Zhelonkin et al. [2012] for unpenalized two-stage M-estimators. □

**Proof of Proposition 4**

For ease of notation we will write $(\theta_1^m, \theta_2^m) = (S(F; p_m^{(1)}), T(F; p_m^{(2)}))$, $\psi_p = \psi(z^{(2)}, \theta_2^m, z^{(1)}, \theta_1^m)$, $S_p = E_F[\psi^{(2)}(Z^{(2)}, \theta_2^m, Z^{(1)}, \theta_1^m)] + \nabla_{\theta_2} p^{(2)}(T(F; p_m^{(2)}))$, $D = F - \Delta_x$, $\text{IF}(p_m^{(1)}) = \text{IF}(p_m^{(1)}(z^1; F, S))$ and $\text{IF}(p_m^{(2)}) = \text{IF}(p_m^{(2)}(z; F, T))$. Further let $\{p_m^{(i)}\}_{m \geq 1}$ and $\{p_m^{(i)}\}_{m \geq 1}$ be to sequences in $C^\infty(\Theta_i)$ converging to $p^{(i)}$ in $W^{2,2}(\Theta_i)$ for $i = 1, 2$. By Lemma 3 and (A3)
\[
\text{IF}(p_m^{(2)}) - \text{IF}(p_m^{(2)}) = S_p^{-1} \left( E_D[\psi_p] + \nabla_{\theta_2} E_F[\psi_p] + \nabla_{\theta_2, \theta_1} p^{(2)}(z^{(1)}; F, S) \right) - S_{p'}^{-1} \left( E_D[\psi_{p'}] + \nabla_{\theta_2} E_F[\psi_{p'}] + \nabla_{\theta_2, \theta_1} p^{(2)}(z^{(1)}; F, S) \right) = T_1 + T_2 \text{IF}(z^{(1)}; F, S).
\]
By the arguments given in Proposition 1, $T_1 \rightarrow 0$. In particular $\|S_p - S_{p'}\| \rightarrow 0$. Furthermore from Lemma 4 we have $\lim_m T(F; p_m^{(2)}) = \lim_m T(F; p_m^{(2)}) = T(F)$. Therefore (A3) and $\lim p_m = \lim p'_m = p$ guarantee that $\lim_m \left[ \text{IF}(p_m^{(2)'}) - \text{IF}(p_m^{(2)}) \right] = 0$. □

**Proof of Lemma 4**

For a given first stage estimator the proof is similar to the one given for Lemma 2. □
4.8.4 Example of explicit computations with distributional derivatives

Assume an orthogonal linear model i.e.

\[ y_i = x_i^T \theta + \epsilon_i, \quad i = 1, \ldots, n, \]

where \( \epsilon_i \) are zero mean errors, \( y_i \) the responses and \( x_i \) the covariates with \( X^T X = I \) and \( X = (x_1, \ldots, x_n)^T \). We consider the penalized least squares problem

\[
\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \sum_{j=1}^{d} p_\lambda(|\theta_j|) \tag{4.17}
\]

where \( p_\lambda(\cdot) \) is a penalty function and \( \lambda \) the regularization parameter. For the lasso and scad penalties, the resulting estimators have an explicit solution that allows for easy computations of distributional derivatives. In the orthogonal design linear model the lasso and scad functionals have the explicit forms

\[ \hat{\theta}_{lasso}^j = \text{sgn}(\theta_j^{(0)})(|\theta_j^{(0)}| - \lambda)_+, \]

and

\[
\hat{\theta}_{scad}^j = \begin{cases} 
\text{sgn}(\theta_j^{(0)})(|\theta_j^{(0)}| - \lambda)_+ & \text{if } |\theta_j^{(0)}| < 2\lambda \\
(a - 1)\theta_j^{(0)} - \text{sgn}(\theta_j^{(0)})a\lambda / (a - 2) & \text{if } 2\lambda \leq |\theta_j^{(0)}| \leq a\lambda \\
\theta_j^{(0)} & \text{otherwise}
\end{cases}
\]

where \((x)_+ = \max\{0, x\}\) and \(a > 2\). The functional form of the least squares estimator is very simple and the \( j \)th coefficient at \( F_\epsilon \) is

\[ \theta_j^{(0)}(\epsilon) = f(\epsilon) = \int X_j Y dF_\epsilon = A + \epsilon(B - A), \]

where \( A = E_F[X_j Y] \) and \( B = x_j y \). Recall that a distribution is a linear functional on the space \( D \) of test functions, i.e. the space of infinitely differentiable functions with compact support. The Dirac delta function, for instance, is the linear functional defined by \( \langle \delta, \phi \rangle = \phi(0), \phi \in D \). The distributional derivative of a function \( g : \mathcal{R} \to \mathcal{R} \) is defined as

\[ \langle g', \phi \rangle = -\langle g, \phi' \rangle = -\int_\Omega g(x)\phi'(x)dx, \]

where \( \Omega \) is the support of \( \phi \) and \( \phi' \) denotes the derivative of \( \phi \).

Let’s first compute the influence function of the lasso. The distributional derivative of \( \hat{\theta}_{lasso}^j \) is by definition

\[
\langle \nabla \hat{\theta}_{lasso}^j, \phi \rangle = -\int \text{sgn}(|f(\epsilon)|)(|f(\epsilon)| - \lambda)_+ \phi'(\epsilon) d\epsilon. \tag{4.18}
\]
In order to obtain an expression for the influence function of the lasso estimator, it suffices to integrate (4.18) over a small interval $[0, \varepsilon]$ because we are only interested in the local behavior of the distributional derivative. Clearly for $f(0) = 0$, $\lambda > 0$ and $\varepsilon$ small enough

$$\langle \nabla \hat{\theta}_j^{\text{lasso}}, \phi \rangle = 0.$$  

(4.19)

When $f(0) > 0$ and $\varepsilon > 0$ is the smallest value such $f(\varepsilon) = 0$, we have

$$\langle \nabla \hat{\theta}_j^{\text{lasso}}, \phi \rangle = -\int_{[0, \varepsilon]} (f(\varepsilon) - \lambda)\phi'(\varepsilon)d\varepsilon = -\phi(\varepsilon)(f(\varepsilon) - \lambda)\bigg|_0^\varepsilon + \int_{[0, \varepsilon]} f'(\varepsilon)\phi(\varepsilon)d\varepsilon$$

$$= -\phi(\varepsilon)(f(\varepsilon) - \lambda) + \phi(0)(f(0) - \lambda) - (A - B)\int_0^\varepsilon \phi(\varepsilon)d\varepsilon.$$ 

Hence for $\varepsilon' \in [0, \varepsilon]$ we obtain

$$\nabla \hat{\theta}_j^{\text{lasso}}(\varepsilon') = -(A - B) - \phi(\varepsilon')(f(\varepsilon') - \lambda)\bigg(\delta(\varepsilon') - \delta(0)\bigg).$$  

(4.20)

An analogous argument yields for $f'(0) < 0$ and $\varepsilon'$ small enough

$$\nabla \hat{\theta}_j^{\text{lasso}}(\varepsilon') = -(A - B) + \phi(\varepsilon')(f(\varepsilon') + \lambda)\bigg(\delta(\varepsilon') - \delta(0)\bigg).$$  

(4.21)

Noticing that $A = -\lambda \text{sgn}(\hat{\theta}_j^{\text{lasso}})$ when $\hat{\theta}_j^{\text{lasso}} \neq 0$, we conclude from (4.19)-(4.21) that

$$\nabla \hat{\theta}_j^{\text{lasso}}(0) = \begin{cases} 0 & \text{if } |\theta_j^{(0)}| \leq \lambda \\ \psi(z, \hat{\theta}_j^{\text{lasso}}) + \lambda \text{sgn}(\hat{\theta}_j^{\text{lasso}}) & \text{otherwise.} \end{cases}$$

Let us now turn to the derivation of the influence function of the scad estimator. It is clear from the form of the $\hat{\theta}_j^{\text{scad}}$ that $\nabla \hat{\theta}_j^{\text{scad}}(0)$ has the same form as $\nabla \hat{\theta}_j^{\text{lasso}}(0)$ when $|\theta_j^{(0)}| < 2\lambda$. It is also obvious that $\nabla \hat{\theta}_j^{\text{scad}}(0) = \psi(z, \hat{\theta}_j^{\text{scad}})$ when $|\theta_j^{(0)}| > a\lambda$. We therefore only have to consider the case $2\lambda \leq |\theta_j^{(0)}| \leq a\lambda$. Suppose that $f(0) > 0$ and denote by $\varepsilon$ the smallest positive real number such that $f(\varepsilon) = 0$. Then

$$\langle \nabla \hat{\theta}_j^{\text{scad}}, \phi \rangle = -\int_{[0, \varepsilon]} \frac{(a - 1)f(\varepsilon) - a\lambda}{a - 2}\phi'(\varepsilon)d\varepsilon$$

$$= -\frac{a - 1}{a - 2}f(\varepsilon)\phi(\varepsilon)\bigg|_0^\varepsilon + \frac{a - 1}{a - 2}\int_{[0, \varepsilon]} f'(\varepsilon)\phi(\varepsilon)d\varepsilon - \frac{a\lambda}{a - 2}\phi(\varepsilon)\bigg|_0^\varepsilon$$

$$= \frac{a - 1}{a - 2}(A - B)\int_{[0, \varepsilon]} \phi(\varepsilon)d\varepsilon + \frac{a - 1}{a - 2}\bigg(f(0)\phi(0) - f(\varepsilon)\phi(\varepsilon)\bigg)$$

$$+ \frac{a\lambda}{a - 2}\bigg(\phi(0) - \phi(\varepsilon)\bigg).$$

Hence for sufficiently small $\varepsilon'$ we have

$$\nabla \hat{\theta}_j^{\text{scad}}(\varepsilon') = \frac{a - 1}{a - 2}(A - B) + \frac{a - 1}{a - 2}f(\varepsilon')\phi(\varepsilon')\bigg(\delta(0) - \delta(\varepsilon')\bigg)$$

$$+ \frac{a\lambda}{a - 2}\phi(\varepsilon')\bigg(\delta(0) - \delta(\varepsilon')\bigg)$$

and thus

$$\nabla \hat{\theta}_j^{\text{scad}}(0) = \frac{a - 1}{a - 2}\bigg(\psi(z, \hat{\theta}_j^{\text{scad}}) + p'_\lambda(\hat{\theta}_j^{\text{scad}})\bigg).$$
The same expression is obtained when \( f(0) < 0 \). Remember that for \( t > 0 \) the scad penalty is defined by

\[
p_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a - 1)\lambda} I(t > \lambda) \right\}.
\]

Hence for \( 2\lambda \leq t \leq a\lambda \) we have \( 1 + p''_\lambda(t) = 1 + 1/(a - 1) = (a - 2)/(a - 1) \). We have therefore established via direct calculations, the following corollary of Proposition 6.

**Corollary 1:** Let \( \theta^{(0)} = T^{(0)}(F) \) denote the least square estimator of \( \theta \). Then the influence function of the functional \( \theta^* = T(F) \) associated with the minimizer of (4.17) exists as a distributional derivative for the lasso and scad penalties, and an explicit computation yields

\[
IF(z; F, T) = \begin{cases} 
0 & \text{if } |\theta^{(0)}_j| \leq \lambda \\
\psi_j(z, \theta^*) + \lambda \sgn(\theta^*_j) & \text{otherwise}
\end{cases}
\]

and

\[
IF(z; F, T) = \begin{cases} 
0 & \text{if } |\theta^{(0)}_j| \leq 2\lambda \\
\left( \psi_j(z, \theta^*) + p'_\lambda(|\theta^*_j|) \sgn(\theta^*_j) \right) / \left( 1 + p''_\lambda(|\theta^*_j|) \right) & \text{otherwise}
\end{cases}
\]

respectively, where \( z = (y, x) \), \( \psi(z, \theta) = (y - x^T \theta^*) x_j \), and \( p'_\lambda(t) \) and \( p''_\lambda(t) \) denote the first two derivatives of \( p_\lambda(t) \) for \( t > 0 \).
Chapter 5

Robust penalized M-estimators for generalized additive models

General remarks

We introduce new estimation and variable selection techniques in the flexible semiparametric context of generalized additive models. Our proposal builds on the group lasso and is valid in high dimensional settings where the number of covariates is larger than the sample size. Furthermore our method estimates simultaneously the scale parameter and is robust towards small deviations from the model assumptions. We justify our methods theoretically by establishing asymptotic theory that shows consistency for the group lasso estimator and oracle properties of our adaptive group lasso estimator. Furthermore, we show that our sparse estimators have a bounded influence function. Formally showing the result requires the framework introduced in Chapter 4. This allows us to define rigorously the influence function via a sequence of approximating functionals and generalizes the notion of influence function in a natural way to semiparametric penalized M-functionals. We illustrate both the consistency and robustness of our proposal in a simulation study.

5.1 Introduction

Many modern data problems include a large number of variables $d$ that can be larger than the sample size $n$. This has made high dimensional statistics a very active area of research. In this paper, we focus on the high dimensional set up for the family of Generalized Additive Models (GAM), a popular class of statistical methods for modeling continuous and discrete data in a flexible semiparametric way. They model a properly transformed mean response as a sum of smooth functions of individual covariates (Hastie and Tibshirani [1990], Wood [2006]). Our goal is to provide valid estimation techniques for GAM in high dimensional settings that are not too sensitive to the presence of deviations from the model. In this sense our work can be viewed as an extension of the robust statistics tools (Huber [1981] and 2nd edition by Huber and Ronchetti [2009], Hampel et al. [1986] and Maronna et al. [2006]), that have been mainly focused on low dimensional statistics, to the more modern high dimensional framework.

Numerous previous works have also tried to bridge some notions of robustness to high dimensional regression set ups. Here we will limit ourselves to point out only a few representative contributions that are most closely related to our proposal. Since
the lasso of Tibshirani [1996] was introduced, a large body of the literature emerged establishing good empirical and theoretical properties of \( \ell_1 \) penalized regression in high dimensions. Our methods can be seen as one possible extension of \( \ell_1 \) penalized procedures to GAM. From this perspective our work is most closely related to GAMlet (Sardy and Tseng [2004]), COSSO (Lin and Zhang [2006]), SpAM (Ravikumar et al. [2009]) and the penalized maximum likelihood of Meier et al. [2009]. Alternatively our methods can also be seen as a robustification of likelihood based penalized procedures. From this other perspective, our work is more related to the work of Sardy et al. [2001], Wang et al. [2007a], Li et al. [2011], Alfons et al. [2013], Wang et al. [2013] and Fan et al. [2014a].

