Approaches for Outlier Detection in Sparse High-Dimensional Regression Models

Doctoral Thesis
by
Luca Insolia

A dissertation submitted in May 2022 to the
Faculty of Sciences, Scuola Normale Superiore,
in partial fulfillment of the requirements for the
Doctoral Program in Data Science, XXXIII cycle

Supervisors:

Prof. Francesca Chiaromonte

Sant'Anna School of Advanced Studies

Pennsylvania State University

Prof. Marco Riani

University of Parma

Abstract

Modern regression studies often encompass a very large number of potential predictors, possibly larger than the sample size, and sometimes growing with the sample size itself. This increases the chances that a substantial portion of the predictors is redundant, as well as the risk of data contamination. Tackling these problems is of utmost importance to facilitate scientific discoveries, since model estimates are highly sensitive both to the choice of predictors and to the presence of outliers. In this thesis, we contribute to this area considering the problem of robust model selection in a variety of settings, where outliers may arise both in the response and the predictors. Our proposals simplify model interpretation, guarantee predictive performance, and allow us to study and control the influence of outlying cases on the fit.

First, we consider the co-occurrence of multiple *mean-shift* and *variance-inflation* outliers in low-dimensional linear models. We rely on robust estimation techniques to identify outliers of each type, exclude mean-shift outliers, and use restricted maximum likelihood estimation to down-weight and accommodate variance-inflation outliers into the model fit. Second, we extend our setting to high-dimensional linear models. We show that mean-shift and variance-inflation outliers can be modeled as additional fixed and random components, respectively, and evaluated independently. Specifically, we perform feature selection and mean-shift outlier detection through a robust class of nonconcave penalization methods, and variance-inflation outlier detection through the penalization of the restricted posterior mode. The resulting approach satisfies a robust oracle property for feature selection in the presence of data contamination – which allows the number of features to exponentially increase with the sample size – and detects truly outlying cases of each type with asymptotic probability one. This provides an optimal trade-off between a high breakdown point and efficiency. Third, focusing on high-dimensional linear models affected by mean-shift outliers, we develop a general framework in which L_0-constraints coupled with mixed-integer programming techniques are used to perform simultaneous feature selection and outlier detection with provably optimal guarantees. In particular, we provide necessary and sufficient conditions for a *robustly strong* oracle property, where again the number of features can increase exponentially with the sample size, and prove optimality for parameter estimation and the resulting breakdown point. Finally, we consider generalized linear models and rely on *logistic slippage* to perform outlier detection and removal in binary classification. Here we use L_0-constraints and mixed-integer conic programming techniques to solve the underlying double combinatorial problem of feature selection and outlier detection, and the framework allows us again to pursue optimality guarantees.

For all the proposed approaches, we also provide computationally lean heuristic algorithms, tuning procedures, and diagnostic tools which help to guide the analysis. We consider several real-world applications, including the study of the relationships between childhood obesity and the human microbiome, and of the main drivers of honey bee loss. All methods developed and data used, as well as the source code to replicate our analyses, are publicly available.

Publications

1. Insolia, L., Molinari, R., Rogers, S. R., Williams, G. R., Chiaromonte, F., and Calovi, M. (2022). Honey bee colony loss linked to parasites, pesticides and extreme weather across the United States. *Under revision*

2. Insolia, L. and Perrotta, D. (2023). Tk-merge: Computationally efficient robust clustering under general assumptions. In García-Escudero, L. A., Gordaliza, A., Mayo, A., Lubiano Gomez, M. A., Gil, M. A., Grzegorzewski, P., and Hryniewicz, O., editors, *Building Bridges between Soft and Statistical Methodologies for Data Science*, volume 1433 of *Advances in Intelligent Systems and Computing*, pages 216–223. Springer International Publishing

3. Insolia, L., Chiaromonte, F., Li, R., and Riani, M. (2021a). Doubly robust feature selection with mean and variance outlier detection and oracle properties. *arXiv preprint, arXiv:2106.11941*

4. Insolia, L., Kenney, A., Calovi, M., and Chiaromonte, F. (2021c). Robust variable selection with optimality guarantees for high-dimensional logistic regression. *Stats*, 4(3):665–681

5. Insolia, L., Kenney, A., Chiaromonte, F., and Felici, G. (2021d). Simultaneous feature selection and outlier detection with optimality guarantees. *Biometrics*, Forthcoming:1–12

6. Insolia, L., Chiaromonte, F., and Riani, M. (2021b). A robust estimation approach for mean-shift and variance-inflation outliers. In Bura, E. and Li, B., editors, *Festschrift in Honor of R. Dennis Cook: Fifty Years of Contribution to Statistical Science*, pages 17–41. Springer

7. Ferigato, C., Insolia, L., Perrotta, D., and Sordini, E. (2018). First report on the collaboration E.2/I.3 on the application of robust statistical techniques to the computation of position from satellite data in the context of project 346 GALILEO, WP 848 GAL-innova. Technical Report JRC110442, Joint Research Centre, European Commission

Conferences and talks

1. EMbeDS 2022 Workshop – Statistical approaches for the analysis of large and structured data. *Simultaneous Feature Selection and Outlier Detection with Optimality Guarantees.* EMbeDS, Sant'Anna School of Advanced Studies, Pisa, 5-6 Jul 2022

2. SMAC Talk – Stochastic Modeling and Computational Statistics. *A Robust Estimation Approach for Mean-Shift and Variance-Inflation Outliers.* Dept. of Statistics, Penn State University, University Park, PA, USA. Virtual event, 1 Apr 2022

3. 6th AIROYoung Workshop – Operation Research and Data Science in Public Services. *Discrete Optimization for Robust Feature Selection.* Roma Tre University, Engineering Department, Rome, Italy, 23-25 Feb 2022

4. 43rd Data Science Leuven Meetup – Leuven, Belgium. *Investigating the Main Drivers of Honeybee Colony Loss Through Discrete Optimization.* Virtual event, 17 Feb 2022

5. DMS Statistics and Data Science Seminar – Department of Mathematics and Statistics. *Parasitic Mites, Pesticides and Extreme Weather Linked to Honey Bee Loss: a Study Across the United States Through Multiple Open Data Sources.* Auburn University, Auburn, AL, USA. Virtual event, 2 Dec 2021

6. ICORS 2021 – International Conference on Robust Statistics. *Doubly Robust Feature Selection with Mean and Variance Outlier Detection and Oracle Properties.* TU Wien, Vienna, Austria, 20-24 Sep 2021

7. ICSA 2021 – International Chinese Statistical Association – Applied Statistics Symposium (invited talk). *Simultaneous Feature Selection and Outlier Detection with Optimality Guarantees.* Virtual event, 12-15 Sep 2021

8. JSM 2021 – Joint Statistical Meetings. *Doubly Robust Feature Selection with Mean and Variance Outlier Detection and Oracle Properties.* Virtual conference, 8-12 Aug 2021

9. ENAR 2021 – Eastern North American Region International Biometric Society – Spring Meeting. *Simultaneous Feature Selection and Outlier Detection with Optimality Guarantees.* Virtual conference, 14-17 Mar 2021

10. YES@IASI – The Young Experts Seminars. *Simultaneous Feature Selection and Outlier Detection with Optimality Guarantees.* Italian National Research Council (CNR-IASI). Virtual seminar, 24 Feb 2021

11. ERCIM 2020 – 13th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2020). *Simultaneous Feature Selection and Outlier Detection with Optimality Guarantees.* Virtual conference, 19-21 Dec 2020

12. CoMoS Climate Change – Complexity Modellers' Society. *The Main Drivers of Honeybees Loss in the Unites States.* EMbeDS, Sant'Anna School of Advanced Studies, Pisa, 15 Oct 2020

13. SMAC Talk – Stochastic Modeling and Computational Statistics. *A Robust Estimation Approach for Mean-Shift and Variance-Inflation Outliers.* Dept. of Statistics, Penn State University, University Park, PA, USA, 25 Oct 2019

Acknowledgments

This thesis would not have been possible without the help and support of a number of people. I am thankful beyond words of having Francesca Chiaromonte and Marco Riani as my supervisors. They shaped my professional and personal development in the most desirable way, and instilled in me their love for Statistics and scientific research. I am extremely grateful for their continuous encouragement, inspiration and precious advice. My academic journey would not have been the same without their mentorship and guidance.

I am thankful to Runze Li for the support he has provided me, and for his mentorship during my visiting period at Penn State University. A special thanks also goes out to Giovanni Felici, who offered invaluable advice and support over the years. My gratitude extends to Domenico Perrotta, Ana Kenney, Martina Calovi and Roberto Molinari for all of our fruitful discussions. You have been wonderful collaborators.

I thank our Ph.D. coordinator Dino Pedreschi and the members of the Ph.D. board for making all of this possible. I am grateful to Tommaso Cucinotta, Nicola Salvati and Monica Pratesi for accepting their role as panel members. I truly appreciate your insightful comments and suggestions.

I have been very lucky to have met many other wonderful collaborators, colleagues and friends, to whom I cannot express my gratitude for the wonderful time spent together. Thank you so much for bringing much happiness throughout these years. Warmest thanks go to Jisu for all the kind care and support.

Last but not least, I am deeply grateful to my family for their enduring encouragement and support. I am glad that you have always been there for me.

Contents

List of Figures

List of Tables

"Although we often hear that data speak for themselves, their voices can be soft and sly"

F. Mosteller, S. Fienberg, R. Rourke (1983)

Chapter 1

Introduction

Contemporary social and scientific endeavors produce a steadily increasing amount of data that can provide invaluable information, but also poses critical challenges. From a statistical standpoint, contemporary data demand special attention since they frequently do not fit well within the traditional paradigm – motivating the development of more sophisticated theories and tools. In particular, regression studies often include a vast number of candidate predictors, potentially far larger than the available sample size. This raises the odds that a substantial portion of the predictors be redundant, as well as the risk of data contamination – that is, outlying cases that may come from a different generating model. Therefore, the development of *robust model selection* techniques, which are still in their infancy, is of utmost importance in a variety of research fields and applied domains.

This chapter highlights the motivation for our research, and summarizes some background information, as well as our main objectives, contributions, and the structure of the thesis.

1.1 Motivation

The so-called "big-data revolution" has impacted virtually all parts of our society, sectors of our economy, and scientific domains – imposing new and fascinating challenges to the field of Statistics. Large and increasingly complex data are collected and analyzed at an astonishing rate, thus motivating the development of effective

methodologies in terms of predictive performance, interpretability, robustness/stability, and computational burden.

Regression analysis is a widespread approach for fitting predictive models to data. Indeed, investigating the relationship between an outcome of interest and a set of predictors is a very old idea, which is still pivotal in numerous domains. Two contrasting goals are typically pursued in this setting: (i) good prediction accuracy, and (ii) high interpretability – e.g., judging the strength/contribution of different subsets of predictors in explaining the response variable. Finding a good balance between these two goals is non-trivial. The problem is exacerbated in high-dimensional models containing noisy predictors and/or data contamination, which can make traditional estimation approaches, such as maximum likelihood estimators, unfeasible or ineffective.

1.1.1 Why model selection?

In high-dimensional regression models, containing a large number of candidate predictors (also called features or explanatory variables), it is likely that only a subset of such predictors have an actual relationship with the response. The exclusion of irrelevant or redundant features from the model may facilitate interpretation and scientific discoveries, and improve the prediction of unobserved outcomes.

In a broad sense, *model selection* is the process of selecting among a group of statistical models. This is often referred to also as sparse estimation, feature or variable selection, or feature elimination. The selection of parsimonious model representations can ease the interpretation of results, and at the same time reduce estimation variability (i.e., ameliorate overfitting) while ensuring high predictive power (Hastie et al., 2015). The so-called *sparsity assumption*, which postulates that only a fraction of the predictors can be effectively used in modeling the response variable, is realistic in several domains of application. For instance:

- Gene expression: it is expected that only a limited number of genes are associated with any given pathology. Based on data on the expression of 4718 genes on samples from 349 patients affected by 15 forms of cancer, Hastie

et al. (2015, p. 4) showed that only 254 genes suffice in predicting each type of cancer with an overall accuracy of 90%.

- Portfolio selection: Markowitz model aims at constructing a financial portfolio with maximal expected return and minimum variance, with an upper bound on the number of open positions. Bertsimas et al. (2021a) provided an optimal sparse solution (e.g., selecting only 5 stocks) for the $\approx$ 3200 securities included in the Wilshire 5000 index.

- Retail sales: consumer demand for products depends on marketing strategies that are pursued across product categories. Analyzing 926 products in 15 food categories for 320 weeks, Ma et al. (2016) showed that identifying relevant features at the intra- and inter-category level improves sales forecasting accuracy and helps highlight complementarity and substitution relationships.

Intuitively, model selection techniques identify the "most significant" subset of predictors, assuming that the majority of them have a negligible effect on the response variable. Their development dates back to the works of Cauchy in 1835 (Seal, 1967), and relies on the idea of fitting different sub-models, which have to be compared according to a predetermined criterion. Ideally, one should rely on the combinatorial evaluation of all possible models of any given size (known as *best subset selection*), which is typically unfeasible for large problems by a naïve brute-force algorithm. Indeed, until the last decade, this was considered computationally intractable for regressions containing more than 40 predictors (Hastie et al., 2009, p. 57). For this reason, several approximations of best subset selection were introduced, such as stepwise procedures that select variables in a path-dependent way in order to effectively reduce the search space (Miller, 2002).

Many researchers remained skeptic about the feasibility to efficiently and optimally solve model selection problems. As stated by Plackett: "*If variable elimination has not been sorted out after two decades of work assisted by high-speed computing, then perhaps the time has come to move on to other problems*" (Miller, 1984). However, continuous penalization methods such as the *Lasso* (Tibshirani, 1996) provided an alternative and computationally leaner avenue to tackle the problem and, more

recently, advances in *mixed-integer programming* (MIP, Bertsimas et al. 2016) allowed researchers to approach best subset selection with much improved efficiency. Notwithstanding all this progress, model selection is still an open problem in Statistics and is receiving renewed attention. Both "soft" feature selection procedures (based on continuous penalties) and "hard" ones (which rely on combinatorial enumeration) achieve varying degrees of sparsity, accuracy and computational efficiency under different scenarios (Hastie et al., 2020).

An optimal sparse estimator, i.e., one computed on the truly relevant (but unknown) set of predictors, is often called an *oracle estimator*. In order to retrieve its solution, several *oracle properties* have been developed in the literature (Fan and Li, 2001; Bradic et al., 2011; Bühlmann and Van De Geer, 2011; Fan et al., 2014a). Under suitable conditions, they guarantee that a certain estimator asymptotically behaves as if the set of relevant features were known in advance (weak oracle property), and model parameters are estimated efficiently (strong oracle property). Notably, nonconvex penalization methods require less stringent assumptions to achieve oracle properties compared to convex methods (Zhang and Zhang, 2012). Focusing on nonconcave penalties, Fan and Peng (2004) extended oracle properties to settings where the number of features diverges with the sample size itself.

1.1.2 Why statistical robustness?

Statistical models provide an imperfect representation of the real phenomena under investigation, whose approximation accuracy is tied to the underlying (albeit often implicit) assumptions. Therefore, it is of utmost importance to study the stability/robustness of estimation results against "reasonable" deviations from model assumptions.

In a broad sense, *robust statistics* encompasses a number of modeling techniques that can pinpoint model deficiencies, and studies robust estimators that behave similarly to classical methods (e.g., maximum likelihood estimators) when the underlying assumptions are fully satisfied, but at the same time provide reliable statistical results under model misspecification and/or the presence of data contamination. Robust methods effectively reveal data structures that would remain hidden otherwise,

and the detection of spurious observations often provides relevant domain-specific insights. For instance:

- Anomaly detection: not all observations in a data set often come from the same population or data generating mechanism. For the 47 stars contained in the star cluster CYG OB1, Rousseeuw and Hubert (2018) analyzed the Hertzsprung–Russell diagram (the logarithm of their light intensity against the logarithm of their surface temperature) showing that robust methods effectively detect giant stars, whereas classical techniques cannot distinguish them from the rest.

- Anti-fraud: data on economic transactions often contain information on fraudulent behaviors. Based on international trade data, Perrotta et al. (2020) developed robust procedures for the estimation of *"fair" import prices* which highlight possibly fraudulent transactions and in turn aid customs operations. Relatedly, Rousseeuw et al. (2019) detected potential frauds and level shifts in time series data of imports into the European Union.