This paper provides three main contributions to the statistical literature. First, we show the asymptotic validity of our proposal for GAM in high dimensions. These type of results had only been established for additive models in previous works. In particular, we extend the asymptotic results of Huang et al. [2010] from the additive model setting to GAM, for both classical and robust estimators. As a side product of the derivation we establish consistency results for a group lasso estimator and oracle properties for an adaptive version of it. Furthermore, unlike the analysis of Huang et al. [2010] for additive models, we can allow the number of nonzero smooth functions of the covariates to diverge with the sample size. Second, we extend the consistency results for the estimation of the scale parameter obtained in Sun and Zhang [2012]. We generalize their construction, which was proposed for penalized least squares problems, to the wider class of classical and robust GAM. This result is interesting on its own as the estimation of the scale parameter is a challenging problem in high dimensional set ups (Fan et al. [2012], Reid et al. [2016]). From a robustness standpoint this result is crucial and had not been formally addressed. Indeed, in robust regression one of the key ideas that yields robustness towards outliers is to bound the effect large standardized residuals (Huber and Ronchetti [2009]). Without proper scaling, “Huberizing” a loss function in the hope of achieving robustness can be either highly inefficient or yield disappointingly poor results in terms of robustness. Therefore, in practice it is necessary to either estimate the scale parameter or know its value before hand which in many statistical applications is a very unrealistic assumption. Third, we introduce a rigorous definition of the influence function for semiparametric penalized M-estimators and show that our proposals are robust from this standpoint. We achieve this by adapting the parametric construction of Chapter 3 to our set up. The main result regarding the validity of the derivation of the influence function via the computation of the limit of an approximating sequence is of general interest. Indeed, this approach allows for an easy assessment of the robustness properties of a penalized M-estimator.

In the next section we will set up the notation and describe some robust fitting procedures that will serve as building blocks for the construction of our main estimators described in Section 3. We establish consistency and oracle properties for our main proposal in Section 4. In Section 5 we introduce the limiting influence function for sparse GAM and show that our procedure is bounded influence robust. Practical issues regarding the implementation of our methods and a numerical example are given respectively in Section 6 and 7. All the proofs are given in the Appendix.
5.2 Notation and background

In GAM we assume that we observe independent copies \((X_1, Y_1), \ldots, (X_n, Y_n)\) of the pair \((X, Y)\), where \(X \in [0, 1]^d\) and the conditional responses \(Y_i|X_i\) for \(i = 1, \ldots, n\) are drawn from a distribution of the type

\[
f(y_i; \theta_i, \phi^2) = \exp \left( \frac{y_i \theta_i - b(\theta_i)}{\phi^2} + c(y_i, \phi^2) \right), \quad \forall i = 1, \ldots, n
\]

where \(a(\cdot), b(\cdot)\) and \(c(\cdot)\) are some specific functions and \(\phi\) a nuisance parameter. Moreover, writing \(E[Y_i|X_i] = \mu_i, \ V[Y_i|X_i] = V(\mu_i)\) we have that

\[
g(\mu_i) = \eta_0(X_i) + \alpha + \sum_{j=1}^p f_j(X_{ij}),
\]

where \(X_i \in \mathbb{R}^p\) is the set of explanatory variables, \(g(\cdot)\) the link function, \(f_j : \mathbb{R} \mapsto \mathbb{R}, \ j = 1, \ldots, p\) are unspecified but smooth univariate functions and \(\alpha\) is a constant.

We denote by \(\nu\) the distribution of \(X_i\) and \(\nu_j\) the marginal distribution of \(X_{ij}\) for each \(j = 1, \ldots, d\). For a function \(f_j\) on \([0, 1]\) we denote its \(L_2(\nu_j)\) and \(L_2\) norm respectively by

\[
\|f_j\|_{\nu_j} = \sqrt{\int_{[0,1]} f_j^2(x)d\nu_j(x)} = \sqrt{\mathbb{E}\{f_j^2(X_{ij})\}} \quad \text{and} \quad \|f_j\|_2 = \sqrt{\int_{[0,1]} f_j^2(x)d(x)}
\]

For simplicity, when the variable \(X_{ij}\) is clear from the context we write \(\|f_j\|\) instead of \(\|f_j\|_{\nu_j}\). For \(j \in \{1, \ldots, d\}\), let \(\mathcal{H}_j\) denote the Hilbert subspace \(L_2(\nu_j)\) of measurable functions \(f_j(x_j)\) of a the scalar variable \(x_j\) with zero mean i.e. \(\mathbb{E}\{f_j(X_{ij})\} = 0\). Therefore \(\mathcal{H}_j\) has the inner product

\[
(f_j, f'_j) = \mathbb{E}\{f_j(X_{ij})f'_j(X_{ij})\}
\]

and \(\|f_j\| = \sqrt{\mathbb{E}f_j^2(X_{ij})} < \infty\). Let \(\mathcal{H}_0\) be the space of square integrable constant functions on \([0, 1]\) and denote by \(\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1 \oplus \ldots \oplus \mathcal{H}_d\) the Hilbert space of functions of \(x = (x_1, \ldots, x_d) \in \mathbb{R}^d\) that have the additive form \(m(x) = \alpha + \sum_{j=1}^d f_j(x_j)\) with \(f_j \in \mathcal{H}_j, \ j = 1, \ldots, d\) and \(\alpha \in \mathcal{H}_0\). Note that \(\|f\|_\mathcal{H} = \min_{j=1,\ldots,d} \|f_j\| + \cdots + \|f_d\|\).

Let \(\{\varphi_{jk}, k = 0, 1, \ldots\}\) denote a uniformly bounded, orthonormal basis with respect to \(L_2[0, 1]\). Unless stated otherwise, we assume that \(f_j \in \mathcal{T}_j\) where

\[
\mathcal{T}_j = \left\{ f_j \in \mathcal{H}_j : f_j(x_j) = \sum_{k=1}^\infty \beta_{jk}\varphi_{jk}(x_j), \sum_{k=0}^\infty \beta^2_{jk} k^4 \leq C^2 \right\}
\]

for some \(0 < C < \infty\). We denote by \(\lambda_{\min}(A)\) and \(\lambda_{\max}(A)\) the minimum and maximum eigenvalues of a square matrix \(A\). For a vector \(v \in \mathbb{R}^m\) we use the norms

\[
\|v\|_2 = \sqrt{\sum_{j=1}^k v_j^2}, \quad \|v\|_1 = \sum_{j=1}^k |v_k|, \quad \|v\|_\infty = \max_j |v_j|.
\]

We now describe an approach to robust estimation for generalized linear models (GLM) that have served as a building block for the construction of robust generalized additive models. We will build on the latter to propose our variable selection strategy.
5.2.1 A class of robust M-estimators for generalized linear models

GLM can be viewed as a special case of GAM when the smooth functions \( f_j \) in (5.2) are assumed to be linear. Therefore the model becomes

\[
g(\mu_i) = \eta_0(X_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}.
\]

The classical approach to GLM based on the likelihood and quasilikelihood (Wedderburn [1974], McCullagh and Nelder [1989]) leads to estimators and tests that are sensitive to small deviations from the underlying distributional assumptions. Robust alternatives have been studied extensively in the literature. For our purpose, we concentrate on a class of M-estimators proposed by Cantoni and Ronchetti [2001] which maximize with respect to \( \beta \) a robust quasilikelihood defined by:

\[
Q(\beta) = Q_M(y, \mu) = \frac{1}{n} \sum_{i=1}^{n} Q_M(y_i, \mu_i),
\]

where

\[
Q_M(y_i, \mu_i) = -\left( \int_{\tilde{s}}^{r_i} \nu(y_i, t)w(x_i)dt - \frac{1}{n} \sum_{j=1}^{n} \int_{\tilde{t}}^{r_j} E\left[ \nu(y_j, t) \right]w(x_j)dt \right),
\]

\[
\nu(y_i, \mu_i) = \psi(r_i) \frac{1}{\nu(\mu_i)^{1/2}}, \quad r_i = (y_i - \mu_i)/\nu(\mu_i)^{1/2}, \quad \tilde{s} \text{ such that } \psi((y_i - \tilde{s})/\nu(\tilde{s})^{1/2}) = 0,
\]

and \( \tilde{t} \) such that \( E\left[ \psi((y_i - \tilde{s})/\nu(\tilde{s})^{1/2}) \right] = 0 \). The function \( \psi(\cdot) \) is bounded and limits the influence of outlying points in the space of the response variables and \( w(\cdot) \) is a function which downweights possible leverage points in the covariates. Taking the derivative with respect to \( \beta \) in (5.4) gives the estimating equation

\[
\frac{1}{n} \sum_{i=1}^{n} \Psi_i(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \psi(r_i) \frac{1}{\nu(\mu_i)^{1/2}} w(x_i) \frac{\partial \mu_i}{\partial \beta} - a(\beta) \right] = 0,
\]

where \( a(\beta) = \sum_{i=1}^{n} E[\psi(r_i)] \frac{1}{\nu(\mu_i)^{1/2}} w(X_i) \partial \mu_i / \partial \beta \) is a constant ensuring Fisher consistency. The solution of (5.5) is an M-estimator with influence function proportional to \( \Psi_i \). Since this last expression is bounded in both the covariates space and the response space, the influence function of the robust GLM estimator is bounded and hence robust in the infinitesimal robustness sense of Hampel et al. [1986].

Note that the solution of (5.5) can be found using an iteratively reweighted least squares algorithm

\[
||\sqrt{W} (z - \eta)||^2_2
\]

with weights \( W = \text{diag}(W_1, \ldots, W_n) \) and pseudo data \( z = (z_1, \ldots, z_n)^T \),

\[
W_i = \frac{1}{n} E[\psi(r_i)r_i] V(\mu_i)^{-1} w(x_i) \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 \quad \text{and}
\]

\[
z_i = \eta_i + \frac{\psi(r_i) - E[\psi(r_i)]}{E[\psi(r_i)r_i]} V(\mu_i)^{-1/2} \frac{\partial \eta_i}{\partial \mu_i}.
\]

These are the robust counterparts of the usual weights and pseudo data appearing in the well known iterative reweighted least squares algorithm; cf. Heritier et al. [2009], Appendix E.3.
5.2.2 Robust approaches to generalized additive models

Although the flexibility of GAM versus GLM is a desirable property, outlying points in the space of the response variables and bad leverage points can have a large effect on the performance of the nonparametric estimator. For this reason two methods to obtain a resistant fit for GAM have been proposed. Azahed and Salibian-Barrera [2011] proposed to robustify the Generalized Local Scoring Algorithm of Hastie and Tibshirani [1986]. This procedure consists of iteratively fitting additive models using the backfitting algorithm until convergence. It is easily robustified by replacing the maximum likelihood based pseudo data and weights of the iterative reweighted least squares algorithm by those associated with the robust quasilikelihood equations of Cantoni and Ronchetti [2001], i.e. (5.6) and (5.7). A second approach considered by Croux et al. [2012] and Wong et al. [2014] combines the penalized splines approach to GAM fitting (see e.g. Green and Silverman [1994] or Wood [2006]) with the robust approach of Cantoni and Ronchetti [2001] for GLM. They present a method that allows for a resistant nonparametric estimation of both the mean and variance function in extended GAM, where the nonparametric functions are estimated through P-splines (Eilers and Marx [1996]). In the estimation procedure the smooth functions \( f_j \) are approximated by polynomial splines. For each \( 1 \leq j \leq p \), let \( v_j \) be a partition of the interval \([0,1]\), with \( k_n \) interior knots \( v_j = \{ 0 = v_{j,0} < v_{j,1} < \ldots < v_{j,k_n} < v_{j,k_n+1} = 1 \} \). Taking \( v_j \) as knots, the resulting polynomial splines of order \( q \geq 1 \) are \( q-1 \) continuously differentiable functions constructed from piecewise \( q \)-degree polynomials on intervals \([v_{j,l}, v_{j,l+1})\), \( l = 0, \ldots, k_n - 1 \) and \([v_{j,l}, v_{j,l+1}]\). Denote the space of such spline functions by \( \varphi_{nj} = \varphi^0([0,1], v_j) \) and \( \varphi_{nj}^0 = \{ s \in \varphi_{nj} : \int_{[0,1]} s(x) dx = 0 \} \) the space of centered spline functions. Croux et al. [2012] proposed the approximation \( f_j(\cdot) \approx f_{nj} = \sum_{l=1}^{m_n} \beta_{jl} B_{jl}(\cdot) = B_j^T \beta_{j} \) for all \( j = 1, \ldots, d \) where \( m_n = k_n + q \) and \( \{ B_{jl}(\cdot) \}_{l=1}^{m_n} \) is a set of spline bases of \( \varphi_{nj}^0 \).

The resistant estimator for GAM of both Croux et al. [2012] and Wong et al. [2014] can be viewed as the maximizer of the penalized robust quasilikelihood

\[
PQ(\beta) = Q(\beta) - \frac{1}{2} \beta^T P \beta .
\]

or equivalently as the solution of

\[
\frac{1}{n} \sum_{i=1}^{n} \Psi_i(\beta) - P \beta = 0 ,
\]

where \( \Psi_i(\beta) \) is as in (5.5) with design matrix \( B(X) = (B(x_1), \ldots, B(x_n))^T \) and \( B(x_i) = (1, B_{11}(x_{i1}), \ldots, B_{1m_a}(x_{i1}), \ldots, B_{d1}(x_{id}), \ldots, B_{dm_a}(x_{id}))^T \). The penalty matrix is \( P = \text{blockdiag}(\gamma_1 P_1, \ldots, \gamma_d P_d) \) with each \( P_i \) being a prespecified penalty matrix and \( \Gamma = (\gamma_1, \ldots, \gamma_d) \) the tuning parameters governing the smoothness of the components.

5.3 Variable selection for robust generalized additive models

The population optimization problem that one would like to solve in a perfectly specified GAM is

\[
\max \left[ \mathbb{E}\left\{ l(Y, f_0 + \sum_{j=1}^{d} f_j(X_j)) \right\} \right],
\]
where \( l \) denotes the log-likelihood of a single observation and \( f \in \mathcal{H} \). In order to account for robustness towards deviations from the stochastic assumptions and introducing regularization via a sparsity inducing penalty \( p(\cdot, \lambda) \) one can consider the alternative problem

\[
\max_f \left[ \mathbb{E} \left\{ Q_M(Y, g^{-1} \left( f_0 + \sum_{j=1}^d f_j(X_j) \right) \right\} - \sum_{j=1}^d p(\|f_j\|; \lambda) \right].
\tag{5.10}
\]

### 5.3.1 Robust group and adaptive group lasso for GAM

We consider the approximation \( f_j(\cdot) \approx \sum_{i=1}^{K_j} \beta_{ij} B_i(\cdot) = B_j^T(\cdot) \beta_{[j]} \) for all \( j = 1, \ldots, d \) as discussed in Section 5.2.2. However we look instead at the empirical version of (5.10) with the penalty function \( p(\|f_j\|; \lambda) = \lambda w_j \|f_j\| \) for \( j = 1, \ldots, d \), where \( w_j \) is some weight function. Specifically, we define a robust group lasso estimator for GAM via

\[
\hat{\beta} = \arg \max_{\beta} \left[ \mathbb{E}_n \left\{ Q_M(Y, g^{-1} \left( f_0 + \sum_{j=1}^d B_j^T(X_j) \beta_{[j]} \right) \right\} - \lambda_n \sum_{j=1}^d \sqrt{d_j} \|\beta_{[j]}\|_2 \right],
\tag{5.11}
\]

where \( \mathbb{E}_n \) denotes the expectation with respect to the empirical distribution and \( d_j \) is the dimension of the \( j \)th group. This group lasso penalty was considered for instance in Huang et al. [2010]. Alternatively variants of the penalty \( \|\hat{\beta}_{[j]}\|_{K_j} = \hat{\beta}_{[j]}^T K_j \hat{\beta}_{[j]} \) with \( K_j = \mathbb{E}_n \{ B(X_j) B^T(X_j) \} \) have been considered for instance in Ravikumar et al. [2009], Xue [2009] and Xue et al. [2010]. Note that \( \|\hat{\beta}_{[j]}\|_{K_j} \) has the nice feature of being an approximation to the empirical \( L_2 \)-norm \( \|\hat{f}_j\|_n = \mathbb{E}_n \{ \hat{f}_j(X_j) \} \). Still, we tend to prefer the problem (5.11) as it is slightly easier to understand from a robustness standpoint since the penalty function does not depend on the data generating process.