- Signal processing: a signal transmission can be corrupted by natural or adversarial occurrences. In the context of global navigation satellite systems, robust estimation techniques can be effective in coping with different kinds of interference, such as multipath (Medina et al., 2019), urban canyons (Gaglione et al., 2017), and jamming (Borio, 2017).

Interestingly, the term "robust" was popularized in the statistics community by Box (1953), but some primitive robust estimation procedures can be traced back to 2,400 years ago, during the Peloponnesian war in Ancient Greece (Cerioli et al., 2011). The theoretical framework for robust statistics was established in the pioneering works of Tukey (1960), Huber (1964), Hampel (1968) and the ever-growing literature on the subject has been extensively discussed in Cook and Weisberg (1982); Rousseeuw and Leroy (1987); Barnett and Lewis (1994); Belsley et al. (2004); Maronna et al. (2006); Morgenthaler (2007); Huber and Ronchetti (2009); Hampel et al. (2011).

Since statistical robustness refers to "*insensitivity against small deviations from the assumptions*" (Huber, 1996, p. 1), robust methods have to consider a hypothetical model and a specific meaning for *smallness* (Huber and Ronchetti, 2009, p. 21), which "*should be statistically meaningful, not just mathematically convenient*" (Hampel et al., 2011, p. 9). In this thesis, we consider as (small) deviations the presence of influential outlying observations; that is, of units that do not come from the same population from which the majority of the sample is drawn, and for which the modeling assumptions do not hold. This is a rather realistic scenario, that can be due to errors in data recording (human or machine errors) or to the fact that the sample at hand is a mixture of different populations (e.g., healthy and unhealthy individuals or fraudulent and non-fraudulent transactions). In particular, one or more observations are defined as *influential* if – singularly or jointly – they have a larger impact on some model estimates of interest compared to the remaining cases in the sample (Belsley et al., 2004). In a regression setting, influential points (often simply called outliers) deviate from the "true" conditional distribution of the response variable. For instance, in linear regression, they correspond to points that have a large distance from typical response values or a large leverage, i.e., a large distance from typical predictors' values (Hoaglin and Welsch, 1978; Cook and Weisberg, 1982). Outlier detection and treatment is essential since they can affect estimates and inferential results provided by non-robust techniques – indeed, even a single outlying case can disrupt the performance of maximum likelihood methods.

Robust estimation and *outlier detection* are traditionally considered as complementary (and at times also alternative) tools in dealing with influential units. They have the same goal but proceed from opposite directions. The former focuses on the "core" of the data, suppressing ill-effects from influential units which can be identified through the analysis of robust residuals. The latter focuses on aberrant cases, which are identified through perturbations on a classical fit (e.g., case-deletion methods based on influence measures, Atkinson 1985; Chatterjee and Hadi 1988). Robust estimators, which are traditionally considered superior (Huber and Ronchetti, 2009), can be categorized in two main groups: *soft trimming methods*, that typically down-weight all points, and *hard trimming procedures* which provide binary weights

(Cerioli et al., 2016, 2018). Importantly, robust estimators have to face a trade-off between high tolerance to the presence of outliers and high efficiency. The former is typically measured by the notion of *breakdown point* (Donoho and Huber, 1983), which captures the estimator stability when a fraction of observations is replaced by arbitrary values, and motivates the use of non-convex loss functions (Bernholt, 2006). The latter measures the estimation variability under the "true" model, which is smaller when estimation is based on as many typical units as possible. Indeed, Anscombe described robustness through an insurance metaphor, because one needs to "*sacrifice some efficiency at the model, in order to insure against accidents caused by deviations from the model*" (Huber and Ronchetti, 2009, p. 5).

1.1.3 The need for robust model selection

As the model dimensionality grows, one can reasonably expect that both the prevalence of irrelevant features and the risk of data contamination raise as well. *Robust feature selection* procedures aim at balancing prediction accuracy and sparsity for uncontaminated data, while maintaining stability in the presence of influential units (Smucler and Yohai, 2017). Notably, only in the last decades some robust feature selection procedures have been introduced in the literature; as claimed by Freue et al. (2019) "*the development of penalized robust estimation methods is still in its early stages*".

Existing methods rely on soft penalizations and replace maximum likelihood estimation with a robust counterpart. Both hard (Alfons et al., 2013; Kurnaz et al., 2017) and soft (Smucler and Yohai, 2017; Freue et al., 2019) trimming approaches have been explored, with evidence that they both do tolerate data contamination, and convex penalties have generally been used to enforce sparsity into model estimates.

1.2 The thesis in a nutshell

In this thesis we focus on the study of sparse, high-dimensional regression models affected by different forms of data contamination. We are interested in the devel-

opment of sound and computationally lean model selection techniques which can detect and limit the influence of outlying cases on the fit, as well as in the study of their theoretical properties and their applications to real-world problems.

1.2.1 Objectives and contributions

To establish our framework, we first focus on low-dimensional linear regression models where outliers affect both the response and the predictors. Here, as opposed to the existing dichotomy in the use of hard or soft trimming estimators, we are interested in a principled combination of these approaches to exploit their strengths. Specifically, in Insolia et al. (2021b) we assume that multiple outliers can arise simultaneously from a *mean-shift* (MSOM, Beckman and Cook 1983) and a *variance-inflation outlier model* (VIOM, Cook et al. 1982) – which lead to the exclusion or the down-weighting of outlying cases, respectively. We develop a novel procedure based on a forward search (Atkinson and Riani, 2000) and restricted maximum likelihood estimation (Harville, 1977) to detect and treat both types of outliers, while attributing full weight to non-outlying cases. We demonstrate the effectiveness of our procedure through Monte Carlo simulations and real-world applications (e.g., loyalty cards data), and introduce graphical diagnostic tools which help to guide the analysis. To our knowledge, this is the first study tackling the presence of multiple VIOM outliers, as well as considering the co-occurrence of multiple MSOMs and VIOMs, since these contamination mechanisms are traditionally considered as alternatives.

Unlike the VIOM, the MSOM has also been exploited in the context of high-dimensional linear models during the last decade (albeit not explicitly at times) through its equivalence with hard trimming methods (Alfons et al., 2013; Kurnaz et al., 2017). However, existing methods are sub-optimal in terms of feature selection and outlier detection, as they rely on heuristic algorithms based on resampling to solve the non-convex robust loss and use convex penalties to enforce sparsity. We are thus interested in developing an optimal estimation strategy based on combinatorial enumeration which could be tackled by modern discrete optimization tools. Specifically, in Insolia et al. (2021d) we consider high-dimensional regression models

contaminated by multiple MSOM outliers affecting both the response and the design matrix. We develop a general framework and use L_0-constraints coupled with mixed integer programming to simultaneously perform feature selection and outlier detection with provably optimal guarantees – meaning that the global optimum is indeed achievable, and even if the algorithm is stopped before convergence one can certify the goodness of its solution. Robust estimation and outlier detection are thus equivalent within this framework (i.e., fitting only the "best" h points out of n is equivalent to excluding the "worst" $n - h$ points) since no resampling technique is involved. In terms of theoretical properties, we provide necessary and sufficient conditions for what we call *robustly strong oracle property* – meaning that one can recover the true sets of relevant features and outlying cases, with the number of features allowed to increase exponentially with the sample size. We also prove the high breakdown point and optimality of the regression coefficient estimates produced by our approach. Notably, our proposal provides stronger theoretical results under weaker assumptions compared to existing methods. We further show its superior performance through Monte Carlo simulations and real-world applications, where we use it to study the relationships between childhood obesity and the human microbiome. Computationally efficient procedures to tune integer constraints and warm-start our algorithm are also developed.

Next, we consider the co-occurrence of multiple MSOM and VIOM outliers in high-dimensional linear models (Insolia et al., 2021a). Building upon the work of Fan and Li (2012), we show that they can be modeled as additional fixed and random components, respectively, and evaluated independently. Our proposal performs feature selection while detecting and down-weighting VIOMs, detecting and excluding MSOMs, and retaining non-outlying cases with full weights. While feature selection and MSOM detection are performed through a robust class of nonconcave penalization methods, VIOM detection is based on the penalization of the restricted posterior mode for an over-parametrized model. To our knowledge, both the theory behind the detection and treatment of multiple VIOM outliers (possibly with MSOMs) and the study of sparse models affected by VIOMs have not been developed to date. Also the use of penalization methods for VIOM detection had not been explored be-

fore – and it turns out to be computationally lean and very effective. The resulting approach satisfies a *doubly robust strong oracle property* for feature selection in the presence of data contamination – which allows the number of features to exponentially increase with the sample size – and detects truly outlying cases of each type with asymptotic probability one. Our procedure improves estimation of the error variance, provides a bridge between robust methods based on convex penalties and the use of combinatorial enumeration, as well as an optimal trade-off between a high breakdown point and efficiency. We also introduce computationally lean heuristic algorithms, and we demonstrate finite-sample performance through synthetic data (simulations) and real-world applications related to the Boston housing market and glioblastoma gene expression data.

For classification problems, there are even fewer approaches for robust model selection than for linear regression, and those that exist have comparable limitations. In Insolia et al. (2021c) we extend our approach based on L_0-constraints and discrete programming (Insolia et al., 2021d) to high-dimensional logistic regression models affected by data contamination. Here, we rely on an over-parametrized *logistic slippage model* (Bedrick and Hill, 1990), which mimics the MSOM and leads to the removal of outlying cases from the fit. The presence of a non-quadratic loss function requires a different approach to the problem. We propose a *mixed integer conic programming* formulation to solve the underlying double combinatorial problem of simultaneous feature selection and outlier detection in a framework that allows one to pursue optimality guarantees. We show the superior performance of our proposal through Monte Carlo simulations and, in a real-world application, use it to investigate the main drivers of honey bee (*Apis mellifera*) overwintering loss using data from the state of Pennsylvania (USA).

1.2.2 Outline and replicability

This thesis consists of four self-contained contributions produced during the PhD program in Data Science – a consortium program coordinated by the Scuola Normale Superiore, in partnership with the Sant'Anna School of Advanced Studies, the Italian National Research Council (CNR), the University of Pisa, and the IMT School for

Advanced Studies of Lucca. The work presented benefited also from a visiting period at The Pennsylvania State University (University Park, PA, USA), and a research fellowship funded by the Sant'Anna School of Advanced Studies (Pisa) and the Institute for Systems Analysis and Computer Science of the CNR (Rome).

The remainder of the thesis is structured as follows:

- **Chapter 2**: we focus on low-dimensional linear models and rely on the mean-shift and variance-inflation outlier models to combine "hard" and "soft" trimming estimators. This contribution was published in Insolia et al. (2021b).

- **Chapter 3**: we study "hard" trimming methods for sparse, high-dimensional linear models, and we develop mixed integer programming techniques for simultaneous feature selection and outlier detection. This contribution was published in Insolia et al. (2021d).

- **Chapter 4**: we consider sparse, high-dimensional linear regression models affected by the co-occurrence of multiple mean-shift and variance-inflation outliers, and develop a doubly robust class of nonconcave penalization methods. This contribution is currently available as a preprint in Insolia et al. (2021a).

- **Chapter 5**: we extend our approach to logistic regression models, where we simultaneously perform feature selection and outlier detection through mixed integer conic programming techniques. This contribution was published in Insolia et al. (2021c).

- **Chapter 6**: we provide some final remarks, highlighting a number of potential extensions of our work and discussing some future challenges for robust statistics in high-dimensions.

Source code for the implementation of all the procedures proposed in this thesis, as well as code to replicate all simulation studies and the applications presented throughout Chapters 2–5, are publicly available at `https://github.com/LucaIns`.

"I see no way of drawing a dividing line between those [observations] that are to be utterly rejected and those that are to be wholly retained"

Daniel Bernoulli (1777)

Chapter 2

A Robust Estimation Approach for Mean-Shift and Variance-Inflation Outliers

This chapter is based on: Insolia, L., Chiaromonte, F., and Riani, M. (2021b). A robust estimation approach for mean-shift and variance-inflation outliers. In Bura, E. and Li, B., editors, *Festschrift in Honor of R. Dennis Cook: Fifty Years of Contribution to Statistical Science*, pages 17–41. Springer. Reprinted/adapted by permission from *Springer Nature Customer Service Centre GmbH*: Springer, Festschrift in Honor of R. Dennis Cook by Bura E., Li B. (eds), Copyright 2021.

Reproducible and documented code for this chapter is available at: `https://github.com/LucaIns/VIOM_MSOM`.

2.1 Introduction

We consider procedures for detecting and treating outliers in a regression setting. With a slight abuse of language, we call outliers those observations that affect estimation and inference on the parameters of the regression model utilized in the analysis. This, which is also known as influence (Cook and Weisberg, 1982), depends on how the position of an observation in the predictor space (leverage) and along the response range (vertical departure from the regression surface) combine

to make it "extreme" relative to the bulk of the data. Critically, it depends also on the presence of other observations whose influence may mask or swamp that of the observation under consideration. In a way, the notion of influence depends on the overall collection of observations at hand, the model being fitted, and the specific parameters of interest.

Simple and, where possible, automated and computationally inexpensive approaches to detect and treat outliers are very important in the practice of regression. These approaches often rely on postulating an outlier generating mechanism and provide a weighting system for the observations. In full generality, weighting here includes removing or trimming observations (attributing a weight of 0), keeping them in the analysis as they are (attributing a weight of 1) and, between the two extremes, down-weighting them to control their influence. The literature on these subjects is extensive (Barnett and Lewis, 1994; Beckman and Cook, 1983; Belsley et al., 2004; Maronna et al., 2006; Huber and Ronchetti, 2009; Atkinson, 1985; Chatterjee and Hadi, 1988; Hampel et al., 2011). In particular, two main frameworks have been utilized: the *mean-shift outlier model* (MSOM) and the *variance-inflation outlier model* (VIOM).

MSOM, which assumes that outliers are generated by shifts in mean (Cook and Weisberg, 1982), has been the historically predominant framework. Traditionally individual mean shift outliers were detected through *deletion-based approaches*, computing prediction residuals (Cook, 1977). Using *maximum likelihood estimation* (MLE) the individual outlier position corresponds to the unit with largest Studentized residual based on an *ordinary least squares* (OLS) fit. Notably, MLE for a MSOM can be reformulated as OLS for the underlying regression model augmented by a dummy for the presence/absence of the observation under evaluation. The estimated coefficient for the dummy is the prediction residual, and the corresponding t-statistic is the externally Studentized residual (also called *deletion residual*; Atkinson (1985)) which tests the "outlying-ness" of an individual observation. This formulation speeds up computation from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2)$ based on the Sherman-Morrison formula, which is critical when many individual tests have to be performed (Chatterjee and Hadi, 1988).

Because of masking and swamping effects, individual deletion residuals can fail to detect multiple MSOM outliers, and thus lead to sub-optimal regression estimation and inference. A simple extension of deletion-based approaches to groups of observations implies a combinatorial increase in computation – to explore the deletion of all subsets of any given size, and a range of sizes, say from 1 to $n/2$ (if one is willing to assume that at least half of the observations are not outliers). Much like best subset feature selection, this was computationally intractable in the 80s and 90s for realistically large problems, and triggered the development of proposals based on penalized fits of a model augmented by n dummies (one for each observation; Menjoge and Welsch 2010; She and Owen 2011; McCann et al. 2006). It is also important to remark that MSOM detection is customarily followed by *outlier removal*; that is, outliers are attributed a weight of 0 in estimation and inference on the regression parameters.

VIOM is generally considered as an alternative to the MSOM framework. It assumes that outliers are generated by an inflation in the error variance (Cook et al., 1982; Thompson, 1985). In a way, *VIOM is the "random effect" version of MSOM*; instead of being generated by a fixed effect (mean shift, to be estimated) the outlier is generated by a random effect with a certain variance (again, to be estimated). In fact, an equivalent parametrization of the regression problem in the presence of variance-inflation outliers can be given in the form of a mixed-effects linear model. In the VIOM framework *outliers are not removed*; they are retained in a *weighted fit*, where the weight for each observation is inversely proportional to the variance of its random effect (i.e. its variance inflation). In general, because it uses down-weighting instead of discarding (or failing to discard) observations, VIOM can achieve higher accuracy than MSOM in estimation and inference on the regression parameters.

Assuming that the data comprise (at most) a single variance-inflated outlier, its detection and treatment (i.e. its variance and thus optimal weight estimation) can be performed through a closed-form MLE. Notably, on a given data set, the outlier identified through individual deletions in MSOM and the outlier identified through MLE in VIOM need not to coincide (unless this observation has both the largest absolute residual and the largest absolute Studentized residual). This illustrates

how the statistical handling of outliers can depend on our assumptions concerning the mechanism that generates them. However, if *restricted maximum likelihood estimation* (REMLE, Harville 1977) is used instead of MLE to detect the outlier and estimate its variance, this coincides with the MSOM outlier identified through individual deletions.