It will be seen in the next section that the group lasso estimator is consistent for estimation, but can only achieve suboptimal rates of convergence. A simple alternative estimator that can be consistent for variable selection without strengthening the assumptions required for consistency is the adaptive group lasso. Furthermore it is shown to be consistent for variable selection. Take as initial estimates the \( \hat{f}_j, j = 1, \ldots, d \) by maximizing (5.11) and define a vector of weights as \( \hat{W} = (\hat{w}_1, \ldots, \hat{w}_p)^T \) where

\[
\hat{w}_j = \begin{cases} 
1/\|\hat{\beta}_{[j]}\|, & \text{if } \|\hat{\beta}_{[j]}\| > 0 \\
\infty, & \text{if } \|\hat{\beta}_{[j]}\| = 0
\end{cases}
\tag{5.12}
\]

We now can define a robust adaptive lasso for GAM via

\[
\hat{\beta} = \arg \max_{\beta} \left[ \mathbb{E}_n \left\{ Q_M(Y, g^{-1} \left( f_0 + \sum_{j=1}^d B_j^T(X_j) \beta_{[j]} \right) \right\} - \lambda_n \sum_{i=j}^d \hat{w}_j \sqrt{d_j} \|\beta_{[j]}\|_2 \right],
\tag{5.13}
\]

where we defined \( 0 = \infty \cdot 0 \) and therefore excluded the components estimated as 0 in the first group lasso fit. The final adaptive group lasso fit are \( \hat{f}_j = B_j^T \hat{\beta}_{[j]} \) for \( j = 1, \ldots, d \).

### 5.3.2 Simultaneous scale estimation

The nuisance parameter \( \phi \) appearing in (5.1) can be ignored in the classical maximum likelihood approaches to GAM and in some cases, such as the Binomial or the Poisson distributions, it is even a known constant. In some other cases such as the normal or the gamma distribution, this nuisance parameter is unknown and is very important for
estimation when we consider the robust quasilikelihood approach. Indeed in these two cases the variance functions \( V(\mu_i) = \phi^2 \nu(\mu_i) \) are given respectively by \( \phi^2 \) and \( \phi^2 \mu_i^2 \), which implies that in practice \( \phi \) should be estimated to bound in a meaningful way Pearson residuals \( r_i = \frac{y_i - \mu_i}{\phi \sigma_i} \) in (5.5). When the nuisance parameter \( \phi \) is unknown, we propose to simultaneously estimate the smooth functions \( f_j \) and \( \phi \) by iterating until convergence the two steps
\[
\begin{align*}
\beta^{(k+1)} &= \arg \max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \phi^{(k)} Q_M (y_i, \mu_i; \phi^{(k)}) - \lambda_n \sum_{j=1}^{d} \| \beta_j \|_2, \\
\phi^{(k+1)} &= \frac{1}{n} \sum_{i=1}^{n} \chi \left( \frac{y_i - \mu_i^{(k+1)}}{\phi v_i^{1/2} (\mu_i^{(k+1)})} \right) = 0,
\end{align*}
\]
where \( \chi(u) = \psi^2(u) + c \) for some constant \( c \). This iterative scheme is to be viewed as a penalized counterpart of Huber’s Proposal 2; c.f. Huber and Ronchetti [2009]. In the additive model and with the choice \( \psi_c \) in (5.4), the robust quasilikelihood becomes the Huber loss and getting a solution via the iterative procedure described above is equivalent to solving the problem
\[
(\tilde{\beta}^T, \tilde{\phi}) = \arg \max_{\beta, \phi} \frac{1}{n} \sum_{i=1}^{n} \phi Q_M (y_i, \mu_i; \phi) + \phi - \lambda_n \sum_{j=1}^{d} \| \beta_j \|_2. \tag{5.16}
\]
Note that optimization problems analogous to (5.16) have already appeared in the lasso literature with the sum of squares loss in the work of She and Owen [2011] and Sun and Zhang [2012]. Our proposal generalizes these previous works in three different directions: it extends the scope of models covered from linear models to GLM, it extend the approach from lasso type penalties to group lasso penalties and finally the estimators are extended from a classical likelihood based approach to an outlier robust one.

5.4 Asymptotic properties

5.4.1 Conditions

Let \( \{(Y_i, X_i)\}_{i=1}^{n} \) denote independent pairs, each having the same distribution. Throughout the rest of paper, unless otherwise stated, \( X_i \) enters in \( \Psi_i(\beta) \) through \( B(X_i) \). We assume the following set of conditions:

(A1) The density function of \( X_i \) is absolutely continuous with compact support. Without loss of generality, let \( \mathcal{X} = [0, 1]^d \) be its support. Moreover there exists constants \( c \) and \( C \) such that the marginal density function \( g_j \) of \( X_j \) satisfies \( 0 < c \leq g_j(x_j) \leq C < \infty \) on \( [0, 1] \) for every \( j = 1, \ldots, d \).

(A2) For each \( j = 1, \ldots, d \), \( f_j \) is \( r \) times continuously differentiable for some \( r \geq 2 \).

(A3) The knot sequences \( v_j = \{0 = v_{j,0} < v_{j,1} < \ldots < v_{j,m_n} < v_{j,m_n+1} = 1\}, j = 1, \ldots, d \), are quasi-uniform, that is, there exists \( c > 0 \), such that
\[
\max_{j=1,\ldots,d} \frac{\max(v_{j,l+1} - v_{j,l}, l = 0, \ldots, m_n)}{\min(v_{j,l+1} - v_{j,l}, l = 0, \ldots, m_n)} \leq c.
\]
Furthermore the number of interior knots \( m_n \asymp n^{1/(2r+1)} \), where ‘\( \asymp \)' means both sides are of the same order.
Chapter 5. Robust penalized M-estimators for GAM

(A4) There is a sufficiently large open set $\mathcal{O}$ that contains the parameter $\beta^*$ such that $\forall \beta \in \mathcal{O}$ the matrices $M(\beta) = \mathbb{E}\{\partial \Psi_i(\beta)/\partial \beta\}$ and $Q(\beta) = \mathbb{E}\{\Psi_i(\beta)\Psi_i^T(\beta)\}$ satisfy

$$0 < c_1 \leq m_n \lambda_{\min}\left\{M(\beta)\right\} \leq m_n \lambda_{\max}\left\{M(\beta)\right\} \leq c_2 < \infty \quad \text{and} \quad Q(\beta) < \infty.$$

(A5) The function $\chi$ is an odd function that can be viewed as the limit of strictly monotone functions. Furthermore it is (stochastically) differentiable and has a zero at $u = 0$.

(A6) $|\psi'(r)|$ and $|\psi'(r)r|$ are bounded. Furthermore for all $\beta \in \mathcal{O}$

$$\frac{\partial Q_M(Y_i, \mu_i)}{\partial \eta_i} \leq A_1(X_i) \quad \text{and} \quad \frac{\partial^2 Q_M(Y_i, \mu_i)}{\partial \eta_i^2} \leq A_2(X_i)$$

where $A_1(X_i), A_2(X_i) < \infty$.

Conditions (A1)-(A3) are common in the polynomial spline estimation literature. See for instance Huang [2003], Xue [2009] or Kauermann et al. [2009]. Conditions (A4) and (A5) are fairly natural and have analogous standard counterparts in the robustness literature. Condition (A6) implies that $\left|\frac{\partial^3}{\partial \eta_i^3}\left(Q_M(Y_i, \mu_i)\right)\right| \leq b(X_i^T\beta)$ for some function $b$. Such $\psi$ functions can be obtained for instance by “smoothing” Huber’s function around $\{-c, c\}$ or by using an appropriately scaled logistic function of the form $a(e^r-1)/(e^r+1)$ where the constant $a$ is positive and regulates the weights given to the observations. Incidentally, a smoothing of Huber’s function has been suggested on statistical grounds in Hampel et al. [2011]. Condition (A6) is similar to the regularity condition 2 of Theorem 4 in Zou [2006].

5.4.2 Consistency and oracle properties

We will first establish the consistency of the group lasso estimator with concomitant scale estimation defined as the stationary point of the iteration of equations (5.14) and (5.15). We will then show the oracle properties of the adaptive group lasso estimator (5.16). Combining this two results we can show that our GAM estimators $\hat{f}_j$ and $\tilde{f}_j$ converge with optimal rates to their true values and $\hat{f}_j = 0$ for $j \in A = \{j : f_j = 0, j = 1, \ldots, d\}$ with probability tending one as the sample size goes to infinity. We denote the cardinality of $A$ by $k = |A|

An important preliminary result needed to show the consistency of our GAM estimator is to show the identifiability of the estimating equations corresponding to the robust quasilikelihood with concomitant scale estimation via Huber’s Proposal 2. This is given in the proposition.

Proposition 1. Under conditions (A1) and (A2), the Jacobian with respect to $(\beta^T, \phi)$ of the following system is positive definite and admits an unique solution

$$\mathbf{\Xi}(\beta, \phi) = \begin{pmatrix} \int \psi \left( \frac{\mu - \mu}{\sigma \sqrt{v/2(\mu)}} \right) \frac{1}{\sigma \sqrt{v/2(\mu)}} w(X) \partial \mu / \partial \beta dF(y, X) + a(\beta; \phi) \\ \int X \left( \frac{\mu - \mu}{\sigma \sqrt{v/2(\mu)}} \right) dF(y, X) \end{pmatrix} = 0. \quad (5.17)$$

Proposition 1 is a non trivial extension of the well established identifiability result for linear regression M-estimators that can be found for instance in Huber and Ronchetti [2009]. It extends Huber’s result to the class of M-estimators for GLM of Cantoni and
Ronchetti [2001]. Proposition 1 also complements the asymptotic normality assessments in Cantoni and Ronchetti [2006] and Aeberhard et al. [2014] in that it gives conditions under which their concomitant scale estimator is unique asymptotically. We denote by \((\beta^*, \phi^*)\) the solutions of (5.17).

We now give the main theoretical results for the group and adaptive group estimators of \((\beta^T, \phi^*)\).

**Theorem 1.** Under conditions (A1)-(A6), iterating (5.14) and (5.15) until convergence gives estimators \((\hat{\beta}^T, \hat{\psi})\) of \((\beta^T, \phi^*)\) such that if \( \lambda = O(\sqrt{n^{-1}\log(m_n d)}) \) and \( n \) is sufficiently large we have

\[
\left\| (\hat{\beta}^T, \hat{\psi}) - (\beta^* T, \phi^*) \right\|_2 = O\left( \frac{m_n^2 k \log(m_n d)}{n} \right)
\]

with probability at least \( 1 - 2e^{-\gamma \kappa_n} \) for some \( \gamma > 0 \) and \( \kappa_n = \max\{n/(m_n k)^2, \log(m_n d)\} \).

**Theorem 2.** Assume conditions (A1)-(A6) and let \( \hat{\beta} \) be as in Theorem 1 and let it define weights (5.12) for \( j = 1, \ldots, d \). Further let the dimensionality of the problem satisfy \( \log(m_n d) = O(n^\alpha) \) for \( \alpha \in (0, 1/2) \) and let the non-sparse dimensionality be of order \( k = o(n^{1/3}/m_n) \). Finally, assume \( \min_{j \in \mathcal{A}} \| f_j \|_2 \geq c f \), \( \lambda_n \sqrt{m_n k} \to 0 \) and \( \lambda_n r_n \gg \max\{\sqrt{m_n k}/n, n^{-\alpha} \sqrt{\log n}\} \), where \( r_n = \sqrt{n^{-1} m_n^2 k \log(m_n d)} \). Then, with probability tending to 1 as \( n \to \infty \), the robust adaptive lasso estimator \( \hat{\beta} = (\hat{\beta}_A^T, \hat{\beta}_{A^c}^T)^T \) satisfies the following properties:

(i) Sparsity: \( \hat{\beta}_{A^c} = 0 \).

(ii) Consistency: \( \| \hat{\beta} - \beta \|_2 = O_p(\sqrt{m_n^2 k/n}) \).

Theorems 1 and 2 are interesting results on their own in that they extend the results of Fan and Lv [2011] to robust GLM with group penalties and with concomitant scale parameter estimation. Note that in the convergence rates given in our theorems there is an additional \( \sqrt{m_n} \) factor appearing in the rates with respect to the usual optimal rates in sparse regression. It is a consequence of the design matrix generated by the spline basis. The above two theorems serve as building blocks for establishing the asymptotic properties of our robust and sparse GAM proposals stated next.

**Theorem 3.** Under the assumptions of Theorem 2, then:

(i) \( \sum_{j=1}^d \| \hat{f}_j - f_j \|_2 = O_p\left( kn^{-r/(2r+1)} \log(m_n d) \right) \).

(ii) \( \sum_{j=1}^d \| \hat{f}_j - f_j \|_2 = O_p\left( kn^{-r/(2r+1)} \right) \).

(iii) \( \mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \to_p 1 \), where \( \hat{\mathcal{A}} = \{ j : \hat{f}_j \neq 0, j = 1, \ldots, d \} \).

Note that the rates of convergence of the adaptive lasso estimator obtained in Theorem 3 are optimal for \( r + 1 \) times continuously differentiable functions as shown in Stone [1982]. Theorem 3 is similar in spirit to the results obtained in Huang et al. [2010] for the additive model. It improves on previous work by extending the result to the more general framework of GAM with concomitant scale estimation and including robustness considerations. Furthermore, unlike in the theory developed in Huang et al. [2010], we can let the cardinality of the set \( \mathcal{A} \) diverge with the sample size.
5.5 Robustness properties

In the statistical literature different mathematical characterizations of the idea of robustness have been proposed. The most prominent examples are the seminal minimax approach of Huber [1964], the sensitivity curve Tukey [1977], the influence function Hampel et al. [1986], the maxbias curve (Huber [1964], Hampel et al. [1986], Maronna et al. [2006]) and the finite sample breakdown point (Donoho and Huber [1983]). In this section we will focus on the approach based on the influence function. Obtaining the influence function of sparse parametric penalized M-estimators is a nontrivial task because the nondifferentiability inherent to sparsity inducing penalties impedes using well established theory. We will therefore extend to GAM the notion of limiting influence function introduced in Chapter 4 for parametric penalized M-estimators.