The extension of MLE or REMLE approaches to multiple VIOM outliers also poses computational issues, because the closed-form expressions used in the case of individual outliers cannot be straightforwardly generalized. Most recently, leveraging the mixed-effects linear model formulation, Gumedze (2019) showed that a closed-form REML for variance estimation cannot be derived in the case of multiple outliers, but numerical estimation procedures used for mixed models can be applied. This is promising, though still computationally costly in problems of realistic size.

In summary, to date, penalization approaches offered progress toward the computationally viable detection of multiple MSOM outliers – which are then removed from the regression. On the other hand, the computational viability of state-of-the-art techniques for detecting and down-weighting multiple VIOM outliers is still a concern.

Notably, the literature on outliers has always been closely related to that on *robust estimation* (Maronna et al., 2006; Huber and Ronchetti, 2009; Hampel et al., 2011); outliers can be thought of as a form of (adversarial) perturbation of the data, due to either errors in data recording or contamination with statistical units belonging to a population different from the one of interest. The mechanism generating *contamination* is critical for studying the properties of robust estimators and has an important role also in our developments.

The traditional paradigm is the case- (or row-) wise contamination mechanism, also known as *Tukey-Huber mixture model*. An outlier is thought of as comprising values that do not conform with the bulk of the data in all its dimensions. Unlike the case of a MSOM outlier, the contamination can affect both response and predictors, i.e. the design matrix. In full generality, one assumes that the mixture data distribution is $Z \sim (1-\epsilon)F + \epsilon C$, so that each individual observation is drawn from the "true" distribution F with probability $(1-\epsilon)$ and from a contaminating distri-

bution C with probability ϵ (the scheme can be extended to multiple contaminating components).

In a way, difficulties in dealing with multiple outliers were a key motivation for the development of high *breakdown point* (BdP) robust estimators, which produce good estimates without assumptions on the nature of the outliers in the data (Donoho and Huber, 1983; Rousseeuw, 1984). Robust estimation is characterized by a trade-off between the reduction of biases due to outlier removal and the increase in estimates variability, or inefficiency, due to (possibly) not leveraging the entire information contained in the data. Indeed, Anscombe compared robustness to an insurance policy where one "*sacrifices some efficiency at the model, in order to insure against accidents caused by deviations from the model*" (Huber and Ronchetti, 2009, 5). In order to obtain a good compromise, state-of-the-art robust methods employ a preliminary high-BdP estimator (that is, a possibly inefficient estimator capable of withstanding high levels of contamination), whose outcome is then refined with a second high-efficiency estimator to retain in the fit as much "uncontaminated" information as possible (Maronna et al., 2006; Rousseeuw and Leroy, 1987). So-called soft-trimming methods down-weight all units and implicitly account for both VIOM and MSOM outliers, while hard-trimming methods provide binary weights and are specialized to account only for MSOM outliers (Cerioli et al., 2016).

Relatedly, outlier problems formulated as mixture contamination models were studied also in the Bayesian literature. De Finetti (1961) investigated a general framework, and Box and Tiao (1968) focused on a VIOM where both the total fraction of contamination and the inflation parameter were assumed to be known constants.

We propose to combine techniques from robust estimation with techniques designed for outliers down-weighting into a novel approach that detects and treats multiple outliers in an effective and computationally viable fashion. Our approach can be fully automated; however, since it utilizes an iteration, we describe criteria to monitor its progression through a graphical diagnostic tool. Importantly, our approach is applicable also to scenarios with a mix of MSOM and VIOM outliers and comprises a step that, under reasonable assumptions, can separate the corresponding

observations.

Specifically, we rely on the *forward search* (FS) (Atkinson and Riani, 2000) for outlier detection, and use REMLE to perform down-weighting. The FS is an adaptive hard-trimming method based on an iterative algorithm. It starts from a clean subset of observations identified with a (possibly inefficient) high-BdP estimator. At each iteration, it uses OLS to fit the regression (which is fully efficient on the current subset) and extends the subset recovering the observation that is closest to the fit in terms of (robust) residuals. Notably, assuming that the fraction of contamination is lower than the BdP of the initial estimator, the FS algorithm provides consistent estimates (Cerioli et al., 2014) and an "optimal" ranking of the observations in terms of their distance from the model (Johansen et al., 2016). Outliers are generally recovered only in the last iterations, and we exploit this fact to detect MSOM and VIOM outliers. Moreover, the FS allows one to monitor the influence exerted on the estimation of regression parameters by the observations retrieved in each iteration – starting from the high-BdP estimator and ending with the OLS computed on the entire sample. This can be used to adaptively settle on a final clean subset of observations, which in many practical scenarios guarantees a better trade-off between BdP and efficiency (Riani et al., 2014).

The reminder of the article is organized as follows. Section 2.2 introduces the classical regression model and its contaminated counterpart, provides some technical background, and details our proposal. Section 2.3 presents a simulation study where various techniques are compared, under different scenarios, in terms of estimation accuracy and computing time. Our graphical diagnostic tools are illustrated focusing on one of the more complex simulation scenarios. Section 2.4 applies our approach to real-world data, in both its automated and its "monitored" versions. Section 2.5 provides final remarks and pointers for future extensions.

2.2 Our proposal and some background

2.2.1 A generalized setting

Consider the classical linear regression model of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is a vector of observable responses, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a full rank design matrix with $n > p$ containing observable predictors (these are customarily considered as given, even when they comprise randomness), $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter vector, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of unobservable random errors. Classical assumptions specify that such errors are uncorrelated, homoscedastic and Gaussian; $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$, where $\sigma^2 > 0$ and $\boldsymbol{I}_n$ is the identity matrix of size n. Under these assumptions the MLE for $\boldsymbol{\beta}$ corresponds to the OLS, which is the *uniformly minimum variance unbiased estimator* (UMVUE). The MLE for σ^2 is biased by the factor $n/(n-p)$, while REMLE provides a UMVUE also for σ^2.

The absence of any (systematic or stochastic) deviation from (2.1) is an implicit assumption. We relax it through a parametric outlier model affecting both means and variances. In particular, we allow the presence of two distinct groups of outliers: m_V observations generated from a VIOM, and m_M observations generated from a MSOM. We index the two groups as I_V and I_M, respectively, but we remark that the outliers' *labels*, i.e. which indexes belong to these two sets, as well as their cardinalities, are unknown. In symbols, we have

$$\varepsilon_i \sim \begin{cases} N(0, \sigma^2 v_i) & \forall\ i \in I_V \\ N(\mu_i, \sigma^2) & \forall\ i \in I_M \\ N(0, \sigma^2) & \text{otherwise,} \end{cases} \tag{2.2}$$

where $v_i > 1$ for $i \in I_V$ and $\mu_i \neq 0$ for $i \in I_M$. An equivalent parameterization of

the contaminated model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}_{I_V}\boldsymbol{\gamma} + \boldsymbol{D}_{I_M}\boldsymbol{\phi} + \boldsymbol{\epsilon},$$

where $\boldsymbol{D}_{I_V}$ $(n \times m_V)$ and $\boldsymbol{D}_{I_M}$ $(n \times m_M)$ are matrices composed by dummy column vectors indexing the outliers belonging to the two groups, $\boldsymbol{\gamma} \in \mathbb{R}^{m_V}$ is a random vector $\sim N(\boldsymbol{0}, \sigma^2 \text{Diag}_{m_V}(v_i - 1))$ ($\text{Diag}_k(\cdot)$ stands for a $k \times k$ diagonal matrix), $\boldsymbol{\phi} \in \mathbb{R}^{m_M}$ is a non-stochastic vector, and the random error vector is again $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. This parameterization highlights that MSOM and VIOM outliers can be thought of, respectively, as fixed and random effects in a *mixed linear model*. As noted by Cook et al. (1982), one could envision outliers compounding a mean shift and a variance inflation – but this leads to an over-parametrization in which these compounded outliers are equivalent to MSOMs.

The fact that we focus on an *unlabeled* problem, where not only the identity but also the number and nature (MSOM vs. VIOM) of multiple outliers is unknown, complicates matters because it makes masking and swamping effects more likely. As customary (especially in the robust statistics literature), we assume that MSOM outliers can also be affected by shifts in the predictors – which contaminate entries of the design matrix, affecting leverage (Maronna et al., 2006) – but that the VIOM outliers are not (Cook et al., 1982). Correspondingly, when generating predictors in our simulation experiments, we introduce mean shifts in their distribution; we thus use μ_X to indicate predictors shifts, and μ_ε to indicate errors shifts. We also restrict ourselves to settings in which n is substantially larger than p, and rely on two key additional assumptions; namely, that:

A1 The total fraction of contaminated observations (MSOM or VIOM) is smaller than 50%.

A2 Systematic contaminations, which induce shifts in means (MSOM) have larger influence on the regression compared to stochastic contaminations, which inflate variances (VIOM). Thus, under the uncontaminated model, MSOM outliers are expected to have larger residuals than VIOM outliers.

(A1) allows us to safely rely on the properties of high-BdP equivariant estimators,

and is fairly standard (Maronna et al., 2006). (A2) allows us to take advantage of the FS algorithm to discriminate between the two types of outliers, and may not be appropriate in all applications. However, it reflects an intuitive logic in differentiating the two types of contaminations; e.g., shifts in means may be due to the inclusion in the sample of units that do not belong to the target population, while inflation in variances may be due to inaccuracies in measurements on units that do. Also intuitively, from the perspective of the remedies taken, a shift in mean, resulting in the deletion of an observation, cannot be less consequential than an inflation in variance, resulting in a down-weighting.

2.2.2 Some technical background

Our proposal utilizes the FS (Atkinson and Riani, 2000), an adaptive hard-trimming method based on an iterative algorithm. The FS algorithm starts from a "clean" subset of observations of size, say, b_0. This is identified with a high-BdP estimator, often setting $b_0 = p$ in order to reduce to probability of including outliers. However, unlike in the case of an MM-estimator (see below), the robustness of the FS does not depend on the choice of high-BdP estimators (as long as it unmasks outliers) but on its inclusion strategy. Indeed, computationally fast high-BdP estimators are generally used in the FS, e.g., *least median of squares* (LMS) or *least trimmed squares* (LTS); see Rousseeuw and Leroy (1987); Rousseeuw and Van Driessen (2006).

At each iteration b, with $b_0 \leq b \leq n$, the FS operates on a current subset of observations $S(b)$ of size b. The OLS estimate $\widehat{\boldsymbol{\beta}}(b)$ is computed on observations $i \in S(b)$, and residuals are produced for *all* observations $i = 1, \ldots, n$:

$$e_i(b) = y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}(b). \tag{2.3}$$

In the subsequent iteration of the FS, $S(b+1)$ will comprise the $b+1$ observations with smallest absolute residuals in (2.3). Importantly, the FS strategy for recovering and sometimes removing observations from the current subset (removals can happen especially as outliers are included in the fit late in the process) provides a natural ordering of all observations at each iteration – because the OLS is fully efficient

under the uncontaminated null model. Once all n observations have been included in the process, the FS reaches the full OLS fit. Indeed, the FS comprises a collection of least squares estimators carrying information on a sequence of model fits – from a very robust one, to the classical OLS.

The next objective is to establish a satisfactory compromise between BdP and efficiency along this sequence, pinpointing an iteration where the inclusion of outliers "breaks down" the OLS (Riani and Atkinson, 2007). For a generic iteration, consider the deletion residuals of the $n - b$ observations $i \notin S(b)$, defined as

$$r_i(b) = \frac{y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}(b)}{\sqrt{s^2(b)\{1 + h_i(b)\}}} = \frac{e_i(b)}{\sqrt{s^2(b)\{1 + h_i(b)\}}}, \tag{2.4}$$

where $s^2(b)$ estimates σ^2 on $b - p$ degrees of freedom, $h_i(b) = \boldsymbol{x}_i^T [\boldsymbol{X}(b)^T \boldsymbol{X}(b)]^{-1} \boldsymbol{x}_i$, and $\boldsymbol{X}(b)$ indicates the design matrix restricted to the rows $i \in S(b)$. Let $i_{\min} = \arg\min_{i \notin S(b)} |r_i(b)|$ be the index of the observation that is closest to $S(b)$ in terms of deletion residuals. The idea is that if the absolute value of $r_{i_{\min}}(b)$ is sufficiently large, $i_{\min}$ (and *a fortiori* all other observations $\notin S(b)$) are outliers. The deletion residuals in (2.4) follow a Student's t distribution under the uncontaminated null model if the estimates are based on all $n - 1$ observations (Cook and Weisberg, 1982, 20). But this fact is not directly applicable for assessing the inclusion of some outliers in an FS iteration, because here they depend on order statistics. However, the assessment can be performed, e.g., by bootstrapping. In particular, we utilize an approach proposed by Riani et al. (2009) for multivariate analysis and adapted by Atkinson et al. (2016) to regression problems. It relies on theoretical results from (symmetrically) truncated distributions and order statistics to provide fast and accurate point-wise bounds that approximate bootstrap envelopes. Multiple testing is handled controlling the sample-wise level at around 1%.

When applied to regression settings, the FS coupled with this "automated" strategy to identify a good trade-off between BdP and efficiency is referred to as *FSR* (Riani et al., 2012); see Figure 2.5 for an example. In more detail, starting from any iteration of the FS, FSR implements a two-stage procedure based on consecutive single outlier testing of the values $r_{i_{\min}}(b)$, which adaptively trims outliers.

A first stage detects outliers using all n observations, testing consecutive triplets, couples or single extreme values. If a signal is detected at a a given iteration, say b_1, all observations not belonging to $S(b_1)$ are flagged as possible outliers. A second stage attempts to validate this signal and, if it does, trims a subset of observations flagged in the first stage. This is performed through a superimposition of (forward) confidence bands, starting from the signal potentially detected at iteration b_1 in the first stage, and until a second signal is potentially detected at iteration b_2. All observations not belonging to the subset $S(b_2)$ are, again, trimmed.

We remark that, in principle, we could develop our proposal using soft-trimming approaches such as the MM-estimator (Yohai, 1987) in place of the FS. However, while these estimators have appealing theoretical and empirical properties (Riani et al., 2014; Maronna et al., 2006), they also have substantial drawbacks with respect to our purposes. In particular, they (i) generally rely on a computationally expensive preliminary high-BdP estimator (e.g., a soft estimator of scale; Riani et al. 2014); (ii) down-weight all observations, possibly trimming the most extreme ones, and thus do not separate a subset of "clean" observations from the rest; (iii) estimate weights through a loss function (e.g., the Tukey bisquare), without explicitly relying on a variance inflation model; (iv) require nontrivial choices (e.g., the preliminary estimator and the loss function); (v) comprise tuning parameters which are often pre-specified (e.g., the efficiency level of the MM-estimator); and (vi) can complicate statistical inferences (Cerioli et al., 2016).

2.2.3 Our proposal

Our proposal estimates model parameters, identifies outliers arising from either a VIOM or a MSOM, separates them, and estimates the weights with which they participate to the regression (these are forced to 0 for MSOM outliers).

In using a hard-trimming approach, we rely on the fact that the VIOM can be viewed as a generalization of the MSOM. An asymptotic equivalence can be drawn (Cook et al., 1982), an increasing variance inflation pushes weights to 0. Moreover, based on REMLE, VIOM provides a ranking of outliers equivalent to that of MSOM (Thompson, 1985). Consequently, to assess the presence of either

VIOM or MSOM outliers, we can simply compare externally Studentized residuals – or any monotonic function of them, e.g., Studentized residuals (Cook and Weisberg, 1982). Of course such residuals must be computed from a robust fit in order to avoid masking problems. In this setting the FS ranking is meaningful: VIOM outliers are recovered by the iterations right after the clean units and before MSOM outliers.