5.5.1 Bounded influence approach

From a robust statistics viewpoint, constructing and studying smooth bounded functionals \( T \) is of great interest because they yield stable minimizers within small neighborhoods of \( F \). A bounded derivative \( \nabla T(F) \) of \( T(F) \) implies that the functional \( T(F) \) cannot increase or decrease arbitrarily in small neighborhoods of \( F \). This is in fact the key ingredient in Hampel’s general approach to robustness based on the influence function (Hampel et al. [1986]). Given a distribution space \( F \), a parameter space \( \Theta \) and a functional \( T : F \rightarrow \Theta \), the influence function of \( T \) at a point \( z \in Z \) for a distribution \( F \) is defined as

\[
IF(z; F, T) = \lim_{\epsilon \to 0^+} \frac{T(F_\epsilon) - T(F)}{\epsilon},
\]

where \( F_\epsilon = (1 - \epsilon)F + \epsilon \Delta z \). In our set up we are interested in finding the influence function of functionals \( T(F) \) defined as the solution to the regularized population problem (5.10). In order to do so, we will follow the construction proposed in Chapter 4 and consider a sequence of smooth approximating problems to (5.10) defining a sequence of approximating functionals \( T_m(F) \) for which the influence function is easily obtained. Specifically, the idea is to consider a sequence of penalties \( p_m \) such that \( \lim_{m \to \infty} p_m(f_j; \lambda) = p(f_j; \lambda) \) and denote by \( T_m(F) = T(F; p_m) \) the approximating functionals. We let \( IF_{p_m}(z; T, F) \) be the influence function of \( T(F; p_m) \) and define the influence function of \( T(F) \) as

\[
IF(z; F, T) := \lim_{m \to \infty} IF_{p_m}(z; F, T).
\] (5.18)

In the next subsection we establish some theory for this limiting influence function. We then use it to derive the form of the influence function for group lasso GAM estimators and show that our proposal is bounded-influence robust.

5.5.2 Regularity conditions

For simplicity we will use the notation \( Q(Z, f) = Q_M[Y, g^{-1}\{f_0 + \sum_{j=1}^d f_j(X_j)\}] \), where \( Z = (X, Y) \) and \( f \in \varphi_n = \varphi_0 \oplus \varphi_n^0 \oplus \cdots \oplus \varphi_n^{0d} \), where \( \varphi_0 \) denotes the space of constant functions integrable in \([0, 1]\) and \( \varphi_n^0 \) denote the space of centered spline functions with \( m_n \) parameters for \( j = 1, \ldots, d \). We also use norm \( \|f_j\|_\beta = \|\sum_{l=1}^{m_n} \beta_l B_{jl}\|_\beta = \|\beta_j\|_2 \) for \( f_j \in \varphi_{nj} \) and \( j = 1, \ldots, d \). We require the following set of assumptions to obtain our theoretical results regarding the limiting influence.
\[ \mathbb{E}[Q(Z, f)] \] is continuous and convex in \( \varphi_n \). Furthermore it has two derivatives with respect to \( f \) denoted by \( \mathbb{E}[\Psi(Z, f)] \) and \( \mathbb{E}[\bar{\Psi}(Z, f)] \) respectively.

**B2** There is a unique function \( f \in \varphi_n \) such that \( \mathbb{E}[Q(z, f)] < \inf_{h \in \varphi_n \setminus \mathcal{O}} \mathbb{E}[Q(z, h)] \) for every open set \( \mathcal{O} \) containing \( f \).

**B3** There is a unique \( f^* = \arg \inf \{ \mathbb{E}[Q(Z, f)] + \lambda \sum_{j=1}^{d} p(\|f_j\|) \} \) in an open ball \( \mathcal{O} \) containing \( f \). The functions \( \Psi(z, f) \) and \( \mathbb{E}[\bar{\Psi}(Z, f)] \) are continuous at \( f^* \) for all \( z \in \mathbb{R}^d \times \mathbb{R} \).

The above conditions are analogous to (A1)-(A3) in Chapter 4. Assumptions (B1) and (B2) are regularity conditions on the population risk. Namely, they require some smoothness and a well separated maximum and in that sense are similar to (A4)-(A6). Assumption (B3) imposes regularity conditions on the score function \( \Psi \) and a unique regularized risk minimum on a neighborhood of the global risk minimum. Obviously when both the population risk and the penalty function \( p \) are convex we can take \( \mathcal{O} = \varphi_n \). Nonconvex loss functions and penalty functions restrain the size of the ball \( \mathcal{O} \).

### 5.5.3 Limiting influence function

For the sake of simplicity and space we will only focus on the group lasso penalized functional of the form (5.10) without the concomitant scale formulation. The result can be extended to two stage type of estimators using the same approximation scheme in the two population penalized M-functionals defined by them. A formal derivation for the parametric case was provided in Chapter 4.

We derive the influence function of regularized M-functionals of the form (5.10) for convex penalties \( p(\cdot ; \lambda) \). In order to do so we introduce a sequence of approximating regularized M-functionals of the form

\[
\max_{f_0 \in \mathbb{R}, f_j \in \mathcal{H}_i, 1 \leq j \leq d} \left[ \mathbb{E}\left\{ Q_M(Y, g^{-1}\left(f_0 + \sum_{j=1}^{d} f_j(X_j)\right)) \right\} - \lambda \sum_{j=1}^{d} p_m(\|f_j\|) \right],
\]

where \( p_m \) is a smooth convex function such that \( p_m \to p \) when \( m \to \infty \). From (5.19) we obtain a sequence of approximating functionals \( T_m(F) \) for which we can easily compute the influence function as well as its limiting form. The next theorem establishes the form of the limiting influence function for GAM and states some general intermediate results regarding the uniqueness of its definition. We use the notation \( C_{\leq 0}(\mathbb{R}) \) for the set of continuous and infinitely differentiable functions \( p : \mathbb{R} \to \mathbb{R}_{\geq 0} \) and \( W^{2, 2}(\mathbb{R}) \) for the Sobolev space given by the subset of functions in \( L^2(\mathbb{R}) \) whose second weak derivative is integrable in \( L^2 \) norm.

**Theorem 4.** Let \( T(F) = (f_1^*, \ldots, f_d^*) \) be obtained as the solution to the problem (5.10). Let \( \{p_m\}_{m \geq 1} \) be a sequence in \( C^\infty(\mathbb{R}) \) converging to \( |t| \) in the Sobolev space \( W^{2, 2}(\mathbb{R}) \) when \( m \to \infty \). Further suppose that the problem (5.19) satisfies (A1)-(A3) for \( m \geq 1 \). Then, we have that for any choice of \( p_m \)

(i) There is a unique limit \( \lim_m T(F; p_m) = T(F) \).

(ii) The influence function (5.18) of \( T(F) \) is unique and has the form

\[
\text{IF}(z; T, F) = -B(x)^T S^{-1}\{\Psi(z, f^*) + \lambda \gamma(f^*)\} B(x),
\]
where \( \gamma(f^\star) \) is a fixed deterministic \( m_n d \) dimensional vector depending only on \( f^\star \), 
\[ S^{-1} = \text{blockdiag}\{(M_{A^\star})^{-1}, 0\} \] 
and \( M_{A^\star} = E_F[B_{A^\star}(X)\Psi_{A^\star}(Z, f^\star)B_{A^\star}(X)^T] \).

Theorem 4 shows that our GAM estimator is locally robust in the sense that its influence function is bounded. Indeed it is immediate from the form of the influence function that it will be bounded as long as \( \Psi \) is bounded. This is inline with classical results for parametric M-estimators (Hampel et al. [1986]) and the semiparametric generalizations of the influence function discussed next. It would be interesting to investigate whether one can define the influence function of the functional \( T \) defined by (5.10) mapping from the space \( \mathcal{P} \) to \( \mathcal{H} \) instead of \( \varphi_n \). This seems to be a far more complicated enterprise that would require a careful inspection in another work.

### 5.5.4 Other approaches to the influence function

The question of deriving the influence function for non sparse semiparametric estimators is not a trivial one. There are two different approaches to this problem in the literature. None of them can be straightforwardly adapted to our setting mainly because of the non-differentiability of our penalty function. The first approach was considered by Christmann and Steinwart [2004] and Christmann and Steinwart [2007] in the context of regularized regression in reproducing kernel Hilbert spaces (RKHS). In this approach the object of interest is the functional \( T \) defined by the map \( T : \mathcal{P} \to \mathcal{H} \), where \( \mathcal{P} \) denotes the space of probability distributions in \( \mathbb{R}^d \times \mathbb{R} \) and \( \mathcal{H} \) denotes a RKHS. Therefore the derivation of the influence function in this work relies on the implicit function theorem in Banach spaces. A second approach was studied in a partially linear model setting by Boente and Rodriguez [2010], where the nonparametric component of the model is estimated with kernel smoothers. In this work, the proposed influence function is related to the idea of empirical influence function of Mallows [1975] and Tukey [1977]. A somehow similar approach has also been developed by Newey [1994] and Ichimura and Newey [2015] for estimators of parameters that depend on unknown functions such as density functions or conditional expectations. Note that the RKHS approach can be easily adapted to GAM. It therefore seems to be more related to our work.

### 5.6 Implementation

#### 5.6.1 A pathwise algorithm

Estimating the robust lasso for generalized linear models when \( d > n \) requires an appropriate algorithm. We propose a pathwise algorithm to compute the group lasso solution path of our robust and sparse GAM based on iterative block thresholding. The algorithm is based on successive expected quadratic approximations of the robust quasilikelihood about the current estimates. Specifically for a given value of the tuning parameter we successively solve via the iteratively block thresholding algorithm of She [2009] the weighted least squares problem given by

\[
\|\sqrt{W}(z - B(X)\beta)\|_2^2 + \lambda \sum_{j=1}^{d} \sqrt{d_j} \|\beta_{[j]}\|_2,
\]

where \( W = \text{diag}(W_1, \ldots, W_n) \) is the weight matrix and \( z = (z_1, \ldots, z_n)^T \) the vector of pseudo data given respectively by (5.6) and (5.7). Our pathwise iterative block thresholding algorithm is therefore a sequence of nested loops:
5.6. Implementation

1. Outer loop: decrease $\lambda$.

2. Middle loop: Update $W$ and $z$ in (5.20) using the current parameters $\hat{\beta}_\lambda$.

3. Inner loop: run the iterative block thresholding algorithm on the reparametrized weighted least squares problem (5.20).

There are two main differences between our algorithm and the ones in the spirit of Friedman et al. [2010]. First, in the quadratic approximation step the expectation of the usual approximation is computed. This small modification turns out to be crucial for the robust group lasso computation in the inner loop because it guarantees that $W$ has only positive components in Poisson and Binomial regression when using Huberized residuals. In other words, this guarantees the convergence of the inner loop. Second, we solve the inner loop via the iterative block thresholding algorithm described in She [2009] and outlined in the next subsection. Note that it is very similar to the composite gradient algorithm of Nesterov [2007] and the iterative thresholding algorithm of Daubechies et al. [2004] and Beck and Teboulle [2009]. In fact all these algorithms belong to the class of proximal algorithms discussed in the review paper Polson et al. [2015].

In the grid of tuning parameters used by our algorithm, we choose as the largest starting point $\lambda_0 = n^{-1} \max_{1 \leq j \leq d} \| \sqrt{W} B(X)_{[j]} z \|_2 / \sqrt{d_j}$ where $B(X)_{[j]}$ is the submatrix of the design matrix $B(X)$ corresponding to $j$th group and $W$ and $z$ are computed with $\beta = 0$. It is such that 0 is the maximizer of the robust lasso. We then follow the three steps mentioned above using warm starts and active set cycling to speed up computations as discussed in Friedman et al. [2010].

5.6.2 An iterative block thresholding algorithm for the group lasso

Let $X$ be a design matrix of dimension $n \times p$ and $y \in \mathbb{R}^n$ the response vector. Provided $\|X^T X\|_2 < 1$, She [2009] shows that penalized regression problems of the form

$$\min_{\beta} \| y - X \beta \|_2^2 + P(\beta ; \lambda)$$

can be solved by iterating until convergence

$$\beta^{(m+1)} = \Theta \left( (I - X^T X) \beta^{(m)} + X^T y; \lambda \right),$$

where $P(\cdot ; \lambda)$ is a penalty function with regularization parameter $\lambda$ and $\Theta(\cdot ; \lambda)$ its corresponding thresholding rule. In the particular case of the group lasso penalty we are interested in solving the problem

$$\min_{\beta} \| y - X \beta \|_2^2 + \lambda \sum_{j=1}^G \sqrt{d_j} \| \beta_{[j]} \|_2,$$

where $d_j$ is the size of the group of parameters $\beta_{[j]}$ and $G$ the number of groups. Using the algorithm of She [2009], this can be done by iterating until convergence

$$\beta^{(m+1)} = \Theta_{BT} \left( (\frac{1}{k_0} X^T X) \beta^{(m)} + \frac{1}{k_0} X^T y; \lambda \frac{1}{k_0} \right),$$
Chapter 5. Robust penalized M-estimators for GAM

where $k_0 > \|X^TX\|_2$ and $\Theta_{BT}(\cdot; \lambda)$ denotes the group block thresholding operator given by

$$\Theta_{BT}(\Gamma; \lambda) = \left\{ \left( 1 - \lambda \sqrt{\frac{d_j}{\|\Gamma[j]\|_2}} \right) + \Gamma[j] \right\}_j^G.$$ 

We use this algorithm with $k_0 = 10\|X^TX\|_2$ in the implementation of the inner step of the algorithm described in the previous subsection.

### 5.6.3 Tuning parameter selection

We choose the tuning parameter $\lambda_n$ based on a robust Schwarz information criterion. Specifically we select the parameter $\lambda_n$ that minimizes

$$\text{BIC}(\lambda_n) = \rho_n(\hat{\beta}_{\lambda_n}) + \frac{\log n}{n} |\text{supp} \hat{\beta}_{\lambda_n}|,$$

where $|\text{supp} \hat{\beta}_{\lambda_n}|$ denotes the cardinality of the support of $\hat{\beta}_{\lambda_n}$. Note that we write $\hat{\beta}_{\lambda_n}$ to stress the dependence of the minimizer of the penalized M-estimator on the tuning parameter. In an unpenalized set up, an information criterion of this form was considered by Machado [1993] who provided solid theoretical justification for it by proving model selection consistency and robustness. In the penalized regression literature Lambert-Lacroix and Zwald [2011] and Li et al. [2011] have also used a robust BIC to select the tuning parameter. The soundness of this choice relies on the work, among others, of Wang et al. [2007b], Wang et al. [2009] and Fan and Tang [2013]. Other popular choices for the selection of the tuning parameter include generalized information criteria (see for instance in Zhang et al. [2010] and Flynn et al. [2013]) and $K$-fold cross-validation.

### 5.7 Simulations

In this section we explore the behavior of our proposals in a simulated example. For the robust estimators based on the robust quasilikelihood we used $\psi(r) = \psi_c(r)$, i.e. the Huber function defined by

$$\psi_c(r) = \begin{cases} r & \text{if } |r| \leq c \\ c \cdot \text{sign}(r) & \text{otherwise.} \end{cases}$$

We took $c = 1.5$ throughout the simulations. We measure the ability of our classical and robust proposals to select the true model with and without contamination. For each simulation scenario described below we generated 100 samples and computed the prediction accuracy of the method by computing:

$$\text{PE} = \mathbb{E}_X \{ (\hat{f}(X) - f(X))^2 \}.$$

The above expectation is approximated by a sample of 2'000 points from the distribution of $X$. We consider a Poisson regression model with canonical link $g(\mu_i) = \log(\mu_i)$ and with an underlying generating model of the form

$$\log \mu_i = f_1(x_{i1}) + f_4(x_{i4}) + f_6(x_{i6}),$$

where the smooth functions are respectively $f_1(x) = 2\sin(\pi x)$, $f_4(x) = \frac{1}{16.174} \{ x^{11} [10(1-x)]^6 + 10(10x)^3(1-x)^{10} \}$ and $f_6(x) = \frac{1}{1.125} \{ \cos(2\pi x) + \sin(\pi x) \}$. They are plotted in
Figure 5.1. The covariates $X_{ij}$ were generated from standard uniforms with correlation $\text{cor}(x_{ij}, x_{ik}) = \rho^{|j-k|}$ and $\rho = 0.5$ for $j, k = 1, \ldots, p$. The additive component $\eta_i$ is then rescaled to yield values in the interval $[0.2, 3]$. Note that this rescaling operation was considered by Marra and Wood [2011] in order to control the signal to noise ratio through the squared correlation coefficient between $\mu_i$ and $y_i$. Our rescaling corresponds to the low signal to noise ratio scenario in the aforementioned work. The response variables $Y_i$ were then generated according to the Poisson distribution $P(\mu_i)$ and a perturbed distribution of the form $(1-b)P(\mu_i) + bP(\nu\mu_i)$, where $b \sim \text{Bin}(1, \epsilon)$. The latter represents a situation where the distribution of the data lies in a small neighborhood of the model that can produce for instance overdispersion. We set $\nu = 10$ and $\epsilon = 0, 0.05$. The number of observations $n$ was set to 200 and the number of covariates $p = 400$. We used cubic B-splines with 5 evenly distributed knots for all the functions $f_j$. We replicated the simulation 100 times.

<table>
<thead>
<tr>
<th>Method</th>
<th>PE</th>
<th>size</th>
<th>#FN</th>
<th>PE</th>
<th>size</th>
<th>#FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon = 0$, $\nu = 0$</td>
<td>0.196</td>
<td>3</td>
<td>0</td>
<td>0.782</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(0.051)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0.190)</td>
<td>(8.895)</td>
<td>(1.482)</td>
</tr>
<tr>
<td>$\epsilon = 0.05$, $\nu = 105$</td>
<td>0.163</td>
<td>3</td>
<td>0</td>
<td>0.216</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.042)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0.051)</td>
<td>(1.482)</td>
<td>(0)</td>
</tr>
<tr>
<td>group lasso</td>
<td>0.057</td>
<td>3</td>
<td>0</td>
<td>0.701</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0.171)</td>
<td>(7.413)</td>
<td>(1.482)</td>
</tr>
<tr>
<td>group r-lasso</td>
<td>0.060</td>
<td>3</td>
<td>0</td>
<td>0.083</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0.037)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
</tbody>
</table>

Table 5.1: Summary of the performance of the penalized estimators computed under two different high dimensional Poisson regression scenarios. The median of each measure over 100 simulations is given with its MAD in parenthesis.

We show the performance of the classical and robust group lasso estimators as well as their adaptive counterparts and the respective oracles. Table 5.1 summarizes the overall performance of the different penalized procedures in terms prediction error of the estimated coefficients, the size of the selected model and the number of false negatives. Figure
Figure 5.2: The plots show the prediction error performance of the group lasso (L), the robust group lasso (RL), the adaptive group lasso (AL), the robust adaptive group lasso (RAL), the classical oracle (O) and the robust oracle (RO). The plot on the left correspond to $\epsilon = 0$ while the one on the right has $\epsilon = 0.05$.

Figure 5.3: The plots show the cardinality of the support of the group lasso (L), the robust group lasso (RL), the adaptive group lasso (AL) and the robust adaptive group lasso (RAL). The plot on the left correspond to $\epsilon = 0$ while the one on the right has $\epsilon = 0.05$.

5.2 gives a more detailed view of the prediction errors of the different estimators while Figure 5.3 illustrates the model selection properties. It is clear that without contamination the classical procedures and their robust counterparts have a very similar performance. Interestingly in this example the group lasso estimators already recover successfully the true model. They are however outperformed by their adaptive counterparts in terms of prediction error. Notice that the adaptive estimators perform even slightly better than
the oracle estimators in term of prediction. This is not entirely surprising because the oracles are unregularized estimators and from the nonparametric statistics literature it is clear that some sort of regularization is desirable. Typically this is done via a ridge type of penalty on the spline coefficients that captures a notion of roughness in the estimated functions. Even though the group lasso penalty might not be the best one for penalizing the roughness of a function, it seems to bring a slight improvement over an entirely unpenalized fit. Under contamination for the classical estimators yield very poor prediction errors compared to their robust counterparts. Furthermore they tend to keep large models while estimating as zero some of the relevant smooth functions. In this example even the robust group lasso outperforms the classical oracle estimator under contamination.

5.8 Appendix

5.8.1 Proof of main results

Proof of Proposition 1

We adapt the argument given in Huber and Ronchetti [2009] p.136 for the location-scale case. If we show that the Jacobian $J$ of the map

$$\xi : (t, \phi) \mapsto \left( \int \psi(y - t) \frac{1}{\phi v^{1/2}(t)} dF(y), \int \chi(y - t) \frac{1}{\phi v^{1/2}(t)} dF(y) \right)$$

(5.21)

is positive definite, then the existence of a unique solution of $\xi(t, \phi) = 0$ follows from the intermediate value theorem provided $\chi(0) < 0 < \chi(\pm \infty)$ and $F$ does not have a pointmass $\varpi \geq \chi(\pm \infty)/\{\chi(\pm) - \chi(0)\}$. Defining the matrix

$$A = \begin{pmatrix} X & 0 \\ 0 & X^T \end{pmatrix}$$

we can see that $\Xi(\beta, \phi)$ is essentially $E(AJ(t, \phi)A^T)$ with $t = g^{-1}(X^T \beta)$ and is therefore also positive definite.

It remains to show that $J$ is positive definite in order to complete the proof. Note that

$$J = -\left( \int \psi'(r) \left\{ \frac{1}{\phi v(t)} + \frac{1}{2} \phi' v(t) \right\} r dF(r) \int \chi'(r) \left\{ \frac{1}{\phi v(t)} + \frac{1}{2} \phi' v(t) \right\} r dF(r) \right)$$

(5.22)

with $r = \phi^{-1}(y - t)/v^{1/2}(t)$. Defining a new probability measure $F^*$ by

$$F^*(dr) = \frac{\psi'(r)}{E_F \{ \psi'(r) \}} F(dr)$$

we can rewrite (5.22) as

$$-E_F \{ \psi'(r) \} \left( \frac{1}{\phi v(t)} + \frac{1}{2} \phi' v(t) \int E_{F^*}(r) + E_{F^*} \{ \psi(r)/\psi'(r) \} \right) \left( \frac{1}{\phi v(t)^2} \int E_{F^*}(r) \right)$$

$$= \frac{1}{\psi'(r)} \int E_{F^*} \left( \chi'(r)/\psi'(r) \right) + \frac{1}{2} \phi' \int E_{F^*} \left( r \chi'(r)/\psi'(r) \right) + \frac{1}{\phi} \int E_{F^*} \left( r \chi'(r)/\psi'(r) \right).$$
Its determinant can therefore be written as
\[
|J| = \left| \mathbb{E}_F \{ \psi'(r) \} \right|^2 \left\{ \frac{1}{\phi' v} \left[ \mathbb{E}_F \{ r \chi'(r)/\psi'(r) \} - \mathbb{E}_F \{ \chi'(r)/\psi'(r) \} \right] \\
+ \frac{\nu'}{2\phi' v^{3/2}} \left[ \mathbb{E}_F \{ r \chi'(r)/\psi'(r) \} \mathbb{E}_F \{ \psi(r)/\psi'(r) \} \right] \\
- \mathbb{E}_F \{ r \chi'(r)/\psi'(r) \} \mathbb{E}_F \{ \psi(r)/\psi'(r) \} \right\}
\]

Assuming that \( \psi' > 0 \), it follows that \( \chi'/\psi' > 0 \) and therefore that \( \text{cov}_F \{ r, \chi'(r)/\psi'(r) \} > 0 \) because for any two strictly monotone functions \( f \) and \( g \) and a random variable \( Z \) we have \( \text{cov} \{ f(Z), g(Z) \} > 0 \). Furthermore both \( \mathbb{E}_F \{ r \chi'(r)/\psi'(r) \} = \mathbb{E}_F \{ \psi(r) \}/\mathbb{E}_F \{ \psi'(r) \} \)
and \( \mathbb{E}_F \{ r \chi'(r)/\psi'(r) \} = 2\mathbb{E}_F \{ 2\psi(r) \}/\mathbb{E}_F \{ \psi'(r) \} \) will be larger or equal to zero if \( r \) is symmetric or right skewed. Since \( \psi \) can be approximated by a sequence of function \( \{ \psi_n \}_{n \geq 1} \) such that \( \psi_n' > 0 \), we have that \( |J| > 0 \). \( \square \)

**Proof of Theorem 1**

We will obtain the desired rates of convergence by showing that the conditions of Theorem 1 in Loh and Wainwright [2015] hold with the claimed probability. Note that the set \( \Omega \), the function \( g : \mathbb{R}^p \to \mathbb{R}_+ \) and the positive constant \( c \) used in their theorem can be respectively replaced in our set up by \( \mathbb{R}^{m \times d} \times \mathbb{R}_+ \), \( \| \cdot \|_1 \) and a positive term proportional to \( 1/\lambda \). In order to apply their result we show that with probability at least \( 1 - 2e^{-\gamma_n/(m_\epsilon k)^2} \), the robust quasilikelihood satisfies the restricted strong convexity discussed in their article. This would complete the proof in view of Lemma 4.

Let \( \theta = (\beta^T, \phi)^T \) and \( \Upsilon_i(\theta) = [\Psi_i((\beta^T, \phi)^T), \chi_i(\beta, \phi)]^T \). Our estimator \( \tilde{\theta} \) is defined as a vector \( \theta' \) that satisfies the equation
\[
\langle \Upsilon_i(\theta') + \nabla \mathcal{R}(\theta'), \theta - \theta' \rangle \text{ for } \theta \in \mathbb{R}^{m \times d} \times \mathbb{R}_+,
\]
where \( \mathcal{R}_\lambda((\beta^T, \phi)^T) = \lambda \sum_{j=1}^d \| \beta_j \|_2 \) and \( \nabla \mathcal{R}_\lambda \) denotes its subgradient. The restricted strong convexity condition requires
\[
\mathcal{E}_n(\Delta) \geq \begin{cases} 
\alpha_1 \| \Delta \|_2^2 - \tau_1 \log m_\epsilon \| \Delta \|_1^2 & \text{for all } \| \Delta \|_2^2 \leq 1 \\
\alpha_2 \| \Delta \|_2^2 - \tau_2 \log m_\epsilon \| \Delta \|_1^2 & \text{otherwise.}
\end{cases} \tag{5.23a}
\]

where the \( \alpha_j \)'s are strictly positive constants, the \( \tau_j \)'s are non-negative constants and
\[
\mathcal{E}_n(\Delta) = \frac{1}{n} \left\{ \sum_{i=1}^n \Upsilon_i(\theta + \Delta) - \sum_{i=1}^n \Upsilon_i(\theta) \right\}^T \Delta.
\]

Note that
\[
\mathcal{E}_n(\Delta) = \Delta^T \Xi_n(\theta) \Delta \geq \Delta^T \Xi(\theta') \Delta - \left| \Delta^T \left\{ \Xi(\theta') - \Xi_n(\theta') \right\} \Delta \right|, \tag{5.24}
\]
where $\Xi(\theta) = E\{Y_i(\theta)\}$, $\Xi_n(\theta) = \frac{1}{n} \sum \tilde{Y}_i(\theta)$ and $\theta'$ lies between $\theta$ and $\theta + \Delta$. Let $k' = m_n k$ and denote by $\Xi^{k'}$ and $\Xi_n^{k'}$ the $2k' \times 2k'$ submatrices of $\Xi$ and $\Xi_n$ formed from $2k'$ components. Then by Proposition 1, Lemma 2 and the arguments given in Lemma 3, we have that

$$u^T \Xi_n^{k'}(\theta) u \geq c_1 + u^T \{ \Xi_n^{k'}(\theta') - \Xi^{k'}(\theta') \} u$$

and

$$P\{u^T (\Xi_n^{k'} - \Xi^{k'}) u > \delta\} \geq 1 - 2e^{-c\delta^2 n/k'^2},$$

where $\delta$ and $c$ are some positive constants. In particular we can choose $\delta = c_1/54$ and apply Lemma 5 with $s = \frac{n}{\log(m_n d)}$ to the second term in (5.24). From this we see that

$$E_n(\Delta) \geq c_1 \|\Delta\|^2_2 - \left( \frac{c_1}{2} \|\Delta\|^2_2 + \frac{c_1 \log(m_n d)}{n} \|\Delta\|^2_1 \right)$$

holds for all $\Delta \in \mathbb{R}^d$. This shows that (5.23a) holds with probability at least $1 - 2e^{-c_1^2 n/(m_n k)^2}$ taking $\gamma_1 = c_1^2/54^2$. Finally (5.23b) holds by Lemma 6 and taking a sufficiently large $n$. □

Proof of Theorem 2

The proof follows from arguments very similar to those of Theorem 3 in Chapter 3 and is therefore omitted for the sake of space. □

Proof of Theorem 3

Let $f_n^*$ be the best population approximation in the space $\varphi_n^0 = \oplus_{j=1}^d \varphi_{nj}^0$ of $f \in \mathcal{H}$ in the sense that

$$(\alpha^*, f_n^*) = \arg \min_{\alpha \in \mathbb{R}, f_n \in \varphi_n^0} E \left( Q_M[Y, \varphi^{-1}_{nj}\{\alpha + \sum_{j=1}^d f_j(X_j)\}] \right).$$

Throughout this proof we will use the simplified notation $Q[Y, f(X)] = Q_M[Y, \varphi^{-1}_{nj}\{\alpha + \sum_{j=1}^d f_j(X_j)\}]$. In order to establish (i) it suffices to bound the inequality

$$\sum_{j=1}^d \|\tilde{f}_j - f_j\|_2 \leq \sum_{j=1}^d \|\tilde{f}_j - f_{nj}^*\|_2 + \sum_{j=1}^d \|f_{nj}^* - f_j\|_2$$

using

$$\sum_{j=1}^d \|\tilde{f}_j - f_{nj}^*\|_2 = O_p \left( kn^{-r/(2r+1)} \log(m_n d) \right) \quad (5.25)$$

and

$$\sum_{j=1}^d \|f_{nj}^* - f_j\|_2 = O_p \left( kn^{-r/(2r+1)} \right). \quad (5.26)$$