In addition to utilizing the FS ranking, our proposal also combines trimming and REMLE weighting. In a way, we create a straightforward generalization of the procedure proposed by Thompson (1985) in the presence of a single VIOM outlier; namely: (i) find the largest squared Studentized residual, (ii) estimate v_i and σ^2 with REMLE, and (iii) estimate $\boldsymbol{\beta}$ using weighted least squares. Relying on assumption A2, which postulates that MSOM outliers are more extreme than VIOM outliers, we adapt this procedure as follows:

- We run FSR with standard settings, and take its first detected signal as our "weak" signal, pinpointing the iteration where VIOM outliers start to be included in the fit.
- We increase the standard quantile thresholds[1], and take the second signal detected by FSR as our "strong" signal, pinpointing the iteration where MSOM outliers start being included in the fit.
- We label the group of observations recovered by the FS iterations between the two signals as VIOM outliers, and those excluded from the FS at the second signal as MSOM outliers.
- We use REMLE to estimate the weights of the observations labeled as VIOMs, and trim out the observations labeled as MSOMs (i.e. set their weights to 0).

This procedure, which we refer to as *FSRw*, adaptively identifies both VIOMs and MSOMs, generating a data-driven estimate of the fraction(s) of outliers and without fixing *a priori* a trade-off between BdP and efficiency. Furthermore, given that FSR relies only on consecutive exceedances of single-unit outlying tests, it tackles multiple outliers without resorting to more complex calculations.

[1]Since FSR only aims to trim observations, the default settings can be too weak to separate coexisting VIOM and MSOM outliers – as we wish to do here.

In the current implementation, we use REMLE to estimate (sub-optimal) *single* weights; thus, from now onward we will refer to our procedure more specifically as *FSRws*. However, we point out that in principle one can use REMLE to estimate multiple weights jointly, based on recent proposals in Gumedze (2019). In this article we do illustrate the excellent performance and low computational burden of FSRws (see Section 2.3), but we do *not* compare it to its FSRwj "joint" extension. This is of course of interest, as it may lead to substantially better performance and, with appropriate computational implementations, to affordable increases in running time – but it is left for future work. Evidence from preliminary comparisons on very small simulated data sets (not shown) suggests that, at least at low contamination levels, FSRwj does not produce marked performance gains.

2.2.4 Graphical diagnostics

If one fixes the thresholds used to pinpoint "weak" and "strong" signals along the FS, FSRws is computationally very efficient and *fully automated*. Full automation is particularly useful when multiple outlier detection/treatment and estimation of model parameters must be accomplished rapidly and without human intervention (e.g., fraud detection in international trade data as described in Perrotta and Torti 2010). When full automation is not necessary, graphical diagnostic tools can aid decisions by allowing a user to monitor the FS process, especially when combined with interactive graphical tools (e.g., using brushing and linking techniques as in Riani et al. 2012). Indeed, the FS algorithm embeds information about the influence of every point, at each iteration, on any parameter (or test statistic) of interest. We propose to profile (single) REMLE weights for all the observations not included in the FS at each iteration, creating what we call a *cascade plot* (see Figure 2.6 for an example).

The rationale for the cascade plot diagnostic is the following. As iterations proceed, and one moves along the plot horizontally, estimated weights at the time of inclusion in the FS should be: (i) approximately constant for uncontaminated observations (except for some short and mild dip due to the FS inclusion rule, especially in the last iterations); (ii) markedly decreasing when VIOM outliers start

to be included; and (iii) sharply increasing when MSOM outliers start to be included (due to masking effects).

Unlike diagnostics that inform us on the quality or pitfalls of a specific regression fit (Belsley et al., 2004; Cook and Weisberg, 1982; Chatterjee and Hadi, 1988; Atkinson, 1985), this and other types of *monitoring diagnostics* provide information about a sequence of fits. This often reveals the empirical properties of a robust estimator and provides useful insights about the structure of the data. Indeed, the "phylosophy" of monitoring, which is very natural in the spirit of the FS, can be generalized to other classes of robust estimation procedures, e.g., creating graphical diagnostic plots that profile residuals (or their correlations) along a sequence of BdP or efficiency values – moving from one extreme to the other (Riani et al., 2014; Cerioli et al., 2016, 2018). For instance, when we compare procedures in Section 2.3, we also implement an *MM-weights plot* (see left panel of Figure 2.7) which is similar in spirit to a cascade plot.

MM-estimators are often used with a pre-specified efficiency level (e.g., 0.85, 0.95 or 0.99) – relying on asymptotic results that hold only for data where contaminated and uncontaminated observations are well-separated, and orthogonal predictors (Maronna et al., 2006, 141). In contrast, an MM-weights plot allows us to monitor estimated weights as a function of efficiency, again keeping track of a sequence of fits – from very high BdP to very high efficiency. Of course, the choice of a preliminary high-BdP estimator affects the solution. One of its clear effects is that of shifting the efficiency level required to (possibly) breakdown the MM-estimator. Monitoring weights derivatives is also informative, and can be logically related to the infinitesimal approach to robustness (Hampel et al., 2011) and to local influence (Cook, 1986). We implement this in a *MM-weights derivatives plot* (see right panel of Figure 2.7).

2.3 Simulation study

Here we present the general simulation framework we created to evaluate our proposal, and then focus on selected simulation results illustrating accuracy and com-

putational burden. Reproducible and documented MATLAB code (based on the FSDA Toolbox[2]; Riani et al. 2012) is available at https://github.com/LucaIns/VIOM_MSOM.

First, we generate data following the uncontaminated model (2.1). The $n \times p$ design matrix $\boldsymbol{X}$ comprises all 1's in the first column (for the model intercept) and the remaining entries of each row are drawn independently from a standard $(p-1)$-variate Normal. The p-dimensional coefficient vector $\boldsymbol{\beta}$ is fixed; note that the size of the coefficients is irrelevant as long as we consider regression and affine equivariant estimators (Maronna et al., 2006, 142). The errors are drawn independently from a $N(0, \sigma^2_{\text{SNR}})$, where σ^2_{SNR} is chosen as function of the signal-to-noise ratio with which we want to characterize an experiment; $\text{SNR} = \text{var}(\boldsymbol{X}\boldsymbol{\beta})/\sigma^2_{\text{SNR}}$.

Next, following (2.2), we independently contaminate (uniformly at random, without repetitions) m_M observations with a MSOM (here mean shifts are introduced also in the predictors) and m_V with a VIOM (here predictors are uncontaminated).

To create an *oracle benchmark*, we run a *weighted least squares* (WLS) fit using the true weights (i.e. 0 or $w_i = v_i^{-1}$ for observations contaminated with a MSOM or a VIOM, respectively, and 1 for uncontaminated observations). In the figures reported in this section, this optimal benchmark is indicated with "opt". We compare it with the following procedures:

- The *ordinary least squares* (OLS).
- The *least median of squares* (LMS), a hard-trimming estimator with asymptotic BdP of 50% (the highest achievable for equivariant estimators) (Rousseeuw, 1984).
- The *MM-estimator* (MM), using LMS as preliminary estimator and Tukey's bisquare loss function, with tuning constant fixed as to achieve 85% nominal efficiency (Maronna et al., 2006). Note that using a preliminary hard-trimming estimator such as LMS is sub-optimal in terms of efficiency in MM, but very convenient in terms of reducing computational burden.

[2]The FSDA Toolbox is freely downloadable at http://rosa.unipr.it/fsda.html

- *Forward search regression* (FSR), the adaptive trimming procedure described in Section 2.2.2, with the initial clean subset created again with LMS (based on our assumption A1, we force FSR to search for a signal only after having included 50% of the sample).

- Our *FSRws*, which utilizes a variant of FSR and single REMLE weights as described in Section 2.2.3.

Performance of the procedures is compared across different sample sizes n and fractions of (total) contamination $\frac{m_M+m_V}{n}$. We generate an equal number of VIOM and MSOM outliers ($m_V = m_M$) without overlaps between the two groups. Each simulation scenario is replicated t times and results are averaged.

In terms of performance metrics, we consider the mean squared error (MSE) of $\widehat{\beta}$ (for $p = 1$) partitioned in variance and squared bias:

$$\text{MSE}\left(\widehat{\beta}\right) = \frac{1}{t}\sum_{i=1}^{t}\left(\widehat{\beta}_i - \beta\right)^2 = \frac{1}{t}\sum_{i=1}^{t}\left(\widehat{\beta}_i - \overline{\beta}\right)^2 + \left(\overline{\beta} - \beta\right)^2, \tag{2.5}$$

where $\overline{\beta} = \sum_{i=1}^{t}\widehat{\beta}_i/t$. When $p > 1$, we average the MSE across the coordinates of $\boldsymbol{\beta}$ and take $\text{MSE}\left(\widehat{\boldsymbol{\beta}}\right) = \sum_{j=1}^{p}\text{MSE}\left(\widehat{\beta}_j\right)/p$.

We also consider the MSE of proxy estimates of weights ($\widehat{w}_i = \widehat{v}_i^{-1}$, $i = 1, \ldots, n$) and error variance ($\widehat{s}^2$). Note that comparing weights estimates poses some issues because outlier labeling varies from replication to replication, VIOM outliers may sometimes not carry sizeable residuals, and experiments with larger sample sizes may contain more outlier-like uncontaminated observations by chance (these are especially hard to distinguish from VIOM outliers). For the variance we take a WLS-like proxy estimate of the form

$$\widehat{s}^2 = \frac{1}{(n-p)}\frac{\sum_{i=1}^{n}\widehat{w}_i e_i^2}{\sum_{i=1}^{n}\widehat{w}_i/n}, \tag{2.6}$$

where the e_i's are estimation residuals[3]. This captures the effectiveness of weights estimates, taking into account the outlying-ness of observations regardless of whether

[3]Note here we are not using any consistency factor in $\widehat{s}^2$. Consistency factors are often used in robust estimation (Maronna et al., 2006).

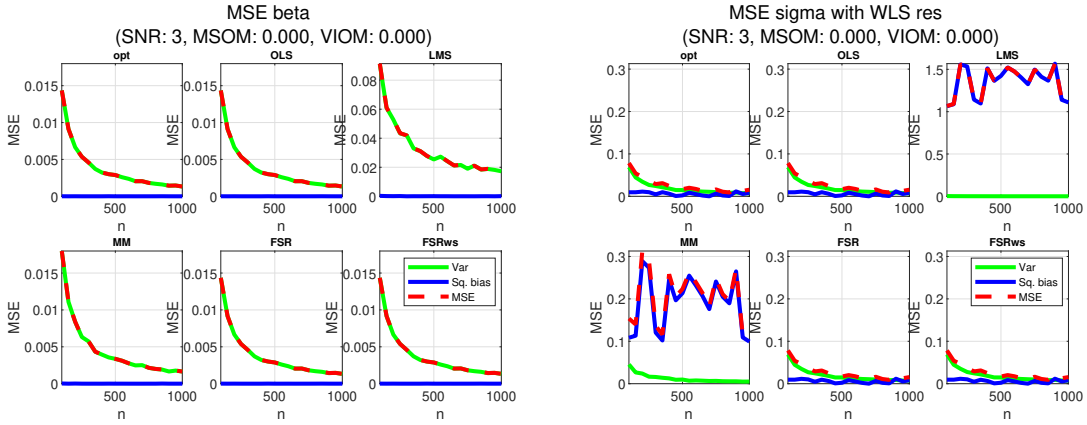


Figure 2.1: MSE comparisons, across procedures and sample sizes, for $\widehat{\boldsymbol{\beta}}$ (left panel) and $\widehat{s}^2$ (right panel) in the absence of contamination.

they are in fact contaminated. The MSE decomposition for $\widehat{s}^2$ is computed as in (2.5), where σ^2_{SNR} and $\widehat{s}^2$ replace β and $\widehat{\beta}$, respectively. Finally, and importantly, we compare procedures in terms of average computing time (in seconds).

In the following, for simplicity, we focus on results for a simulation scenario where the uncontaminated model contains an intercept and a single predictor ($p = 2$), setting $\boldsymbol{\beta} = (2, 2)^T$. The signal-to-noise ratio is set to SNR $= 3$. VIOM outliers are all generated with variance inflation parameter $v = 10$. MSOM outliers are all generated with error mean shift $\mu_\varepsilon = -3$ and predictor mean shift $\mu_X = 3$. We use shifting parameters with opposite signs in order to create bad leverage points which are more likely to disrupt the true positive slope relating response and predictor. We consider increasing sample sizes n ranging from 100 to 1000 (with a step size of 50), and total contamination fractions $(m_V + m_M)/n$ of 0, 0.25 and 0.5. Data for each setting are generated $t = 500$ times, and results are averaged over these replications.

Figure 2.1 shows results for the MSEs of $\widehat{\boldsymbol{\beta}}$ (left panel) and $\widehat{s}^2$ (right panel) across procedures and sample sizes, when there is no contamination; $(m_V + m_M)/n = 0$. Both FSR and FSRws do not detect any signal and lead to optimal OLS estimates. On the other hand, MM and LMS must sacrifice some efficiency under the null uncontaminated model. In particular, the MM $\widehat{\boldsymbol{\beta}}$ estimates are close to optimal, but its $\widehat{s}^2$ estimates are biased. LMS has a lower convergence rate for $\widehat{\boldsymbol{\beta}}$ and even larger biases for $\widehat{s}^2$.

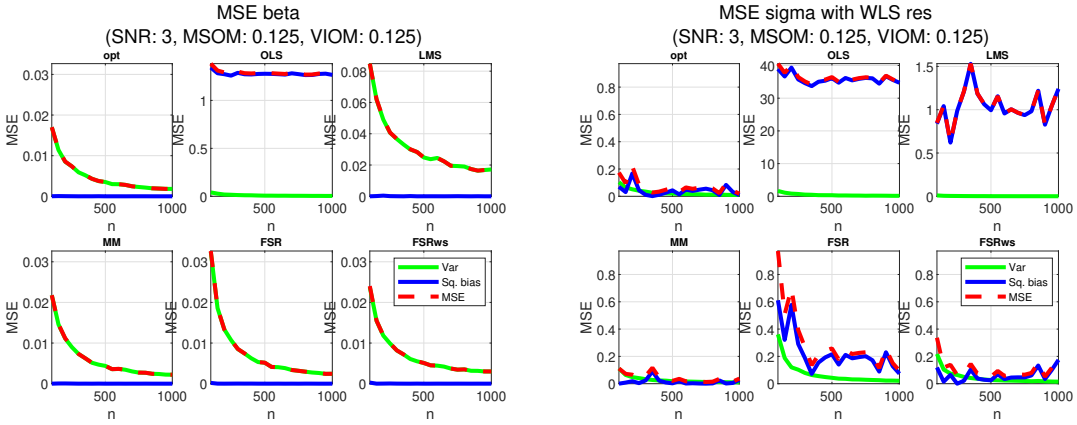


Figure 2.2: MSE comparisons, across procedures and sample sizes, for $\widehat{\boldsymbol{\beta}}$ (left panel) and $\widehat{s}^2$ (right panel) in the presence of an intermediate level of contamination.

Figure 2.2 has the same format of Figure 2.1, but here the total fraction of contamination is set to $(m_V + m_M)/n = 0.25$ $(m_V/n = m_M/n = 0.125)$. The MM outperforms other procedures – and, notably, it often outperforms also the oracle benchmark in terms of $\widehat{s}^2$. This may be due to the fact that some VIOM outliers do not need to be down-weighted because they lie along the bulk of the data. However, FSR and FSRws perform almost on par with MM, and markedly better than LMS (and of course OLS) in terms of $\widehat{\boldsymbol{\beta}}$. LMS performance is in fact similar to the case with no contamination, due to its high BdP. The OLS breaks down due the presence of MSOM outliers, which induce strong biases (and very low variances) in its estimates. Note that, for our FSRws, the MSE of $\widehat{s}^2$ shows a slight increase for large n values. This is due to the fact that FSRws tends to detect more outliers as the sample size increases, including "false positives" (especially "false positive" VIOMs). Consequently, FSRws may down-weight (or even trim out) more observations than needed when the sample size is very large[4]. Nevertheless, the $\widehat{s}^2$ produced by FSRws shows smaller bias than that produced by FSR across sample sizes.

Figure 2.3 has again the same format, but here the total fraction of contamination

[4]This is a feature "inherited" from our use of FSR. As the sample size increases, FSR becomes better (likely due to its strong consistency) at detecting and thus trimming all outliers (both VIOM and MSOM). Hence, when n is large, the first signal (which is the same both for FSR and FSRws) can occur while still including clean observations. In the strong contamination scenario (50% total fraction of contamination) which we consider next we do not notice such phenomenon because we are forcing FSR to find signals in the second half of the search (i.e. we force 50% of the weights to be = 1, motivated by our assumption A1).

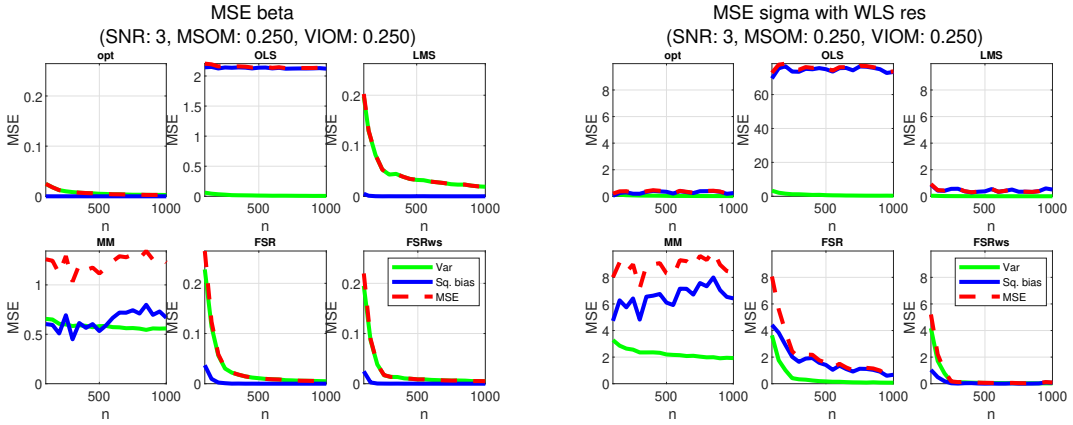


Figure 2.3: MSE comparisons, across procedures and sample sizes, for $\widehat{\boldsymbol{\beta}}$ (left panel) and $\widehat{s}^2$ (right panel) in the presence of a high level of contamination.

is increased to $(m_V + m_M)/n = 0.50$ ($m_V/n = m_M/n = 0.25$). FSRws, FSR and LMS perform comparably well in terms of $\widehat{\boldsymbol{\beta}}$. But at this high contamination level we see a breakdown of the MM, not just the OLS. The situation is similar for $\widehat{s}^2$; LMS provides (nearly) optimal estimates, FSRws does well (better than LMS as n increases) and improves upon FSR (especially in terms of bias), MM and OLS do poorly. These results highlight how the need to use a pre-specified efficiency can seriously hinder the MM-estimator; having fixed efficiency at 85%, we observe a breakdown in the MM only when raising the contamination to a total fraction as high as 50%. However, when efficiency is set at higher levels (95% or 99% are often used), MM performance can seriously deteriorate also with milder contaminations.

Figure 2.4 (left panel) shows average computing times across procedures and sample sizes, in the high contamination setting. All the procedures we compare here run reasonably fast on our simple and (relatively) small simulated data. For all, on average, the running time is less than one second for $n = 1000$ (using MATLAB R2018a on an Intel Core i7-7700HQ CPU at 2.8 GHz $\times$ 4 processors and 16 GB RAM). In general, the computational cost of an MM-estimator is nearly all due to obtaining a preliminary high-BdP estimator. This can in fact be very expensive with standard choices (e.g., soft-scale estimators). However, in MM we use the LMS which is inexpensive. After the LMS is computed, the M-estimation phase of MM takes a negligible amount of time (as running iteratively reweighted LS). Running FSR from a clean subset produced by LMS is more expensive but, importantly,

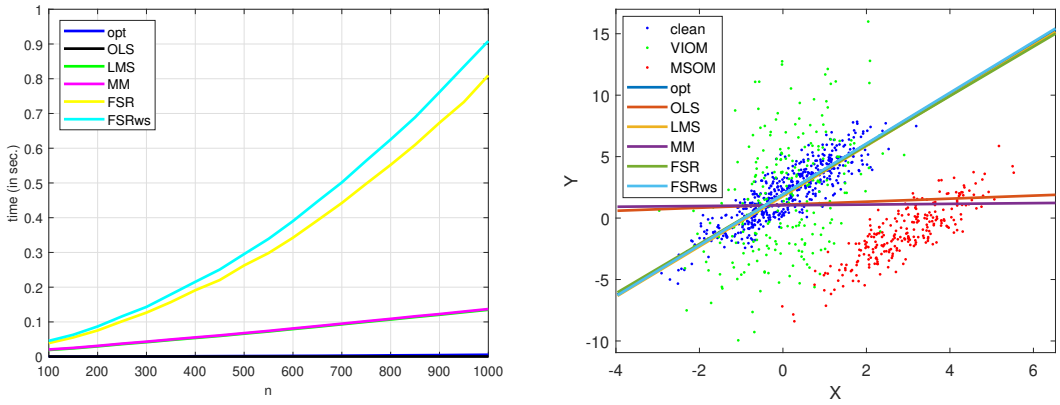


Figure 2.4: Left panel: average computing time comparisons, across procedures and sample sizes, in the presence of a high level of contamination. Right panel: scatterplot of a simulation example comparing different fits in the presence of a high level of contamination. Here $n = 1000$.

it produces information about a sequence of fits. Running our FSRws adds yet some to the computational burden, but not a lot. In our current implementation, FSRws runs fairly inexpensively based on the FSR solution – and with further code optimization, the added cost on top of that of FSR should be nearly equivalent to that of running a WLS.

Importantly, all procedures considered here, including the FSRws (which effectively tackles multiple MSOM and VIOM outliers), are hugely cheaper than any approach for outlier treatment that relies on combinatorial enumeration – especially in the case of VIOM outliers.

Next, we illustrate graphical diagnostics using a simulation with high contamination and $n = 1000$. Figure 2.4 (right panel) compares different fit on this dataset – red and green points represent MSOM and VIOM outliers, respectively. FSR, FSRws and LMS here provide fits much closer to the oracle than MM, which breaks down because efficiency is set at 85%, and of course OLS.

Figure 2.5 shows the corresponding *residuals forward plot* (left panel) and *absolute minimum deletion residuals forward plot* (right panel) which are commonly used as graphical diagnostics for the FS (Atkinson and Riani, 2000; Atkinson et al., 2016). The residuals forward plot tracks residuals trajectories along the FS iterations. On our simulated dataset, it clearly indicates that MSOM residuals start being included around iteration 720, where large residuals (in dark blue) shrink to zero

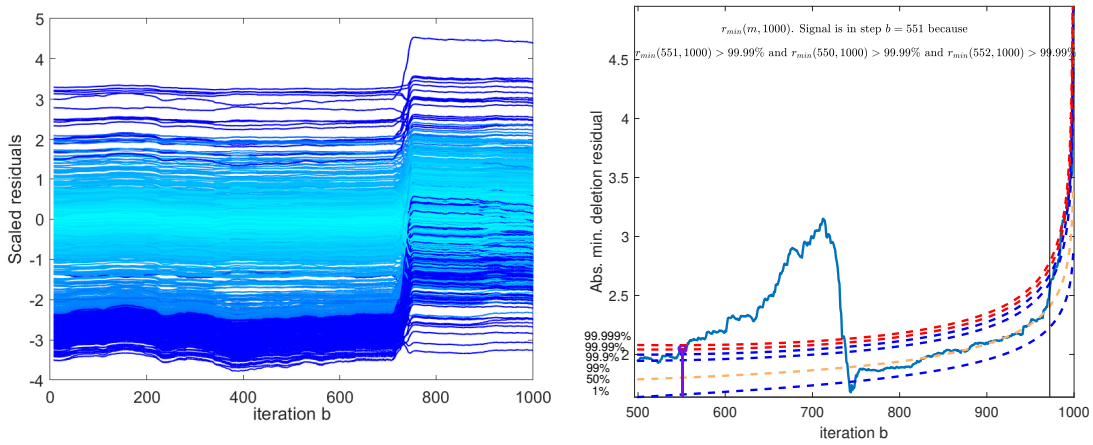


Figure 2.5: Residuals forward plot (left panel) and minimum absolute deletion residuals forward plot (right panel) for the simulated dataset in the right panel of Figure 2.4.

and start masking each other. However, as in this example, a residuals forward plot might become too complex to diagnose the presence of VIOM residuals in large samples. The absolute minimum deletion residuals forward plot tracks a single statistics which depends on the observations excluded from the FS at each iteration, providing a meaningful, simple summary of the information contained in the residuals forward plot. Dashed lines represent different quantiles for the point-wise distribution of the absolute minimum deletion residuals (as described in Section 2.2.2). On our simulated dataset, absolute minimum deletion residuals rapidly increase after iteration ≈ 550 and abruptly fall at iteration ≈ 720. The purple vertical line marks the first "weak" signal identified by FSR at iteration 551; in our FSRws this is when VIOM outliers start being included.

In contrast to these diagnostics our *cascade plot*, which is shown in Figure 2.6, highlights both local information (the influence of each observation at every iteration) and global information (the overall estimator performance). The strong decrease in estimated weights between iterations ≈ 550 and ≈ 720 indicates the inclusion of VIOM outliers, and their abrupt increase after iteration ≈ 720 indicates the inclusion of MSOM outliers. Notably, after iteration ≈ 720 estimated weights increase due to masking effects, followed by a large number of observations' interchanges in the FS subset due to swamping effects. The swamped units (represented by dark blue trajectories in the final part of the cascade plot) were included in earlier

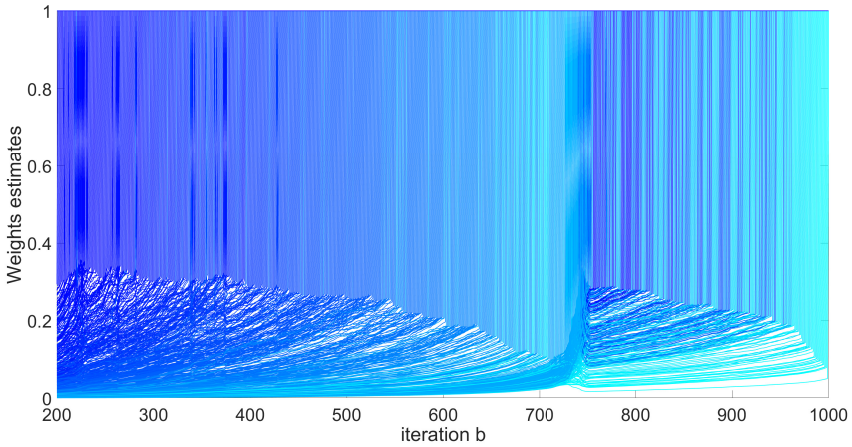


Figure 2.6: Cascade plot for the simulated dataset in the right panel of Figure 2.4.

iterations of the FS and exit the subset as MSOM outliers begin entering it.

Finally, we monitor the performance of the MM using our *MM-weights plot* and *MM-weights derivatives plot*, which are shown in the left and right panel of Figure 2.7, respectively. We track estimated weights along efficiency levels ranging from 0.5 to 0.99. As in the right panel of Figure 2.4, red and green denote MSOM and VIOM outliers. Both plots clearly indicate that the MM breaks down at an efficiency level of ≈ 0.84, where it produces a fit very similar to the OLS. Before this efficiency value, clean units (as well as "non-outlying" VIOM outliers) have large and stable weights, and MSOM outliers have very small weights; these abruptly increase after the threshold. VIOM outliers are in between these two extremes. Intuitively, while trajectories for influential observations that create masking tend to be convex-shaped, the ones for swamping observations are concave-shaped. These effects become even clearer in the MM-weights derivatives plot. Right before the estimator breaks down, one can see a steep increase in derivatives corresponding to outliers that are masking each other (typically MSOMs), and a steep decrease in derivatives corresponding to swamped observations (e.g., good leverage points). In contrast, uncontaminated non-swamped observations have "flat" and small derivatives.

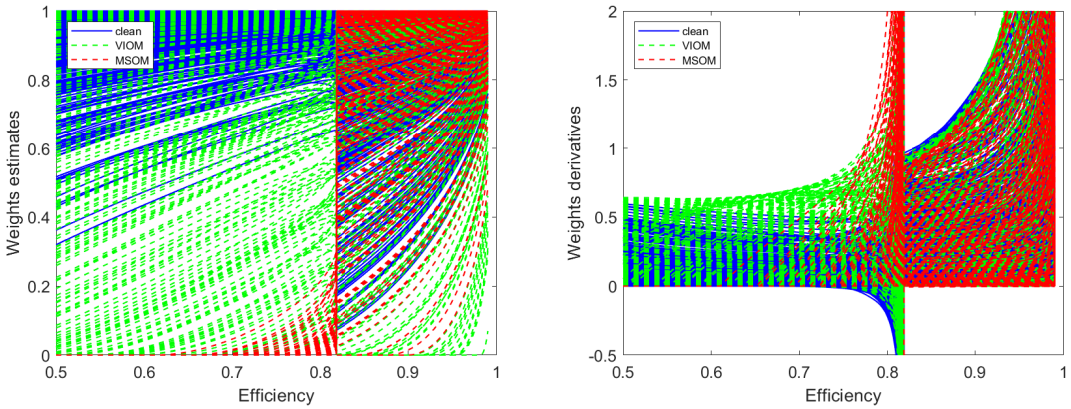


Figure 2.7: MM-weights plot (left panel) and MM-weights derivatives plot (right panel) for the simulated dataset in the right panel of Figure 2.4.

2.4 Real-data examples

We now apply our FSRws and graphical diagnostics to two real-world datasets which pose different levels of challenge, both in terms of sample size and in terms of contamination mechanisms.

The first dataset is very small ($n = 11$), and was used by Cook et al. (1982). It contains measurements of the thickness of non-magnetic coatings of galvanized zinc on iron and steel, obtained with two different procedures; an expensive one (response) and a cheaper one (predictor). Figure 2.8 (top left panel) shows a scatterplot along with OLS, MM and FSRws fits for the regression, which includes an intercept. This simple example motivates the use of robust estimation procedures to deal with multiple outliers arising from a VIOM and/or MSOM. Assuming a single possible VIOM outlier and using MLE, Cook et al. (1982) flagged observation 9, which has the largest absolute residual. Based on REMLE, Thompson (1985) flagged observation 11, which has the largest absolute Studentized residual.

The order of the last observations to enter the FS is: 5, 6, 9 and 7. For this small dataset, the residuals forward plot (Atkinson and Riani, 2000) shown in Figure 2.8 (top right panel) provides very clear information. Observations 9 and 7 have a similar behavior, and so do observations 5 and 6. The inclusion of unit 9 at iteration 10 causes a masking of observation 7 and swamping of observations 11 and 10. These effects become stronger as observation 7 is included in the last iteration, i.e. in the

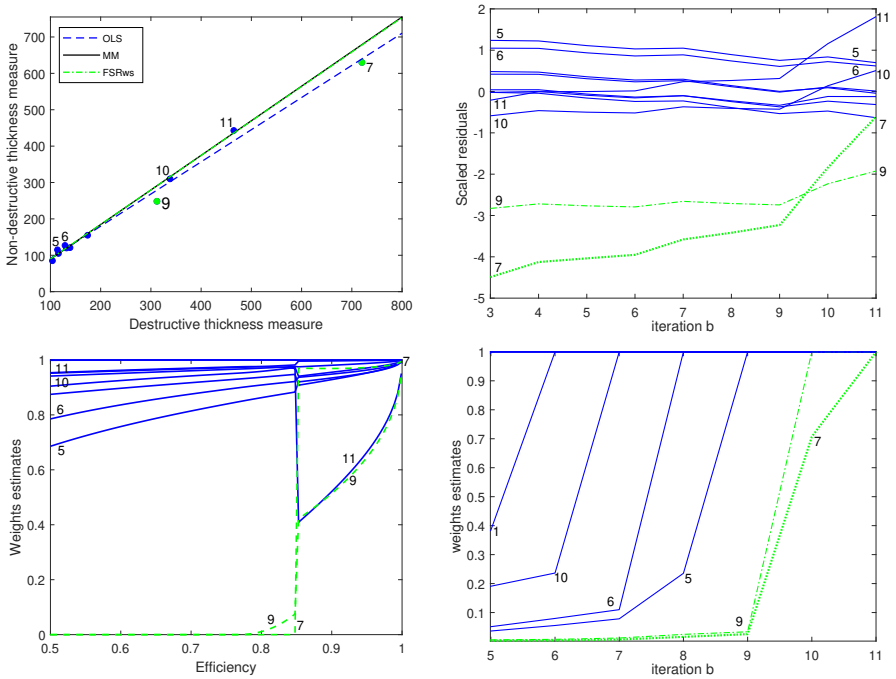


Figure 2.8: Scatterplot of a small dataset ($n = 11$) on coating thickness from Cook et al. (1982) with OLS, MM and FSRws fits superimposed (top left panel). Corresponding graphical diagnostics: residuals forward plot (top right panel), MM-weights plot (bottom left panel) and cascade plot (bottom right panel).