It is easy to see that Theorem 1 and Lemma 1 imply (5.25). Showing (5.26) requires different considerations. Define $f_{n(-j)}^* = f_n^* - f_{nj}^* + \tilde{f}_{nj}$ for some $\tilde{f}_{nj} \in \varphi_{nj}$ and note that by definition of $f_n^*$ we have

$$E \left( Q_M[Y, f_n^*(X)] \right) \leq E \left( Q_M[Y, f_{n(-j)}^*(X)] \right).$$
Then by Lemma 7 it follows that
\[
\mathbb{E}\left(\left\{ f^*_n(X_j) - f_j(X_j)\right\} \frac{\partial}{\partial \eta} Q_M[Y, f(X)] \right|_{f=f^*} + \left\{ f^*_n(X_j) - f_j(X_j)\right\}^2 \frac{\partial^2}{\partial \eta^2} Q_M[Y, f(X)] \right|_{f=f} \leq \mathbb{E}\left(\left\{ \bar{f}_{n_j}(X_j) - f_j(X_j)\right\} \frac{\partial}{\partial \eta} Q_M[Y, f(X)] \right|_{f=\bar{f}} + \left\{ \bar{f}_{n_j}(X_j) - f_j(X_j)\right\}^2 \frac{\partial^2}{\partial \eta^2} Q_M[Y, f(X)] \right|_{f=\bar{f}},
\]
where \(\bar{f}\) and \(\bar{f}\) are a function taking values between \(f^*_n\) and \(f^*_n + f_{n_j}\). Since \(\mathbb{E}[f_j(X_j)] = \mathbb{E}[\bar{f}_{n_j}(X_j)] = \mathbb{E}[f^*_n(X_j)] = 0\), from (A6) and the above inequality we deduce that
\[
\mathbb{E}\left(\left\{ f^*_n(X_j) - f_j(X_j)\right\}^2 \frac{\partial^2}{\partial \eta^2} Q_M[Y, f(X)] \right|_{f=\bar{f}} \leq \mathbb{E}\left(\left\{ \bar{f}_{n_j}(X_j) - f_j(X_j)\right\}^2 \frac{\partial^2}{\partial \eta^2} Q_M[Y, f(X)] \right|_{f=\bar{f}}.
\]
From this last inequality and Condition (A4) we can deduce that there is a constant \(\tilde{C}\) such that
\[
\mathbb{E}\left[\left\{ f^*_n(X_j) - f_j(X_j)\right\}^2 \right] \leq \tilde{C}\mathbb{E}\left[\left\{ \bar{f}_{n_j}(X_j) - f_j(X_j)\right\}^2 \right].
\]
Then, by (A1) we have
\[
c\|f^*_n - f_j\|^2 \leq \tilde{C}\|\bar{f}_{n_j} - f_j\|^2.
\]
Since \(f^*_n = 0\) for \(j \in \mathcal{A}\) we deduce that
\[
\sum_{j=1}^{d} \|f^*_n - f_j\|^2 \leq c\sum_{j=1}^{k} \|\bar{f}_{n_j} - f_j\|^2 \leq c\max_j \|\bar{f}_{n_j} - f_j\|^2
\]
for some constant \(c\). Now it suffices to choose functions \(\bar{f}_{n_j}\) in (5.27) such that \(\|\bar{f} - f_j\| \leq cm_n^{-r}\). We can do so because from de Boor [2001] p.149, it is easy to see that for each \(j = 1, \ldots, d\), there exists a constant \(c > 0\) and a spline function \(f_{n_j} \in \varphi_{n_j}\) such that \(\|f - f_j\|_\infty \leq cm_n^{-r}\). This shows (5.26) and completes part (i).

Given Theorem 2, part (ii) follows from an argument very similar to (i) and is therefore omitted. Finally, (iii) is a consequence of Theorem 2. \(\square\)

**Proof of Theorem 4**

We begin by showing (i). By convexity we can rewrite (5.10) as
\[
\max_{f \in \varphi_n} \mathbb{E}\left\{ Q_M(Y, g^{-1}\left(f_0 + \sum_{j=1}^{d} f_j(X_j)\right)) \right\} \text{ s.t. } \sum_{j=1}^{d} \|f_j\| \leq K.
\]
Since the mapping \(\Gamma : W^{2,2} \mapsto \varphi_n\) given by \(\Gamma p = \{f \in \varphi_n | \sum_{j=1}^{d} p(\|f_j\|) \leq K\}\) is weakly continuous by Lemma 8, we can invoke Berge’s maximum theorem and say that
\[
\Phi p = \{f | f \in \Gamma p, \mathbb{E}\left\{ Q_M(Y, g^{-1}\left(f_0 + \sum_{j=1}^{d} f_j(X_j)\right)) \right\} = M(p)\}
\]
is continuous in \(W^{2,2}\) and the mapping
\[
\mathbb{E}\left\{ Q_M(Y, g^{-1}\left(f_0 + \sum_{j=1}^{d} f_j(X_j)\right)) \right\} = M(p)
\]
is upper hemicontinuous from $W^{2,2}$ to $\mathcal{O}$. Since

$$f^m = \phi p_m = \{f | f \in \Phi p_m, -\mathbb{E}(Q_M[Y, g^{-1}\{f_0 + \sum_{j=1}^d f_j(X_j)\}]) = \sup M(p_m)\}$$

is single valued and upper hemicontinuous, it is continuous. Therefore $\lim_m f^m = f^* = \phi p = \lim_m \phi p_m$.

We now turn to the proof of (ii). We will first establish the form of the influence function of (5.19) and then show that its limit does not depend on the choice of the sequence of smooth functions $\{p_m\}_{m \geq 1}$. In order to obtain the influence function of $T_m(F)$, the maximizer of (5.19), we need to study the map

$$G(\epsilon, f) := \mathbb{E}_{(1-\epsilon)F + \epsilon\Delta}(\Psi(z, f)) - \lambda \nabla_f P_m(f),$$

where $\Delta_z$ denotes the mass point measure at $z$, $\epsilon \in [0, 1]$ and $\nabla_f$ denotes the derivative operator with respect to $f$ and $P_m(f) = \sum_{j=1}^d p_m(\|f_j\|_\beta)$. Indeed, $G(\epsilon, f)$ is the derivative of (5.19) when the expectation of loss is taken with respect to the mass point $\epsilon$-contamination distribution $(1-\epsilon)F + \epsilon\Delta_z$. Therefore by convexity $G(\epsilon, f) = 0$ if and only if $f = f_\epsilon = T_m((1-\epsilon)F + \epsilon\Delta_z)$. A straightforward application of the implicit function theorem proves the existence of a differentiable function $\epsilon \rightarrow f_\epsilon$, defined on a small interval $[-\varepsilon, \varepsilon]$ for some $\varepsilon > 0$, and with derivative

$$IF(z; T_m, F) = \frac{\partial f_\epsilon}{\partial \epsilon}(0) = -S_m^{-1}\{\Psi(z, f^*_m) + \lambda \nabla f P_m(f^*_m)\},$$

where $f^*_m$ is defined by $G(0, f) = 0$, $S_m = E[\Psi(Z, f^*_m)] + \lambda \nabla f P_m(f)$. The uniqueness of the limiting influence function follows from (B3) and the convergence of $p_m$ to $p$. Finally noting that the influence function of $f = B\beta$ is simply proportional to the influence function of $\beta$, the claimed form of the influence function of $T(F)$ follows along the lines of the proof of Proposition 3 in Chapter 4. □

### 5.8.2 Auxiliary lemmas for consistency results

The first two lemmas are known results in the spline smoothing literature. The first is given as fact 3 in Fan et al. [2011] citing the work of Zhou et al. [1998]. The claimed result is not explicitly stated by the latter authors, but can be deduced from the arguments they use to prove their Lemma 6.1. The second one is stated as Lemma 6.2 in Zhou et al. [1998]. Lemmas 3 and 4 are new but simple results that we will require for the proof of Theorem 1. The last two lemmas are due to Loh and Wainwright [2012] (Lemma 12) and Loh and Wainwright [2015] (Lemma 9).

**Lemma 1.** Under conditions (A1) and (A3), there exists constants $C' \geq c' > 0$, such that

$$c' m_n^{-1} \leq \lambda_{\min}\left\{\mathbb{E}(B_jB_j^T)\right\} \leq \lambda_{\min}\left\{\mathbb{E}(B_jB_j^T)\right\} \leq C' m_n^{-1}.$$

**Lemma 2.** Under conditions (A1) and (A3), there exists constants $C' \geq c' > 0$, such that

$$\{c + o(1)\} m_n^{-1} \leq \lambda_{\min}\left\{\mathbb{E}_n(B_jB_j^T)\right\} \leq \lambda_{\min}\left\{\mathbb{E}_n(B_jB_j^T)\right\} \leq \{C + o(1)\} m_n^{-1}.$$
Lemma 3. Let \( \hat{M}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_i(\beta) \) be a \( p \times p \) matrix, where \( p = m_n s \). Under conditions (A1)-(A6) we have with probability at least \( 1 - 2e^{-cn/p^2} \) that \( m_n \lambda_{\min} \left\{ \hat{M}(\beta) \right\} \geq C \) for all \( \beta \in \mathcal{O} \), where \( c \) and \( C \) are some positive constants. Furthermore \( m_n u^T (\hat{M}(\beta) - M(\beta)) u \rightarrow_p 0 \) for any \( p \) dimensional vector \( u \).

Proof: We need to show that \( m_n u^T \sum_{i=1}^{n} \hat{\Psi}_i(\beta) u > 0 \) for any \( p \) dimensional vector \( u \). Without loss of generality we assume \( \|u\|_2 = 1 \). It is easily seen that

\[
\begin{align*}
    u^T \sum_{i=1}^{n} \hat{\Psi}_i(\beta) u &= u^T \left( \sum_{i=1}^{n} \mathbb{E}\{\hat{\Psi}_i(\beta)\} + \sum_{i=1}^{n} \left[ \hat{\Psi}_i(\beta) - \mathbb{E}\{\hat{\Psi}_i(\beta)\} \right] \right) u \\
    &= c_1 m_n^{-1} n + \sum_{i=1}^{n} \tilde{\gamma}_i
\end{align*}
\]

where

\[
\tilde{\gamma}_i = \{B(X_i)^T u\}^2 \gamma_i - \mathbb{E}[\{B(X_i)^T u\}^2 \gamma_i] \quad \text{and} \quad \gamma_i = \partial^2 Q_M \{y_i, g^{-1}(X_i^T \beta)\} / \partial \eta^2.
\]

Note that the terms \( \tilde{\gamma}_i \) have mean zero, finite variance and by Condition (A6) are bounded by \( pK \), where \( K \) is some positive constant. Hence from Hoeffding’s inequality we have

\[
\mathbb{P} \left( \sum_{i=1}^{n} \tilde{\gamma}_i \leq c_1 m_n^{-1} n \right) \geq 1 - 2 \exp \left( - \frac{c_1^2 n^2}{2 np^2 K^2} \right) = 1 - 2e^{-cn/p^2}.
\]

It is readily seen that taking any constant \( t > 0 \) instead of \( c_1 \) in the last inequality, we also have \( m_n u^T (\hat{M}(\beta) - M(\beta)) u \rightarrow_p 0 \). \( \square \)

Lemma 4. Let \( \frac{1}{n} \sum_{i=1}^{n} \Psi_i(\beta^*) \) the \( p \) dimensional estimating equation (5.5) evaluated at the true \( \beta^* \). Under conditions (A1)-(A3), for some positive constants \( c \) and \( K \) we have

\[
\mathbb{P} \left( \left\| \frac{1}{n} \sum_{i=1}^{n} \Psi_i(\beta^*) \right\|_\infty \leq K \sqrt{\frac{\log p}{n}} \right) \geq 1 - 2e^{-c \log p}.
\]

Proof: Note that from Condition (A3) we have \( |\Psi_{ij}(\beta)| \leq A_1 |x_i| |x_j| \leq \tilde{c} < \infty \). Therefore letting \( c = K^2/(2\tilde{c}^2) \) the result follows immediately from Hoeffding’s inequality. \( \square \)

Lemma 5. For a fixed matrix \( \Gamma \in \mathbb{R}^{p \times p} \), parameters \( s \geq 1 \) and \( \delta > 0 \), suppose we have the deviation condition

\[
|v^T \Gamma v| \leq \delta, \quad \forall v \in \mathbb{B}_0(s) \cap \mathbb{B}_2(1),
\]

where \( \mathbb{B}_0(s) = \{ v v \in \mathbb{R}^p : \sum_{i=1}^{p} |v_i| \leq s \} \) and \( \mathbb{B}_2(1) = \{ v v \in \mathbb{R}^p : \|v\|_2 \leq 1 \} \). Then

\[
|v^T \Gamma v| \leq 27\delta (\|v\|^2 + \frac{1}{s} \|v\|^2), \quad \forall v \in \mathbb{R}^p.
\]

Lemma 6. If inequality (5.23a) holds for all \( \| \Delta \|_1 \in \mathbb{R}^p \) and \( n / \log p \geq 4 \tau_1 \| \Delta \|_1^2 \), then (5.23b) hold as well.

Lemma 7. Let \( g \) be a twice continuously differentiable function from \( \mathbb{R} \) to \( \mathbb{R} \) and let \( X \) and \( Y \) be some real valued random variables. Then

\[
\mathbb{E}[g(X)] - \mathbb{E}[g(Y)] = \mathbb{E}[(X - Y)g'(Y)] + \mathbb{E}[(X - Y)^2 \mathbb{E}_U \{(1 - U)g''(Y + (X - Y)U)\}],
\]
where $U$ is uniformly distributed in $[0, 1]$ and is independent of $X$ and $Y$.

**Proof:** Simple manipulations yield
\[
g(x) = g(x_0) + (x - x_0)g'(x_0) + \int_{x_0}^{x} (x - t)g''(t)dt \\
= g(x_0) + (x - x_0)g'(x_0) + (x - x_0)^2 \int_{0}^{1} (1 - u)g''(x_0 + (x - x_0)u)du \\
= g(x_0) + (x - x_0)g'(x_0) + (x - x_0)^2 \mathbb{E}[(1 - U)g''(x_0 + (x - x_0)U)].
\]

The result by replacing $x$ and $x_0$ by $X$ and $Y$ in the last equality and taking the expectation of the resulting expression. $\square$

### 5.8.3 Auxiliary results for Berge’s maximum theorem

We start by recalling some definitions from Berge [1997]. Let $\Gamma$ be a mapping (possibly multivalued) of the topological space $X$ to the topological space $Y$. We say that $\Gamma$ is **lower hemicontinuous at** $x_0 \in X$ if for each open set $O$ meeting $\Gamma x_0$ there is a neighborhood $N(x_0)$ such that $x \in N(x_0) \Rightarrow \Gamma x \cap O \neq \emptyset$. We say that $\Gamma$ is **upper hemicontinuous at** $x_0 \in X$ if for each open set $O$ containing $\Gamma x_0$ there is a neighborhood $N(x_0)$ such that $x \in N(x_0) \Rightarrow \Gamma x \subset O$. If $\Gamma$ is both lower and upper semi-continuous at $x_0$ we say that $\Gamma$ is **continuous at** $x_0$.

If $\Gamma$ is lower semi-continuous at each point of $X$ it is called **lower hemicontinuous in** $X$. We say that $\Gamma$ is **upper hemicontinuous in** $X$ if it is upper semi-continuous at each point of $X$ and if also $\Gamma x$ is compact for each $x \in X$. If $\Gamma$ is both lower and upper hemicontinuous in $X$, then it is called **continuous in** $X$.