OLS fit. The cascade plot in Figure 2.8 (bottom right panel) tells the same story; with this small sample it does not provide a diagnostic advantage with respect to the residuals forward plot. The weights estimates for the observations not included in the FS subset remain approximately constant between iterations 5 and 8. At iteration 9, observations 7 and 9 have very small weights, but as observation 9 enters the FS at iteration 10, the weight of observation 7 increases abruptly due to masking; both observations are outliers. A similar behavior, though less marked, can be seen for observations 5 and 6 at iterations 7 and 8. Notice also that there are no interchanges of observations in FS; the lines tracking the weights do not cross.

The MM-weights plot in Figure 2.8 (bottom left panel) shows that also the MM-estimator is strongly affected by observations 7 and 9. Indeed, their weights abruptly increase after the 0.85 efficiency level, where the weight for observation 11 abruptly decreases due to swamping. Figure 2.8 (top left panel) shows that the MM fit nearly overlaps with the one for FRSws. But as the efficiency level increases to 86%, MM becomes indistinguishable from the OLS. This demonstrates

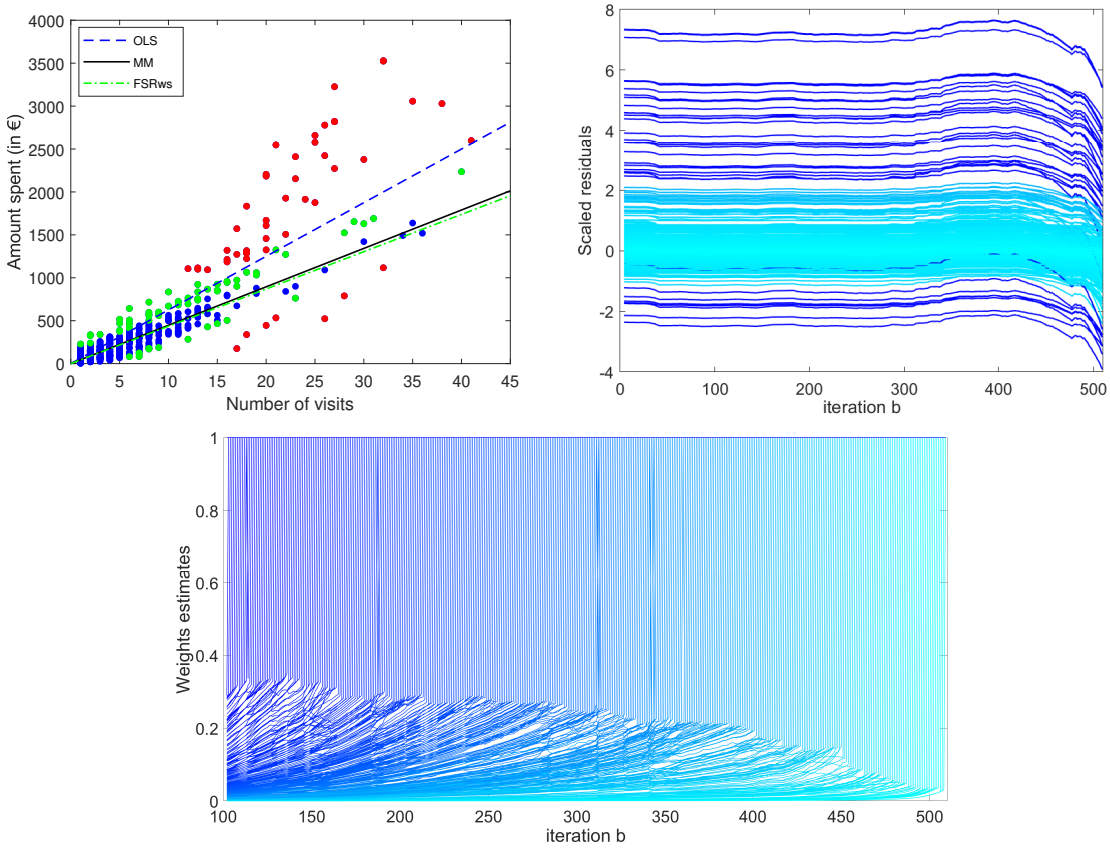


Figure 2.9: Scatterplot of a larger dataset ($n = 509$) on loyalty cards data with OLS, MM and FSRws fits superimposed (top left panel). Corresponding graphical diagnostics: residuals forward plot (top right panel) and cascade plot (bottom panel).

the importance of having a right balance between BdP and efficiency, independently of the choice of a specific preliminary fit for MM-estimators.

Based on the diagnostics discussed above, three strategies could be used for these data: (i) trim observations 7 and 9 and down-weight 5 and 6, or (ii) down-weight observations 5, 6, 7 and 9, or (iii) down-weight only observations 7 and 9. The FSRws fit shown in Figure 2.8 (top left panel) corresponds to (iii)[5]. As we showed before, OLS is influenced by observations 7 and 9 jointly, and for this reason our solution differs from Cook et al. (1982) and Thompson (1985). Indeed, the most outlying unit here is observation 7, which cannot be detected using single outliers methods.

The second dataset we consider is larger, with $n = 509$. It contains loyalty cards

[5]Here, due to the small sample size, FSR does not detect any signal. The FSRws fit shown in the figure is based on "manual" detection; the two down-weighted observations also correspond to the two residuals exceeding 90% confidence intervals in a LMS fit.

information on customers of a supermaket chain in Northern Italy (this data was introduced by Atkinson and Riani 2006 and is available in the FSDA MATLAB Toolbox). The response is the amount spent over a six months period (in Euros) and the predictor is the number of visits to the supermarket in the same period of time. Figure 2.9 (top left panel) shows a scatterplot along with OLS, MM and FSRws fits for the regression, which does *not* include an intercept (the expenditure corresponding to 0 visits is reasonably assumed to be ≈ 0). Here robust fits behave very differently from the OLS, which is affected by multiple outliers. Red and green points represent, respectively observations trimmed (48) and down-weighted (64) by FSRws.

The residuals forward plot in Figure 2.9 (top right panel) shows that some observations have very large residuals during most of the FS – which decrease in absolute terms after iteration 470 due to the inclusion of more extreme outliers. However, this plot is not very informative in terms of diagnosing the joint presence of multiple VIOM and MSOM outliers. The cascade plot in Figure 2.9 (bottom panel) appears to provide more insight: estimated weights decrease markedly after iteration ≈ 400 and the decrease further accelerates after iteration ≈ 450, suggesting the presence of more extreme outliers which are likely to be MSOMs. We also notice that the inclusion of such outliers does not cause any interchanges of observations in the last portion of the FS, indicating that they are not as disruptive as the outliers in our simulation example in Section 2.3 (see the right panel of Figure 2.4 for a comparison).

Figure 2.10 highlights the two classes of outliers detected by FSRws in the MM-weights and weights derivatives plots (left and right panel, respectively). These indicate that the MM-estimator is strongly influenced by outliers for efficiency levels higher than 95%. In particular, the MM-weights plot shows that for most efficiency values trajectories are convex-like for observations labeled as outliers and concave-like for observations labeled as clean (color coding corresponds to FSRws labeling, but by and large this behavior would be visually appreciable even without it). Moreover, the MM-weights derivatives plot shows that units flagged as MSOM have flat derivatives which bump up right before the estimator "breaks down". Units flagged

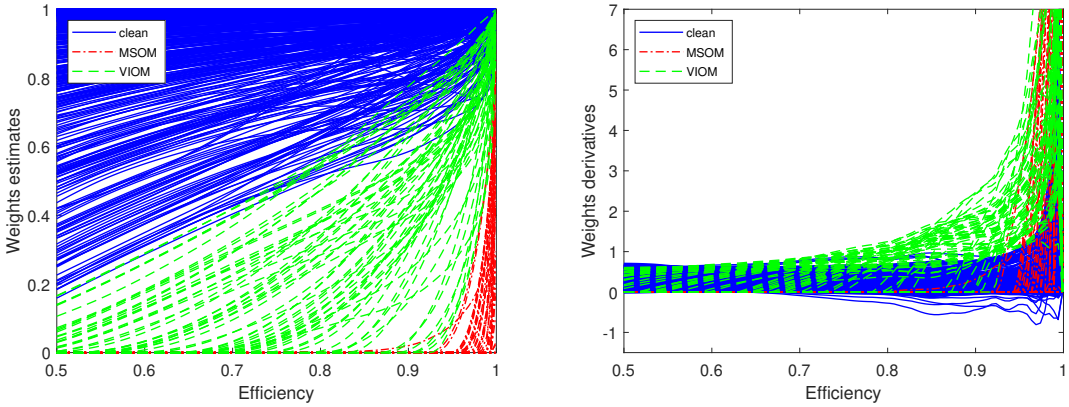


Figure 2.10: MM-weights plot (left panel) and MM-weights derivatives plot (right panel) for Loyalty Cards data in the top left panel of Figure 2.9.

as VIOM have steadily increasing derivatives, and they too accelerate before the breakdown. Non-outlying units have small and constant derivatives for most efficiency values, which eventually become negative for swamped good leverage points.

2.5 Final remarks

Our proposal builds upon different approaches and tools. We use high-BdP and efficient techniques from the robust estimation literature to design a novel procedure that can identify multiple outliers arising from either a MSOM or a VIOM, and provide a way to distinguish between the two. In practice, both soft and hard estimation procedures can deal effectively with VIOM and MSOM outliers. However, soft-trimming procedures can be harder to interpret, because the link between each observation and its influence is blurred by a general down-weighting. Furthermore, choosing the preliminary high-BdP estimator and setting tuning parameters is nontrivial.

Thus, we prefer to focus on hard-trimming procedures, which provide a clear link between each observation and its outlying-ness. In particular, we consider the adaptive hard-trimming approach in FSR and build upon it to construct our FSRws. This provides a meaningful ranking of the observations and a way to detect both a "weak" and a "strong" signal – which we then use to separate VIOM and MSOM

outliers. After this phase, we blend into the mix REMLE techniques from the VIOM literature, which allow us to move from shear trimming to a more general scheme where some observations are trimmed (those identified as MSOMs) and some down-weighted (those identified as VIOMs). This, in a way, "softens" back the trimming.

Quoting Beckman and Cook (1983):

> "*There is a stormy history behind the rejection of outliers. In the past, as is the case today, the lines were fairly well drawn between those who discarded discordant observations, those who gave each observation a different weight, and those who used simple, unweighted averages*".

Combining robust estimation and REMLE techniques these three seemingly separate takes can be effectively joined in a single, principled framework.

In addition to our FSRws, we introduce novel graphical diagnostics. These are monitoring tools that provide information about a sequence of fits – and are similar in spirit to other diagnostics utilized by the FS and FSR. For FSRws we propose the cascade plot, which tracks estimated weights along the FS iterations as they include/exclude observations. The plot aids in the discrimination of VIOM and MSOM outliers, and can complement or replace the automated detection of "weak" and "strong" signals performed as part of FSRws. In a way, it extends existing diagnostic tools depicting both local and global information on the FS process. This can provide critical insights on the structure of the data being analyzed, especially for large sample sizes and in combination combined with interactive tools (e.g., brushing and linking techniques; see Riani et al. 2012). We also propose the MM-weights and MM-weights derivatives plots. These, switching back to soft-trimming and in particular the MM-estimator, allow one to monitor performance as a function of efficiency values, and thus to flag poor decisions leading to the breakdown of the estimator.

Our general approach to the identification of multiple VIOM and MSOM outliers could, in principle, be used with robust estimation procedures other than FSR. This may be particularly useful when the sample size is very small (FSR may not be able to detect signals) or very large (FSR may become computationally demanding). Of course monitoring diagnostics similar, e.g., to the MM-weights and MM-weights

derivatives plots mentioned above, ought to be used also here to achieve a good balance between BdP and efficiency. This is akin to monitoring residuals (or their correlations) for different efficiency levels or BdP as proposed in Riani et al. (2014); Cerioli et al. (2016, 2018), and could lead to a rigorous procedure for data-driven tuning of the MM and other soft-trimming estimators.

Our work is expanding in several other directions. In parallel to investigating theoretical properties, we are analyzing more simulation settings – including high-dimensional scenarios with various degrees of predictors collinearity, different ratios between VIOM and MSOM outliers, observation-specific contamination parameters, and VIOM outliers with contaminated predictors. We also plan to extend the comparisons to additional robust estimation procedures which are computationally more expensive than the ones we considered to date; e.g., preliminary S-estimators for MM-estimators and Tau estimation (Rousseeuw and Yohai, 1984; Yohai and Zamar, 1988). Relatedly, we note that algorithmic advances in *mixed integer optimization*, which have been recently utilized in the feature selection arena (Bertsimas et al., 2016), may offer interesting opportunities also for the computationally viable detection of multiple MSOM outliers. These advances have already been exploited in hard-trimming estimation (e.g., to compute the LMS solution; Bertsimas and Mazumder 2014), and we are investigating their use for performing selection among the dummy features introduced to reparametrize the MSOM. In our context, which comprises both MSOM and VIOM outliers, we can also add a regularization component to "shrink" the latter (She and Owen, 2011; Menjoge and Welsch, 2010).

Looking ahead in a different direction, while for the time being FSRws utilizes single REML weights estimation, an extension is under development to utilize joint REML estimation. As an alternative, the MLE approach proposed in Cook et al. (1982) could be used in place of REMLE. However, this would require a substantial change in the FS algorithm. In fact, since deletion residuals are not necessarily informative, one would need to evaluate a likelihood for each observation excluded in the FS at every iteration.

"The method of Least Squares is seen to be our best course when we have thrown overboard a certain portion of our data – a sort of sacrifice which has often to be made by those who sail upon the stormy sea of Probability"

Francis Ysidro Edgeworth (1887)

Chapter 3

Simultaneous Feature Selection and Outlier Detection with Optimality Guarantees

This chapter is based on: Insolia, L., Kenney, A., Chiaromonte, F., and Felici, G. (2021d). Simultaneous feature selection and outlier detection with optimality guarantees. *Biometrics*, Forthcoming:1–12.

Reproducible and documented code for this chapter is available at: https://github.com/LucaIns/SFSOD_MIP.

3.1 Introduction

High-dimensional regression problems have become ubiquitous in most application domains, and this is especially true in biomedical research where studies are consistently increasing in size and complexity. In these problems the number of features recorded on each observation (or case) is very large – possibly larger than the sample size, and often growing with the sample size itself. The availability of ever larger numbers of potential predictors increases both the chances that some substantial portion of them are irrelevant, and the chances of contamination in the data (i.e., of some cases following a different model). In principle, these risks may be mitigated in very controlled studies targeting specific populations, but these studies often have

smaller sample sizes. In this article, we consider one such study investigating the relationship between childhood obesity and microbiome composition. We use data from Craig et al. (2018) – who studied weight gain in very young children as part of the *intervention nurses start infants growing on healthy trajectories* (INSIGHT) project (Paul et al., 2014). While previous work (Haffajee and Socransky, 2009; Zeigler et al., 2012) focused on the relationship between adult and/or adolescent obesity and microbiome composition, Craig et al. (2018) connected infant weight gain (which is known to be predictive of obesity later in life, Taveras et al. 2009) to microbiota of the child, as well as the mother. As INSIGHT followed children with repeated visits and extensive data collection from birth to around 3 years of age, its sample size was fairly limited (in the hundreds). In such a setting, eliminating redundant features while accounting for potential contamination with estimation approaches that address both *sparsity* and *statistical robustness* is critical.

Two main contamination mechanisms have been traditionally investigated in the literature on low-dimensional linear models: the mean-shift outlier model (MSOM) and the variance inflation outlier model (VIOM; Beckman and Cook 1983; Insolia et al. 2021b). In this work we focus on the MSOM as it is the best developed and most common framework in relatively low dimensions. It operates excluding cases identified as outliers from the fit, and has previously received substantial attention in biomedical research (Alfons et al., 2013; Freue et al., 2019). For high-dimensional settings, the most typical approaches focused on robustifying information criteria or resampling methods (Müller and Welsh, 2005). The last decade has also seen the development of several *robust penalization methods* which rely on a robustification of soft-selection procedures (She and Owen, 2011), adopting a case-wise robust counterpart of maximum likelihood estimation (MLE).

The notion that one can develop methods for *simultaneous feature selection and outlier detection* (SFSOD) stems from the fact that an MSOM can be equivalently parametrized with the inclusion of binary variables, transforming outlier detection into a feature selection problem (Morgenthaler et al., 2004). This is exactly the avenue we pursue in this article. We propose a discrete and provably optimal approach to perform SFSOD based on the use of L_0 constraints – highlighting its connections

with other methods and overcoming the heuristic nature of previous approaches. L_0 constraints have been used separately for feature selection (Bertsimas et al., 2016) and robust estimation (Zioutas et al., 2009) – both of which can be formulated as a *mixed-integer program* (MIP) and solved with optimality guarantees. We combine the two into a novel formulation and take advantage of existing heuristics to produce effective big-$\mathcal{M}$ bounds and warm-starts to reduce the computational burden of MIP.

We provide theoretical guarantees for our approach, including its high breakdown point, necessary and sufficient conditions to achieve a *robustly strong oracle property* – which holds also in the ultra-high dimensional case when the number of features increases exponentially with the sample size – and optimal parameter estimation. In contrast to existing methods, our approach requires weaker assumptions and allows the sparsity level and the amount of contamination to depend on the number of predictors and on the sample size, respectively.

Our results are established under tighter bounds then those derived from direct but naïve extensions of existing results in feature selection. Moreover, we propose criteria to tune, in a computationally efficient and data-driven way, both the sparsity of the solution and the estimated amount of contamination.

The reminder of the article is organized as follows: Section 3.2 provides the relevant background. Section 3.3 details our proposal – including a general framework for SFSOD, the MIP formulation and its theoretical properties. Section 3.4 presents a simulation study comparing our proposal with state-of-the-art methods. Section 3.5 presents our application investigating the relationships between childhood obesity and microbiome composition. Final remarks are included in Section 3.6 and additional details are provided in Appendix A.

3.2 Background

Consider a regression model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ the error vector with a $N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ distribution ($\boldsymbol{I}_n$ is the identity matrix of size n), $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ the design matrix, and $\boldsymbol{\beta} \in \mathbb{R}^p$ the vector of regression

coefficients. In the following, we briefly review methods for outlier detection, and present the equivalent formulation as a feature selection problem. We then discuss approaches for model selection, focusing on the use of an L_0 constraint for best subset selection.

We consider a case-wise contamination mechanism, where each outlying unit might be contaminated in some (or even all) of its dimensions. Specifically, we assume that outliers follow an MSOM, where the set of outliers $M = \{i \in \{1, \dots, n\} : \varepsilon_i \sim N(\mu_{\varepsilon_i}, \sigma^2), \mu_{\varepsilon_i} \neq 0\}$ has cardinality $|M| = n_0$. For a given dimension $p \leq n - n_0$, MLE leads to the removal of outliers from the fit (Cook and Weisberg, 1982). Moreover, as is customary, we assume that the MSOM can also affect the design matrix $\boldsymbol{X}$ with mean shifts μ_{x_i} (Maronna et al., 2006).

If a regression comprises a single outlier, its position corresponds to the unit with largest absolute Studentized residual, which is a monotone transformation of the deletion residual $t_i = (y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{(i)})/(\widehat{\sigma}_{(i)}(1 + \boldsymbol{x}_i^T(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i)^{1/2}$, where the subscript (i) indicates the removal of the i-th unit. Under the null model, a generic t_i follows a Student's t with $n - p - 1$ degrees of freedom, which can be computed from an MLE fit based on all units and used as a test for outlying-ness of single data points (Cook and Weisberg, 1982). This can be easily generalized to regressions with multiple outliers. Operationally though, it was considered ineffective – due to the high likelihood of masking (undetected outlying cases) and swamping (non-outlying cases flagged as outliers) effects – and computationally intractable (Bernholt, 2006). The presence of multiple MSOM outliers motivates the use of high-breakdown point estimators such as the least trimmed squares (LTS), S, and MM (Maronna et al. 2006, see also Section 3.3.3); outlier detection and high-breakdown point estimation are historically distinct but closely related areas of statistical research.

Assuming without loss of generality that outliers occupy the first n_0 positions in the data, the MSOM can be equivalently parametrized as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}_{n_0}\boldsymbol{\phi} + \boldsymbol{\varepsilon}$, where the original design matrix $\boldsymbol{X}$ is augmented with a binary matrix $\boldsymbol{D}_{n_0} = [\boldsymbol{I}_{n_0}, \boldsymbol{0}]^T$ of size $n \times n_0$ indexing the n_0 outliers (Morgenthaler et al., 2004). If $p \leq n - n_0$, the MLE for $\boldsymbol{\phi} \in \mathbb{R}^{n_0}$ provides prediction residuals for the n_0 units excluded from the fit; i.e., their residuals under a model which excludes them from the estimation

process. This is given by $\widehat{\boldsymbol{\phi}} = [\boldsymbol{I}_{n_0} - \boldsymbol{H}_{MM}]^{-1} (\boldsymbol{y}_M - \boldsymbol{X}_M^T \widehat{\boldsymbol{\beta}}_{(M)})$, where $\boldsymbol{y}_M$ and $\boldsymbol{X}_M$ comprise values for the true set of outliers, $\boldsymbol{H}_{MM} = \boldsymbol{X}_M (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_M^T$, and the associated t-statistics $\boldsymbol{t}_M$ provide (multiple) deletion residuals. However, masking and swamping effects can again arise if $\boldsymbol{D}_{n_0}$ does not index all possible outliers.

Outlier detection in low-dimensional problems can be performed substituting the identity matrix $\boldsymbol{I}_n$ in place of $\boldsymbol{D}_{n_0}$ and applying feature selection methods to $\boldsymbol{\phi} \in \mathbb{R}^n$ to identify outlying cases. The literature contains examples of both convex (McCann et al., 2006; Taylan et al., 2014; Liu and Jiang, 2019; Taylan et al., 2021) and non-convex (She and Owen, 2011; Liu et al., 2017; Gómez, 2021; Barratt et al., 2020) penalization methods applied to this problem; notably, the latter are necessary to achieve high-breakdown point estimates.

Penalization methods are also the hallmark of feature selection in high dimensional problems, where they seek to induce sparsity estimating $p_0 < p$ non-zero coefficients in $\boldsymbol{\beta}$ – whose dimension p can exceed n. Soft penalization methods such as lasso (Tibshirani, 1996) and SCAD (Fan and Li, 2001) rely on non-differentiable continuous penalties, which can be convex or non-convex. They can be formulated as $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + R_\omega(\boldsymbol{\beta})$, where the penalty function $R_\omega(\boldsymbol{\beta})$ depends on a tuning parameter ω (or even more).

Best subset selection, a traditional hard penalization method, solves feature selection combinatorially, comparing all possible models of size p_0 (Miller, 2002). It can be formulated as a MIP through an L_0 constraint on $\boldsymbol{\beta}$, where the L_0 pseudo-norm is defined as $\|\boldsymbol{\beta}\|_0 = \sum_j I(\beta_j \neq 0)$ ($I(\cdot)$ is the indicator function). The MIP formulation of best subset selection is computationally intractable (Natarajan, 1995) and was previously considered impossible to solve with optimality guarantees for regression problems of realistic size. Nevertheless, improvements in optimization solvers and hardware components, which experienced a 450 billion factor speed-up between 1991 and 2015, now allow one to efficiently solve problems of realistic size with provable optimality (Bertsimas et al., 2016). Modern MIP solvers rely on implicit enumeration methods along with constraints such as *cutting planes* that tighten the relaxed problem (*branch & bound* and *branch & cut*, Schrijver 1986). Optimality is certified monitoring the gap between the best feasible solution and the problem

relaxation. Notably, MIP methods can recover the subset of true active features (i.e., they satisfy oracle properties, see Section 3.3.3) under weaker assumptions compared to soft penalization methods. Here the MIP formulation is not equivalent to the L_0-penalty due to non-convexity (Shen et al., 2013).

3.3 Proposed methodology

We focus on a regression comprising both outliers and inactive features, where one has to tackle at the same time an *unlabeled* MSOM problem (i.e., one where the identity, number and strength of outliers are unknown, Beckman and Cook 1983) and the sparse estimation of $\boldsymbol{\beta}$. SFSOD can be framed as an optimization problem; namely:

$$\left[\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}\right] = \underset{\boldsymbol{\beta},\boldsymbol{\phi}}{\arg\min} \sum_{i=1}^{n} \rho(y_i, f(\boldsymbol{x}_i; \boldsymbol{\beta}) + \phi_i) \qquad (3.1)$$
$$\text{s.t.} \quad R_\omega(\boldsymbol{\beta}) \leq c_\beta, \quad R_\gamma(\boldsymbol{\phi}) \leq c_\phi,$$

where $\rho(\cdot)$ is a loss function, $f(\cdot)$ defines the relation between predictors and response vector, and $R_\omega(\boldsymbol{\beta})$ and $R_\gamma(\boldsymbol{\phi})$ are penalties subject to sparsity-inducing constraints, which may depend on tuning constants ω and γ. Non-zero coefficients in $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\phi}}$ identify active features and outlying units, respectively. Although in this article we focus on linear regression the framework in (3.1) is very general; it comprises generalized linear models, several classification techniques and non-parametric methods.

Many approaches have been recently developed to solve (3.1) using ordinary least squares (OLS) as the loss function $\rho(\cdot)$. Both penalties $R_\omega(\boldsymbol{\beta})$ and $R_\gamma(\boldsymbol{\phi})$ are generally convex (Morgenthaler et al., 2004; Menjoge and Welsch, 2010; Lee et al., 2012; Kong et al., 2018) although some non-convex procedures have been considered (She and Owen, 2011). Robust soft penalization methods also can be cast into (3.1), abandoning the explicit use of $\boldsymbol{\phi}$ and adopting a robust loss $\rho(\cdot)$ in place of the OLS. These include MM-estimators for ridge regression (Maronna, 2011), sparseLTS (Alfons et al., 2013), bridge MM-estimators (Smucler and Yohai, 2017), enetLTS (Kurnaz et al., 2017), penalized elastic net S-estimators (Freue et al.,

2019), and penalized M-estimators (Loh, 2017; Chang et al., 2018; Amato et al., 2021), as well as their re-weighted counterparts. Indeed, through specific penalties, M-estimators can be equivalently formulated as feature selection problems (She and Owen, 2011).

While (3.1) highlights an important parallel between SFSOD and robust soft penalization, existing heuristic methods suffer several drawbacks. Some rely on restrictive assumptions or their finite-sample and asymptotic performance in terms of feature selection and outlier detection is not well-established. Others rely heavily on an initial subset of non-outlying cases. Yet others provide a down-weighting of all units, which complicates interpretation and the objective identification of outliers, or have an asymptotic breakdown point of 0%, so they in fact do not tolerate outliers in the first place. Finally, some methods require tuning of other parameters in addition to ω and γ, which can severely increase computational burden.

3.3.1 MIP formulation

Our proposal solves (3.1) with optimality guarantees, from both optimization and theoretical perspectives. This preserves the intrinsic discreteness of the problem, facilitating implementation, interpretation, and generalizations. We impose two separate integer constraints on $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ in (3.1), combining in a single framework the use of L_0 constraints for feature selection (Bertsimas et al. 2016; Bertsimas and Van Parys 2020; Kenney et al. 2021) and outlier detection (Zioutas et al. 2009; Bertsimas and Mazumder 2014). In particular, we consider the MIP formulation in (3.2) where $\boldsymbol{\mathcal{M}}^{\beta}$ and $\boldsymbol{\mathcal{M}}^{\phi}$ in constraints (3.2a) and (3.2b) are the so-called big-$\mathcal{M}$ bounds (Schrijver, 1986). In our proposal these are vectors of lengths p and n, respectively, which can be tailored for each β_j and ϕ_i. In the L_0-norm constraints (3.2c) and (3.2d), k_p and k_n are positive integers modulating sparsity for feature selection and outlier detection, respectively – for the latter, we can think of sparsity as a level of trimming (i.e., outlier removal). In the L_2-norm ridge-like constraint (3.2e), $\lambda > 0$ can be used to counteract strong collinearities among the features (Hoerl and Kennard, 1970b). It also modulates a trade-off between continuity and unbiasedness in the estimation of $\boldsymbol{\beta}$, and allows one to calibrate the intrinsic discreteness of

the problem – making its solutions more stable with respect to data perturbations (Breiman, 1995) and weak signal-to-noise ratio regimes (Hastie et al., 2020).

$$\left[\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}\right] = \underset{\boldsymbol{\beta}, \boldsymbol{z}^{\beta}, \boldsymbol{\phi}, \boldsymbol{z}^{\phi}}{\arg\min} \frac{1}{n} \rho(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\phi}) \tag{3.2}$$

$$\text{s.t.} \quad -\mathcal{M}_j^{\beta} z_j^{\beta} \leq \beta_j \leq \mathcal{M}_j^{\beta} z_j^{\beta} \tag{3.2a}$$

$$-\mathcal{M}_i^{\phi} z_i^{\phi} \leq \phi_i \leq \mathcal{M}_i^{\phi} z_i^{\phi} \tag{3.2b}$$

$$\sum_{j=1}^{p} z_j^{\beta} \leq k_p \tag{3.2c}$$

$$\sum_{i=1}^{n} z_i^{\phi} \leq k_n \tag{3.2d}$$

$$\sum_{j=1}^{p} \beta_j^2 \leq \lambda \tag{3.2e}$$

$$z_j^{\beta} \in \{0, 1\}, \quad \beta_j \in \mathbb{R}, \quad j = 1, \ldots, p$$

$$z_i^{\phi} \in \{0, 1\}, \quad \phi_i \in \mathbb{R}, \quad i = 1, \ldots, n.$$

Although solving (3.2) plainly with state-of-the-art software may be computationally intractable for large dimensions, with the appropriate implementation it can be used to tackle many real-world applications optimally and efficiently. In general, what it means for a statistical problem to be small or large depends on its structure – for instance, the "signal-to-noise" ratio plays a fundamental role (see Sections 3.3.3 and 3.4). From an operational standpoint, in this setting a problem can be considered large if the signal-to-noise ratio is small or moderate (e.g., smaller than 2), and the sample size and number of features are in the thousands or more. Another important advantage of our proposal from an application standpoint is that it allows one to easily incorporate additional constraints to leverage structure in the data – such as groups, ranked features, hierarchical interactions, and compositional information. We note that (3.2) could undergo a transformation through perspective cut model (Frangioni and Gentile, 2006) that may be of interest to improve the performance of the solution algorithm, here omitted for brevity.

3.3.2 Some implementation details

Setting the big-$\mathcal{M}$ bounds for (3.2) is made even more complicated due to the "double" nature of SFSOD. A robust estimator of the regression coefficients, say $\widetilde{\boldsymbol{\beta}}$, can be used to set $\boldsymbol{\mathcal{M}}^{\beta} = \widetilde{\boldsymbol{\beta}}c$ and $\boldsymbol{\mathcal{M}}^{\phi} = (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})c = \widetilde{\boldsymbol{e}}c$, where $c \geq 1$ is a suitable multiplicative constant. We generalize this approach using an *ensemble* $\widetilde{\boldsymbol{\beta}}_t$ (for $t = 1, \ldots, T$) of preliminary estimators and setting $\mathcal{M}_j^{\beta} = \max_t(|\widetilde{\beta}_{t_j}|)c$ and $\mathcal{M}_i^{\phi} = \max_t(|\widetilde{e}_{t_j}|)c$. The ensemble guarantees that, if at least one of the $\widetilde{\boldsymbol{\beta}}_t$'s is reasonably close to the optimal solution, the MIP will easily recover such solution. Importantly, having also non-robust or non-sparse estimators in the ensemble does not negatively affect solution quality but only convergence speed.

The MIP formulation in (3.2) critically depends on the big-$\mathcal{M}$ bounds; they should be large enough to retain the optimal solution, yet small enough to avoid unnecessary computations and numerical instability. If identifying suitable bounds is not possible, we use an alternative strategy based on *specially ordered sets of type 1* (SOS-1; Bertsimas et al. 2016). These allow only one variable in the set to be non-zero, e.g., $(1 - z_j^{\beta}, \beta_j) = 0 \iff (1 - z_j^{\beta}, \beta_j)$: SOS-1, which can be solved via modern MIP solvers such as `Gurobi` or `CPLEX`. SOS-1 constraints in (3.2) guarantee that the global optimum can be reached, and generally outperform big-$\mathcal{M}$ bounds when these are difficult to reasonably set.