**Lemma 8.** The mapping $\Gamma p = \{ f \in \varphi_n \mid \sum_{j=1}^{d} p(\|f_j\|_\beta) \leq K \}$ is continuous in $W^{2,2}(\mathbb{R})$.

**Proof:** Let $\|g\|_2 = \left( \int g^2(t)dt \right)^{\frac{1}{2}}$ and note that by Lemmas 1 and 3 in Xue et al. [2010] the norms $\|f_j\|_2$ and $\|f_j\|_\beta$ are equivalent on $\varphi_n$. Further recall that the space $W^{2,2}(\mathbb{R})$ is equipped with the norm $\|p\|_{2,2} = \|\nabla^2 p\|_2 + \|p\|_2$ and let $\|\tilde{p} - p\|_{2,2} < \epsilon$ for some $\epsilon > 0$. Since
\[
\sum_{j=1}^{d} \tilde{p}(\|f_j\|_\beta) - K = \sum_{j=1}^{d} \{ p(\|f_j\|_\beta) - p(\|f_j\|_\beta) + \tilde{p}(\|f_j\|_\beta) \} - K \\
\leq \sum_{j=1}^{d} \{ \tilde{p}(\|f_j\|_\beta) - p(\|f_j\|_\beta) \} \leq \sqrt{\sum_{j=1}^{d} \{ \tilde{p}(\|f_j\|_\beta) - p(\|f_j\|_\beta) \}^2} \\
\leq \|\tilde{p} - p\|_2 \leq \|\tilde{p} - p\|_{2,2} \leq \epsilon
\]

it follows that $\sup \{ \|f - g\|_2 \mid f \in \Gamma \tilde{p} \setminus \Gamma p, g \in \partial \Gamma p \} \leq d + \epsilon$, where $\partial \Gamma p = \{ f \in \varphi_n \mid \sum_{j=1}^{d} p(\|f_j\|_\beta) = K \}$ and $d = \sup \{ \|f - g\| \mid f, g \in \Gamma p \}$. This implies both upper and lower hemicontinuity since we can take $\epsilon$ as small as we want. $\square$

**Berge’s maximum theorem.** If $h$ is a continuous function in $Y$ and $\Gamma$ is a continuous mapping of $\mathcal{X}$ to $\mathcal{Y}$ such that, for each $x$, $\Gamma x \neq \emptyset$, then the function $M$ defined by $M(x) = \max \{ h(y) \mid y \in \Gamma x \}$ is continuous in $\mathcal{X}$ and the mapping $\phi$ defined by $\phi x = \{ y \mid y \in \Gamma x, h(y) = M(x) \}$ is a upper hemicontinuous mapping of $\mathcal{X}$ into $\mathcal{Y}$. 

**Proof:** Simple manipulations yield
\[
\sum_{j=1}^{d} \tilde{p}(\|f_j\|_\beta) - K = \sum_{j=1}^{d} \{ p(\|f_j\|_\beta) - p(\|f_j\|_\beta) + \tilde{p}(\|f_j\|_\beta) \} - K \\
\leq \sum_{j=1}^{d} \{ \tilde{p}(\|f_j\|_\beta) - p(\|f_j\|_\beta) \} \leq \sqrt{\sum_{j=1}^{d} \{ \tilde{p}(\|f_j\|_\beta) - p(\|f_j\|_\beta) \}^2} \\
\leq \|\tilde{p} - p\|_2 \leq \|\tilde{p} - p\|_{2,2} \leq \epsilon
\]

it follows that $\sup \{ \|f - g\|_2 \mid f \in \Gamma \tilde{p} \setminus \Gamma p, g \in \partial \Gamma p \} \leq d + \epsilon$, where $\partial \Gamma p = \{ f \in \varphi_n \mid \sum_{j=1}^{d} p(\|f_j\|_\beta) = K \}$ and $d = \sup \{ \|f - g\| \mid f, g \in \Gamma p \}$. This implies both upper and lower hemicontinuity since we can take $\epsilon$ as small as we want. $\square$
Chapter 6

Final remarks and discussion

In this last chapter we discuss some general issues regarding robust statistics in high dimensions. We start by discussing some extensions of the work provided in this thesis. We then overview some recent developments in the literature that are at the interface between robust and high dimensional statistics that give new insight into some well established statistical problems. We conclude this thesis by describing some major challenges for robust statistics that need to be addressed in future research.

6.1 Extensions

There are many problems related to the work presented in this thesis that deserve further investigation. In this section we describe with certain detail some questions that the author is currently working on.

6.1.1 Robust sure independence screening

We describe a screening procedure that can be of particular interest when we are faced with very large dimensional problems. The basic idea is to extend the sure independence screening idea of Fan and Lv [2008] to the GLM/GAM set up of this thesis. It consists on applying a crude but fast way of reducing the dimensionality of the problem and then using a more sophisticated technique to perform variable selection and estimation. One of the simplest versions of sure independence screening for the linear model is based on estimated parameters $\hat{\beta}_j^M$ obtained by running univariate regression models on each of the standardized covariates. The set of covariates that are kept after a marginal screening is

$$\bar{S}_n = \{1 \leq j \leq d : |\hat{\beta}_j^M| \geq \tau_n\},$$

for a given threshold $\tau_n$. Conditions under which the above screening rule keeps the subset of important variables with high probability have been studied in Fan and Lv [2008]. A non-exhaustive list of extensions of this methodology to more complicated models includes GLM, additive models and varying coefficients proposed respectively in Fan and Song [2010], Fan et al. [2011] and Fan et al. [2014b]. Following this line of work, a natural extension of sure independence screening to our set up is to consider marginal nonparametric fits $\tilde{f}_j^M$ obtained by solving

$$\arg\min_{f_0 \in \mathbb{R}, f_j \in \varphi_n} \mathbb{E}_{\mathcal{M}} \left[ Q_{\mathcal{M}}(Y, g^{-1}(f_0 + f_j(X_j))) \right],$$
where $\varphi_n$ is a space of centered spline functions, and then select a set of variables by the screening rule

$$\hat{S}_n = \{1 \leq j \leq d : \|f_j^M\|_n^2 \geq \tau_n\}. \quad (6.1)$$

Once we have reduced the dimensionality of the initial problem based on the screening rule (6.1) we can apply the more refined group lasso type of methods introduced in Chapter 5. Note that the success of this operation relies on whether the screening procedure does not mistakenly delete some important variables. In other words the procedure should have the sure screening property of Fan and Lv [2008]. It would be interesting to study the performance of this method both theoretically and numerically.

From a robustness standpoint a first step screening procedure such as (6.1) can be particularly appealing for dealing with the problem of propagation of outliers discussed in the last section of this chapter. Indeed, constructing robust estimators of regression with only one covariate can be easily done with the proposals available in the literature. This is dramatically different from the situation of multivariate regression where only a few recent attempts to construct robust estimators towards elementwise contamination have been made. In high dimensions there are currently no proposals and even a simple Mallow’s type of estimator seems to be bound to fail. Therefore reducing the dimensionality by an initial robust screening procedure will have the double advantage of being robust towards elementwise contamination and of giving some indication of which points might be outliers in a more refined second step estimation procedure. It is worth noting that another possible set of popular techniques that might be successfully adapted to robust high dimensional GLM/GAM are regression trees and random forests. Indeed, given that they build on sequential local fits existing low dimensional robust procedures could potentially be integrated successfully in these constructions.

### 6.1.2 Choice of the weight function

An important question that we have not been carefully discussed up to this point is the choice of the weight function for the covariates appearing in the definition of the robust quasilikelihood. A short discussion on this weight function in linear regression can be found in Carroll and Welsh [1988]. More generally in a fixed parameters set up, Cantoni and Ronchetti [2001] point out a few alternatives. A simple one can be to take $w(x_i) = \sqrt{1 - h_i}$, where $h_i$ is the $i$th diagonal element of the hat matrix $X(X^T X)^{-1}X^T$. A more sophisticated choice would be to take $w(x_i)$ to be the inverse of the Mahalanobis distance defined through a high breakdown estimate of the center and of the covariance matrix of the covariates.

All the above choices are naturally no longer possible when the number of observations $n$ is smaller than the number of parameters $d$. A few possible alternatives can be found in Hampel et al. [1986] p. 321. In particular one could for instance take $w(x)$ to be

$$w(x) = \min \left\{1, \frac{a}{\|Ax\|_2} \right\}, \quad \text{or} \quad w(x) = \min \left\{1, \frac{a^2}{\|Ax\|_2^2} \right\},$$

for parameters $a > 0$ and $A \in \mathbb{R}^{d \times d}$. These weight functions are reminiscent of those of Krasker and Welsch [1982]. They correspond to the $w(x)$ of optimal B-robust estimators of regression within the class of Mallows, subject to bounds on the sensitivity and self-sensitivity respectively. Both choices guarantee that $\|w(x)x\|_2$ is bounded for fixed choices of $a$ and $A$. Note that the second one can lead to higher order robustness as in La Vecchia et al. [2012] and robustness of the asymptotic variance functional as pointed
out in Markatou et al. [1991]. The effect of choosing one of these weight functions would be to shrink data points for which $\|x_i\|_2$ is large towards some elliptical shape defined by $A$. This choice is however not the most satisfactory given the discussion of the previous subsection and the problem of robustness towards elementwise contamination described in the last section of this chapter. Indeed a more meaningful choice of the $w(x_i)$ could be given after a first step screening procedure such as (6.1) has been carried on.

### 6.1.3 Tuning parameter selection

Most of the theoretical results derived in this thesis are concerned with fixed sequences of tuning parameters. They are therefore not directly applicable to the case where the amount of regularization is selected based on a data driven procedure. Results of this type are fairly common in the high dimensional literature and it is an issue that goes beyond the family of models presented in our work.

The theoretical results concerning data driven selection methods for the tuning parameter of sparse penalized estimators have focused on BIC-type of selection and K-fold cross-validation. Interestingly the type of results established for these two approaches to tuning reflect the usual way statisticians seem to think of BIC and cross-validation. The former being more a variable selection tool whereas the latter is more geared towards prediction. Indeed BIC-type tuning results establish variable selection consistency of penalized estimators while cross-validation tuning results focus on loss consistency or prediction consistency. Representative work on BIC type of selection includes Wang et al. [2007b], Wang et al. [2009], Zhang et al. [2010], Fan and Tang [2013] and Flynn et al. [2013]. An early result on K-fold cross-validation for a variant of the lasso can be found in Meinshausen [2007]. Recent work on cross-validation for the lasso includes Lecué and Mitchell [2012], Homrighausen and McDonald [2014] and Chatterjee and Jafarov [2015].

It follows from the above remarks that making connections between the existing BIC and cross-validation results would be an interesting step forward. Extending the scope of the existing results on data driven variable selection to our robust GLM and GAM seems to be a promising extension of our work given the numerous consistency results that we have already established. Accounting for data driven selection in our new construction of the influence function would also be a natural extension of our work.

### 6.1.4 Further topics

From a theoretical point of view one big open question is whether some differentiability results in the spirit of Clarke [1986] can be generalized to penalized high dimensional estimators. In particular appropriate von Mises expansions for penalized estimators would be extremely helpful for an easy assessment of their asymptotic biases and level of tests as in Heritier and Ronchetti [1994]. This might also lead to more insight into to the recent work of inference in high dimensions that has followed the work of Javanmard and Montanari [2014], Van de Geer et al. [2014] and Lockhart et al. [2014]. Indeed inference for sparse estimators is only starting to be theoretically explored and understood. Refitting after variable selection is a somehow closely related problem that has been discussed among others by Sun and Zhang [2012] and Belloni and Chernozhukov [2013]. Refitting with low dimensional techniques after model selection is another example, among many more, of issues explored in the high dimensional literature that can be naturally extended to robust counterparts.
6.2 Some recent developments

We give a selective overview of a few recent works that build on key ideas of robust statistics to obtain new insight into different statistical problems.

6.2.1 Exponential concentration without light tails

Robust statistics provides a formal framework for understanding how deviations from an assumed model affect statistical methods. One can however think of other ways of assuring some kind of robustness. In particular, it is fairly natural to seek for methods that extend the validity of some classical methods for larger classes of models. Although this approach does not always give robust estimators in the sense of Huber and Ronchetti [2009], the resulting estimates are natural appealing alternatives to classical estimators.

Concentration inequalities play an important role in the theory of high dimensional statistics and it is very frequent to assume that the distribution of the data generating process has subgaussian tails. This typically allows to obtain an exponential concentration inequality (Boucheron et al. [2013]). This is critical for instance for the empirical mean, since it does not necessarily concentrates exponentially unless a subgaussian assumption is made (Bubeck et al. [2013]). Unfortunately, assuming light tails can be a strong assumption especially in problems where the number of variables is very large. We will briefly describe two estimators of the mean of i.i.d random variables that only requires the existence of a finite variance while at the same time achieving the same exponential concentration one would obtain by averaging the mean of i.i.d. gaussian random variables. Both estimators use respectively some well established robust statistics tools: M-estimators and the median as a surrogate estimator of the mean parameter.

The first type of estimators that we describe was proposed by Catoni [2012]. He provided a family of M-estimators achieving exponential concentration around the mean. Specifically, he considers non-decreasing functions $\psi: \mathbb{R} \rightarrow \mathbb{R}$ satisfying

$$-\log(1 - x + x^2/2) \leq \psi(x) \leq \log(1 + x + x^2/2)$$

defining an estimator $\hat{\theta}$ of $E[Y_i] = \theta$ as the solution to

$$\sum_{i=1}^{n} \psi[\alpha(Y_i - \hat{\theta})] = 0.$$

Catoni [2012] shows that if $Var[Y_i] = v, n \geq 4 \log(1/\delta)$ and

$$\alpha = \sqrt{\frac{-2 \log(1/\delta)}{n(v + \frac{2v \log(1/\delta)}{n-2 \log(1/\delta)})}},$$

then, with probability at least $1 - \delta$,

$$\hat{\theta} \leq \theta + 2 \sqrt{\frac{v \log(1/\delta)}{n}}$$

and a similar bound holds for the lower tail. It is worth noting that this construction uses M-estimators with a tuning parameter that depends on $n$ and is typically much larger that the ones that would be advised by robust statistics. For instance Fan et al. [2016] recently applied this idea to high dimensional linear regression using the Huber
loss function. They showed that assuming errors with only two finite moments and taking
the tuning parameter proportional to $\sqrt{n/\log d}$, the resulting estimator admits the same
$\ell_2$ convergence rate as the optimal one attainable under subgaussian errors. Note that
a typical robust estimator would have instead a fixed tuning parameter that is chosen
based on efficiency considerations at a target model. In a neighborhood of the model one
can still expect exponential concentration but not necessarily around the true parameter.
This is the main difference with Catoni’s construction, i.e. the different approach to the
trades off robustness/consistency with exponential concentration.

The second estimator that we describe is the median of means estimator proposed
by Nemirovsky [1983] and further studied in Lerasle and Oliveira [2011], Bubeck et al.
[2013] and Joly and Lugosi [2016]. In particular the last paper extends the construction
and main results of the median of means estimators to general U-statistics. To be more
precise, let $\{Z_i\}_{i=1}^n$ be an i.i.d. sample with mean $\mu$, $V \leq n/2$ and $B = (B_1, \ldots, B_V)$ be
a regular partition of $\{1, \ldots, n\}$, i.e. $\text{card}(B_v) - n/V \leq 1$, where $\text{card}(B_v)$ denotes the
cardinality of $B_v$ for all $v \in \{1, \ldots, V\}$. Then the median of the mean estimators based on
the regular partition $B$ is $\bar{\mu}_B = \text{med}\{\bar{Z}_{B_v}, v = 1, \ldots, V\}$, where $\text{med}\{a_1, \ldots, a_n\}$ denotes
a median of $\{a_1, \ldots, a_n\}$ and $\bar{Z}_{B_v}$ is the mean of the $Z_i$s on the subsample $B_v$.
This simple construction is surprisingly capable of achieving exponential concentration around the true
mean parameter assuming only a finite variance. Intuitively the means of subsamples
make the estimator less dependent on symmetry while the median makes the solution
somehow more “robust” and yields fast concentration. A somehow related procedure has
been considered in the statistics literature in series of work by Brown et al. [2008], Cai
and Zhou [2009] and Cai et al. [2012]. In these papers a linear sequence model with errors
having 0 median is assumed. These authors propose to transform the data by taking
local medians on subsequent blocks of responses. This transformation can be viewed as
a ”gaussianization” of the responses and enables them to handle the transformed data
as gaussian sequence models. Although this approach is intuitively more closely related
to the spirit of robust statistics than the median of means estimator, the local median
construction relies heavily on the symmetry of the random variables being “gaussianized”.
Therefore it cannot in general guarantee that good mean estimators under asymmetry
are obtained.

6.2.2 Connections between M-estimators and high dimensional
statistics

We now discuss a nice connection between M-estimators of regression and high dimen-
sional statistics. There is a close relationship between penalized estimators and M-
estimators the linear regression context. Assume a model

$$y = X\beta + \gamma + \varepsilon, \quad (6.2)$$

where $y$ is a response vector of dimension $n$, $X$ is a $n \times d$ dimensional design matrix, $\varepsilon$
is an $n$ dimensional vector of i.i.d. noise and $\gamma$ a sparse $n$ dimensional parameter vector
with nonzero entries $\gamma_i$ when observation $i$ is an outlier. The mean-shift model (6.2) was
first considered in McCann and Welsch [2004] and McCann and Welsch [2007]. It can be
shown that minimizing the problem

$$\sum_{i=1}^n (y_i - x_i^T \beta - \gamma_i)^2 + \sum_{i=1}^n p_\lambda(|\gamma_i|) \quad (6.3)$$
over $\beta$ and $\gamma$, we obtain a parameter $\beta$ matching the one obtained from the minimization of

$$\sum_{i=1}^{n} \rho(y_i - x_i^T \beta).$$

(6.4)

We therefore have that for an estimate $(\hat{\beta}, \hat{\gamma})$ solving (6.3), $\hat{\beta}$ is an M-estimate associated with a solution of

$$\sum_{i=1}^{n} \psi(y_i - x_i^T \beta)x_i = 0$$

where $\psi(\cdot)$ denotes the derivative of $\rho(\cdot)$. This result was obtained in She and Owen [2011] and rediscovered recently in Donoho and Montanari [2016]. Note that for the specific case of Huber’s estimate, the corresponding penalty function in (6.3) is the lasso. An early derivation of this specific case was obtained in Sardy et al. [2001]. This rather unexpected relationship with M-estimators has many interesting consequences from a robustness viewpoint. We highlight some of them:

- The formulation of (6.3) opens the door for overparametrization as a new way of obtaining robust estimates. Keeping classical loss functions and taking care of contamination via a sparse parameter is very different in spirit from the typical robust approach where the role of bounded loss functions has played a predominant role. One remarkable example of such a proposal can be found in the work of Candès et al. [2011] in the context of low rank data matrix approximation.

- It is immediate that any low dimensional M-estimator of regression can be reformulated as a high dimensional problem, where the sparse parameter $\gamma$ will roughly have $\lfloor \epsilon n \rfloor$ nonzero components in a contamination neighborhood. There is a large potential for positive spillovers from the sparse modeling literature in robust statistics. Non asymptotic theory, fast algorithms and new insights into the basic problems of robust statistics, are to be counted as possible encouraging side effects.

- It is crucial to develop formal tools in order to characterize the robustness properties of penalized estimators in a high dimensions. This point is further discussed in the next section.

### 6.2.3 New asymptotics in high dimensions

Another very interesting recent new result regarding the old problem of minimizing (6.4) was first pointed out by El Karoui et al. [2013] and subsequently formally proved independently by El Karoui [2013] and Donoho and Montanari [2016]. Specifically, assuming a linear model (6.2) with $\gamma = 0$ and fixed design $X$, Huber [1973] showed that in a regime where $n, d \to \infty$ but $d/n \to 0$, the asymptotic distribution of the solution to (6.4) is normal $N(0, \Sigma)$ with asymptotic variance $\Sigma$ given by

$$\Sigma = V(\psi, F_\epsilon)(X^T X)^{-1},$$

(6.5)

where $V(\psi) = (\int \psi^2 dF) / (\int \psi dF)^2$ is the asymptotic variance functional of Huber [1964] and $F_\epsilon$ denotes the cumulative distribution function of $\epsilon$. Assuming that $d/n \to r \in (0, 1)$ the asymptotic distribution of M-estimators does not necessarily satisfy the well known formula (6.5). Indeed, El Karoui [2013] and Donoho and Montanari [2016] show that if
the entries $x_i$ of $X$ are i.i.d. subgaussian vectors independent of $\varepsilon_i$ then the asymptotic covariance of the M-estimator has the form

$$
\Sigma = V(\tilde{\psi}, \tilde{F}_\varepsilon)(E[X^T X])^{-1},
$$

where $V$ is still Huber’s asymptotic variance functional but $\tilde{\psi}$ is an effective score function that differs from $\psi$ in this new asymptotic regime and $\tilde{F}_\varepsilon$ is a convolution of the noise distribution with an extra Gaussian noise component that is unobserved in the classical setting.

Obviously this new finding implies that the existing formulas used for assessing confidence statements about M-estimates should be reviewed systematically in high dimensions. More importantly, the result implies that maximum likelihood are no longer the most efficient estimators in this scenario. Furthermore the usual Fisher information bound is not necessarily bounded since $I(\tilde{F}_\varepsilon) < I(F_\varepsilon)$, where $I$ denotes the Fisher information functional.

This striking new phenomenon will certainly bring new research and insight into high dimensional problems. There are already a few recent papers showing that many other well established results fail to hold when $d/n \to r \in (0, 1)$. Nevo and Ritov [2015] point out that in this regime M-estimators fail to be consistent in the euclidean norm and propose an empirical Bayes estimator that remedies this undesirable feature. Donoho and Montanari [2015] show that the Huber estimator’s asymptotic variance breaks down at a critical ratio of observations per parameter. In the classical situation this breakdown does not occur (Huber and Ronchetti [2009]). Finally, El Karoui and Purdom [2015] show that the usual residual and pairs bootstrap for regression M-estimates encounter problems. Namely, the residual bootstrap tends to give poor inference on $\beta$ with type I error rates much higher than desired while the pairs bootstrap fails in the opposite direction by yielding overly conservative results.

### 6.3 New challenges for robust statistics

#### 6.3.1 Propagation of outliers

Probably one of the main methodological challenges for robust statistics in high dimensions is to provide new appropriate contamination models. Assume we have a $n \times p$ data matrix of $n$ independent realization of a $p$ dimensional random variable $Z$. In a classical framework we could for instance have some normally distributed variable $Z \sim \mathcal{N}(\mu, \Sigma)$. The Tukey-Huber contamination model introduced in Tukey [1962] and Huber [1964] would consider distributions of the form $F_\varepsilon = (1 - \varepsilon)\mathcal{N}(\mu, \Sigma) + \varepsilon G$, where $G$ is an unspecified outlier generating distribution. It therefore assumes that a fraction $\varepsilon$ of the columns of the data matrix $Z$ is contaminated. This is however very stringent when $p$ is large for at least two reasons. First, although a whole observed $p$ dimensional vector could potentially be contaminated, it seems natural to expect that not all its entries contain non-representative information of the bulk of the data. In this case applying to the whole vector the usual downweighting schemes of robust methods will necessarily lead to losses of efficiency. Second, it is unrealistic to expect that only a tiny fraction of the observations is contaminated if the dimensionality is large. If instead of row-wise contamination of the data matrix implied by the Tukey-Huber contamination we consider cellwise, then a large number observations will contain corrupt entries if each of them is contaminated with a
probability $\epsilon$. This phenomenon was called propagation of outliers and was first studied by Alqallaf et al. [2009].

In order to handle cellwise contamination, some new robust procedures have been proposed in the context of low dimensional multivariate location and scatter estimation (Van Aelst et al. [2011], Van Aelst et al. [2012]), linear models (Oellerer et al. [2016]), linear mixed models (Agostinelli and Yohai [2014]), and cluster analysis (Farcomeni [2014]). Robust estimation of sparse covariance and inverse covariance matrices via pairwise estimates have been recently studied in Tarr et al. [2015], Oellerer and Croux [2014], Agostinelli et al. [2015] Loh and Tan [2015] and Han et al. [2014]. Refer to the latter paper for an account of rank based high dimensional covariance matrix estimation. Candès et al. [2011] and She et al. [2016] successfully recover the principal components of a high dimensional data matrix under arbitrary cellwise corruption by overparametrizing as discussed above. Formal characterizations of the robustness properties under these new scenarios is needed. We discuss this point in the following subsection.

6.3.2 Characterizing robustness and regularization

Formally speaking, cellwise contamination arises in the so-called independent contamination model introduced in Alqallaf et al. [2009]. This model considers a family of distributions

$$\mathcal{F}_\epsilon = \{ F_\epsilon : F_\epsilon \text{ is the distribution of } Z = (I - B_\epsilon)Z_0 + \epsilon \tilde{Z} \},$$

where $Z_0$ is distributed according to the target distribution $F$, $\tilde{Z} \sim G$, and $B_\epsilon = \text{diag}(B_1, \ldots, B_p)$, where the $B_j$ are independent Bernoulli variables with parameter $\epsilon$. This model formulation has nontrivial consequences even when the number of variables is fixed. Notably, it requires new and more involved derivations of (generalized) influence functions and breakdown points because general results given in Huber and Ronchetti [2009], Hampel et al. [1986] and Maronna et al. [2006] do not hold anymore under this setting.

Extending this model to high dimensions makes the robustness assessments even more difficult, partly because the parameters are assumed to have an increasing dimension. However, the main point is that regularization is required and there is not much work on theoretical characterization of robustness for these methods. Breakdown points under the Tukey-Huber contamination model have been computed in the context of low and high dimensional robust sparse linear models respectively in Wang et al. [2013] and Alfons et al. [2013]. The breakdown under cellwise contamination for a pairwise scatter estimator was obtained in Oellerer and Croux [2014]. The derivation of the influence function of lasso-type penalized estimators seems to be somehow more difficult. The reason being that these estimator are not defined implicitly by a system of estimating equations as most of the well known robust estimates. Some progress on this topic has been done in Wang et al. [2013] and Öllerer et al. [2015]. In Chapter 4, we provide a rigorous theoretical derivation by defining an influence function as the limit of a sequence of differentiable approximations and by showing that it can be viewed as a derivative in the sense of distribution theory. It remains to be explored extending to penalized M-estimators the generalized influence function for elementwise contamination considered in Alqallaf et al. [2009].

More generally speaking, the usual way of thinking of outliers as points that are far away from the majority of the data becomes meaningless in high dimensional space
because the curse of dimensionality makes every point be far from each other. One could say that the independent contamination model is a first attempt to define robustness in high dimensions and that it suggests one should still be particularly worried about anomalous data in low dimensions. Other types of contamination models can be advocated such as grouped contamination or mixtures of grouped and independent contamination models. Clearly introducing new contamination models will also require redefining and rederiving basic tools such as the influence function or the breakdown point and it could lead to introducing some new tools.

### 6.3.3 Revisiting old issues

We point out some old problems in robust statistics that should be revisited in light of the challenges of high dimensional statistics. One major issue to be tackled by future research is to develop appropriate algorithms for computing high dimensional robust estimators. Indeed the high computational costs of some robust procedures in small dimensional scenarios would make them prohibitive in high dimensions. At the same time, as pointed out by Tibshirani [2011], the large literature on fast implementations of sparse estimators has been one of the main driving forces of this research field. Hence it seems fundamental to develop fast algorithms that scale well when the dimension of the data at hand is very large.

A closely related problem is to theoretically understand the computational burden of robust estimators. In light of recent work, initiated by the work of Berthet and Rigollet [2013], establishing trade offs between computational complexity and statistical performance, it would be very interesting to explore whether there is some kind of trade off between robustness and computational complexity in the computation of high breakdown point estimators. Indeed, many popular high breakdown estimators of scatter such as the Stahel-Donoho estimator of Stahel [1981] and Donoho [1982], the minimum covariance determinant of Rousseeuw [1985] or the S-estimator of Davies [1987] are notoriously cumbersome to compute for large data sets.

A more philosophical question that seems to be worth revisiting is the close link between robust statistics and parametric statistics. To quote Huber [1981]: “Robust methods […] are destined to work with parametric models; the only differences are that the latter are no longer supposed to be literally true, and that one is also trying to take this into account in a formal way”. The rise of big data might lead to question or perhaps even change to some extent Huber’s premise. Indeed assuming that parametric models are not literally true is certainly a very sensible point of view when we only model a few variables or a response in a regression problem. Working with a central parametric model when we try to model thousands of variables simultaneously can however be very stringent. Suppose for instance that we are interested in estimating the covariance matrix of a thousand dimensional vector. The usual robust statistics approach for this problem would be to fit robust estimates of scatter that are consistent up to a scale factor for elliptical distributions. This approach guarantees, as Huber suggests, that if the generating process is not exactly elliptical but most of the data is well approximated by an elliptical distribution, then we will obtain a reasonable estimator of scatter. It is however apparent that as the dimensionality of the problem gets larger, even a seemingly mild elliptical assumption might already be too strong as it implies a lot of structure and in particular symmetry. More generally, although a typical robust approach takes into account that the model is only approximately true, the price to pay of adopting a parametric structure
might be too high in many situations.

Hopefully this will not necessarily entail that we will be more frequently in uncomfortable situations like the one described in Huber [1981]: “In the relatively rare cases where one is specifically interested in estimating the true population mean, there is little choice except to pray and use the sample mean”. One possible way of escaping this situation could be to consider some kind of sparsity. We might not need to “pray and use the sample mean” if in (most) very large dimensional problems the initial dimension can be effectively reduced, by simple flexible methods, to a more manageable size where the robust statistics toolbox can offer additional insight.
References


F. Han, J. Lu, and H. Liu. Robust scatter matrix estimation for high dimensional distributions with heavy tails. *(Manuscript)*, 2014.