The formulation in (3.2) also, and critically, requires the tuning of k_p, k_n and, if a ridge-like constraint is included in the model, λ. Performing this simultaneously along an extensive grid of values can be computationally unviable for MIP. We therefore proceed as follows: **(i)** fix λ (possibly, in turn, to a few values in a meaningful range); **(ii)** fix k_n to a starting value larger than a reasonable expectation on the amount of contamination in the problem (n_0); **(iii)** holding fixed the k_n starting value from (ii), tune k_p through cross-validation or an information criterion; **(iv)** holding fixed the k_p value selected in (iii), refine downward the value of k_n. See also She and Owen (2011) for a discussion on parameter tuning for outlier detection. To the best of our knowledge, this is still an open research area, especially in high-dimensional settings. In our numerical studies, we found that there

was little difference in choosing one tuning approach over the other. Thus, in this work we mainly focus on a robust counterpart of the BIC. We provide further details concerning feature standardization, cross-validation, selecting a trimming level, and using information criteria in Appendix A.2.

3.3.3 Theoretical results

In this Section, we characterize the theoretical properties of our proposal through two groups of results. The first comprises properties established under the general framework introduced in (3.2). The second comprises key properties established under an L_2-norm loss function $\rho(\cdot) = \|\cdot\|_2^2$; namely, the *robustly strong oracle property* and *optimal parameter estimation* for SFSOD. All proofs are provided in Appendix A.1.

Without loss of generality, we assume that (3.2) has a unique global minimum, and that the loss function is such that $\rho(\boldsymbol{x}) \geq 0$ with $\rho(\mathbf{0}) = 0$ (this is the case for OLS and many other instances, such as estimation in quantile regression and robust estimators). Our first result connects our proposal to a large class of penalized methods based on trimming.

Theorem 3.1 (Sparse trimming)**.** *For any λ, n, p, k_n and k_p, the $\widehat{\boldsymbol{\beta}}$ estimator produced solving* (3.2) *is the same as the one produced solving*

$$\begin{aligned} \arg\min_{\boldsymbol{\beta}} \ & \frac{1}{n}\sum_{i=1}^{n-k_n}\{\rho(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\}_{i:n} = \frac{1}{n}\sum_{i=1}^{n-k_n}\{\rho(e_i)\}_{i:n} \\ \text{s.t.} \quad & (3.2a), (3.2c), (3.2e), \end{aligned} \tag{3.3}$$

where e_i (for $i = 1, \ldots, n$) are the residuals, and $\{\rho(e_1)\}_{1:n} \leq \ldots \leq \{\rho(e_n)\}_{n:n}$ the order statistics of their $\rho(\cdot)$ transformation.

Theorem 3.1 demonstrates the equivalence of our formulation to a trimmed loss problem, where the level of trimming is directly controlled by the L_0 constraint on $\boldsymbol{\phi}$. This extends a well-known result for unpenalized OLS and motivates the formulation in (3.1) as a general framework for SFSOD. In particular, (3.3) includes some trimmed likelihood estimators as special cases (Hadi and Luceño, 1997). Thus,

our proposal inherits their desirable properties, such as equivariance if the points are in general position (Maronna et al., 2006).

The largest proportion of outliers that an estimator can tolerate before becoming arbitrarily biased is referred to as the breakdown point. In symbols, consider a sample $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ with $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i^T)$. The *maximum bias* for an estimator, say $\boldsymbol{\tau}$, is $b^*(n_0; \boldsymbol{\tau}, \boldsymbol{Z}) = \sup_{\widetilde{\boldsymbol{Z}}} \|\boldsymbol{\tau}(\widetilde{\boldsymbol{Z}}) - \boldsymbol{\tau}(\boldsymbol{Z})\|_2$, where $\widetilde{\boldsymbol{Z}}$ represents $\boldsymbol{Z}$ after the replacement of n_0 points by arbitrary values. The *finite-sample replacement breakdown point* (BdP henceforth), defined as $\epsilon^*(\boldsymbol{\tau}, \boldsymbol{Z}) = \min_{n_0}\{n_0/n : b^*(n_0; \boldsymbol{\tau}, \boldsymbol{Z}) \to \infty\}$, is the maximum proportion of observations that, when arbitrarily replaced, still provide bounded estimates (Donoho and Huber, 1983). Our second result shows that our MIP approach for SFSOD achieves arbitrarily large BdP.

Theorem 3.2 (MIP breakdown point)**.** *For any λ, n, p, k_n and k_p, where $(y_i, \boldsymbol{x}_i^T)$ are not necessarily in general position, the BdP of the $\widehat{\boldsymbol{\beta}}$ estimator produced solving* (3.2) *is $\epsilon^* = (k_n + 1)/n$.*

Thus, $k_n \geq n_0$ is the only requirement to achieve the largest possible BdP. Similar results were obtained for the least quantile estimator (Bertsimas and Mazumder, 2014), the LTS estimator with a lasso penalty (Alfons et al., 2013), and MM-estimators with a ridge or elastic net penalty (Maronna 2011; Kurnaz et al. 2017). However, there are two caveats: the BdP can be misleading for non-equivariant estimators (Smucler and Yohai, 2017), and it only guarantees against the worst-case scenario – infinite maximum bias – as it does not account for large but finite biases in $\widehat{\boldsymbol{\beta}}$. This motivates the development of additional theoretical guarantees.

Next, we exclude the ridge-like penalty and take $\rho(\cdot) = \|\cdot\|_2^2$, making (3.2) a *mixed-integer quadratic program* (MIQP). In this setting, we prove that under certain conditions our approach satisfies the *robustly strong oracle property* (see Definition 1, based on Fan et al. 2014a). In the following, we use the L_0 sparsity assumption on $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ as in Zhang and Zhang (2012). Recall that MSOM leads to outlier removal (see Section 3.2), and we showed in Theorem 3.1 that the L_0 constraint on $\boldsymbol{\phi}$ controls the level of trimming from the fit, thus this sparsity assumption on $\boldsymbol{\phi}$ is equivalent to the presence of MSOM outliers. In our SFSOD problem, as customary in feature selection literature, let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\phi}_0^T)^T \in \mathbb{R}^{p+n}$ be the true parameter vector, and

decompose it as $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{\mathcal{S}}^T, \boldsymbol{\theta}_{\mathcal{S}^c}^T)^T = \{(\boldsymbol{\beta}_{\mathcal{S}_\beta}^T, \boldsymbol{\phi}_{\mathcal{S}_\phi}^T), (\boldsymbol{\beta}_{\mathcal{S}_\beta^c}^T, \boldsymbol{\phi}_{\mathcal{S}_\phi^c}^T)\}^T$ where $\boldsymbol{\theta}_{\mathcal{S}}$ contains only the true non-zero regression coefficients. Let the *robust oracle estimator* be $\widehat{\boldsymbol{\theta}}_0 = (\boldsymbol{A}_{\mathcal{S}}^T \boldsymbol{A}_{\mathcal{S}})^{-1} \boldsymbol{A}_{\mathcal{S}}^T \boldsymbol{y}$, where $\boldsymbol{A}_{\mathcal{S}} = (\boldsymbol{X}_{\mathcal{S}_\beta}, \boldsymbol{I}_{\mathcal{S}_\phi})$ is the $n \times (p_0 + n_0)$ matrix restricted to the active features belonging to $\mathcal{S}_\beta$ and the outlying cases belonging to $\mathcal{S}_\phi$. The robust oracle estimator is akin to the oracle estimator in feature selection – where the oracle is simply the OLS solution across the active set, when the features belonging to it are known. Our robust oracle estimator extends this concept taking also outliers into account. Specifically, $\widehat{\boldsymbol{\theta}}_0$ behaves as if the sets of active features and outliers were both known in advance. Indeed, if we know which points are outliers, we can include dummies for them, effectively removing their influence on the fit and making the OLS the optimal estimator.

Definition 3.1 (Robustly strong oracle property)**.** *An estimator $\widehat{\boldsymbol{\beta}}$ with support $\widehat{\mathcal{S}}_\beta$ satisfies the robustly strong oracle property if (asymptotically) there exists tuning parameters which guarantee $P(\widehat{\mathcal{S}}_\beta = \mathcal{S}_\beta) \geq P(\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_0) \to 1$ in the presence of MSOM outliers.*

Such robust version of the oracle property is stronger and more general than the oracle property in the sense of Fan and Li (2001), as it implies both SFSOD consistency and sign consistency (see also Bradic et al. 2011). Thus, SFSOD consistency depends on the achievability of the robust oracle estimator which we investigate by extending the theoretical framework in Shen et al. (2013) for feature selection. This requires weaker assumptions compared to other penalization methods (Zhang and Zhang, 2012), and we generalize it to the presence of MSOM outliers. Intuitively, if the robust oracle estimator is achievable (i.e., if it has the lowest objective for models of the same size), it is also the solution of our MIQP when the integer constraints are set to $k_p = p_0$ and $k_n = n_0$. Achievability depends on the difficulty of the problem, as measured by the *minimal degree of separation* between the true and a least favorable model – indexed by the supports $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, respectively. This is defined as $\Delta_\theta(\boldsymbol{A}) = \min_{\boldsymbol{\theta}_{\widetilde{\mathcal{S}}}} \|\boldsymbol{A}_{\mathcal{S}}\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{A}_{\widetilde{\mathcal{S}}}\boldsymbol{\theta}_{\widetilde{\mathcal{S}}}\|_2^2 / \{n \max(|\mathcal{S} \backslash \widetilde{\mathcal{S}}|, 1)\}$ (for $\boldsymbol{\theta}_{\widetilde{\mathcal{S}}} : \widetilde{\mathcal{S}} \neq \mathcal{S}, |\widetilde{\mathcal{S}}_\beta| \leq p_0, |\widetilde{\mathcal{S}}_\phi| \leq n_0$), which relates to the signal-to-noise ratio and can be bounded as $\Delta_\theta \leq \Delta_\beta + \Delta_\phi$ (with Δ_β and Δ_ϕ defined similarly to Δ_θ using $\boldsymbol{X}$ and $\boldsymbol{I}_n$, respectively). We control this level of difficulty in Theorem 3.3, which provides

a *necessary* condition for SFSOD consistency over $B(u,l) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_0 \leq u, \Delta_\theta \geq l\}$, the L_0-band with upper and lower radii u and l, respectively (a subset of the L_0-ball $B(u,0)$ excluding a neighborhood of the origin).

Theorem 3.3 (Necessary condition for SFSOD consistency)**.** *For any support estimate $\widehat{\mathcal{S}}$ and $u > l > 0$, $\sup_{\boldsymbol{\theta}_0 \in B(u,l)} P(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$ implies that*

$$\Delta_\theta \geq l = \frac{\sigma^2}{n} \max\left\{d_\beta \log(p), d_\phi \log(n)\right\}, \tag{3.4}$$

where $d_\beta > 0$ (which may depend on $\boldsymbol{X}$) and $d_\phi > 0$ are constants independent of n and p.

This lower bound on Δ_θ indicates one can focus on solving the most difficult task between outlier detection and feature selection; if this is achievable, *a fortiori*, the other will be as well. Next, we provide a *sufficient* condition for SFSOD consistency based on a finite-sample result bounding the probability that our proposal differs from the robust oracle estimator.

Theorem 3.4 (MIQP oracle reconstruction)**.** *For any n, p, n_0 and p_0, the $\widehat{\boldsymbol{\theta}}_{L_0}$ estimator produced solving* (3.2) *with $k_p = p_0$ and $k_n = n_0$ is such that*

$$\begin{aligned} P\left(\widehat{\boldsymbol{\theta}}_{L_0} \neq \widehat{\boldsymbol{\theta}}_0\right) \leq & \frac{5e-1}{e-1} \max\left[\exp\left\{-\frac{n}{18\sigma^2}\left(\Delta_\beta - 36\sigma^2\frac{\log(p)}{n}\right)\right\}, \right. \\ & \left. \exp\left\{-\frac{n}{18\sigma^2}\left(\Delta_\phi - 36\sigma^2\frac{\log(n)}{n}\right)\right\}\right]. \end{aligned} \tag{3.5}$$

Based on these results, one can easily prove the robustly strong oracle property as follows.

Theorem 3.5 (MIQP robustly strong oracle property)**.** *Assume that $u_\theta = u_\phi + u_\beta$, where $u_\phi < n - k_p$ and $u_\beta < \min(n - k_n, p)$, and that there exists a constant $d_\theta > 36$ such that $l_\theta = d_\theta \sigma^2 / n \max\{\log(p), \log(n)\}$. Then, under* (3.4) *and for $(n,p) \to \infty$, the estimator $\widehat{\boldsymbol{\theta}}_{L_0}$ produced solving* (3.2) *with $k_p = p_0$ and $k_n = n_0$ satisfies*

1. *Robustly strong oracle property:*

$$\sup_{\boldsymbol{\theta}_0 \in B(u_\theta, l_\theta)} P(\widehat{\mathcal{S}}^{L_0} = \mathcal{S}) \geq \sup_{\boldsymbol{\theta}_0 \in B(u_\theta, l_\theta)} P(\widehat{\boldsymbol{\theta}}_{L_0} = \widehat{\boldsymbol{\theta}}_0) \to 1$$

uniformly over $B(u_\theta, l_\theta) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_0 = (p_0 + n_0) \leq u_\theta, \Delta_\theta \geq l_\theta\}$.

2. *Asymptotic normality:*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{L_0} - \boldsymbol{\theta}_0) \to^d N(\mathbf{0}, \boldsymbol{\Sigma}_\theta),$$

where $\boldsymbol{\Sigma}_\theta = \sigma^2(\boldsymbol{A}_{\mathcal{S}}^T \boldsymbol{A}_{\mathcal{S}}/n)^{-1}$.

Theorem 3.5(1) provides a sufficient condition for SFSOD consistency and the robust oracle reconstruction up to a constant term d_θ. Note that the number of features is allowed to exponentially increase with the sample size – so these properties hold also in ultra-high dimensional settings where $p = \mathcal{O}(e^{n\alpha})$ with $\alpha = \Delta_\theta/(d_\theta\sigma^2) > 0$. Theorem 3.5(2) guarantees asymptotic normality and efficiency of MIQP estimates, which achieve the Cramèr–Rao lower bound as if the true sets of features and outliers were known a priori. Thus, although finite-sample inference with our approach can be problematic, as with other robust and/or regularization approaches, "large sample" statistical inference can be performed. Importantly, existing penalized M-estimators provide weaker results under stronger assumptions (Loh, 2017; Smucler and Yohai, 2017; Amato et al., 2021). We conclude with a result showing that our proposal attains optimal parameter estimation with respect to the L_2-norm in the presence of MSOM outliers.

Theorem 3.6 (MIQP optimal parameter estimation)**.** *Under the same conditions of Theorem 3.5, the estimator* $\widehat{\boldsymbol{\theta}}_{L_0}$ *produced solving* (3.2) *with* $k_p = p_0$ *and* $k_n = n_0$ *provides*

1. *Optimal* L_2*-norm prediction error:*

$$n^{-1}E\|\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_{L_0} - \boldsymbol{\theta}_0)\|_2^2 = \sigma^2(p_0 + n_0)/n.$$

2. *Risk-minimax optimality for parameter estimation:*

$$\sup_{\boldsymbol{\theta}_0 \in B(u_\theta, l_\theta)} n^{-1}E\|\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_{L_0} - \boldsymbol{\theta}_0)\|_2^2 = \inf_{\boldsymbol{\tau}_n} \sup_{\boldsymbol{\theta}_0 \in B(u_\theta, l_\theta)} n^{-1}E\|\boldsymbol{A}(\boldsymbol{\tau}_n - \boldsymbol{\theta}_0)\|_2^2 = \sigma^2 u_\theta/n.$$

Finally, the theoretical guarantees developed in this section can be extended in a similar fashion to other penalization methods, albeit under stronger assumptions. For instance, one might consider the regularized L_0-penalty or the trimmed L_1-penalty. Importantly, our results do hold also when p_0 depends on p and/or n_0 depends on n which has yet to be established for other methods in the literature (Shen et al., 2013). We stress the fact that all results for the proposed formulation rely on the identification of the true k_p and k_n tuning parameters. While this is a standard requirement to establish oracle properties (see Fan and Li 2001), it highlights the importance of proper tuning for these bounds. For this reason, in Section 3.2 we propose robust methods to effectively tune the two integer constraints.

3.4 Simulation study

We use simulations to study the performance of our proposal and compare it with state-of-the-art heuristic methods. The simulated data is generated as follows. The first column of the $n \times p$ design matrix $\boldsymbol{X}$ comprises all 1's (for the model intercept) and we draw the remaining entries of each row independently from a $(p-1)$-variate Gaussian $N(\boldsymbol{0}, \boldsymbol{\Sigma}_X)$, we fix the values of the p-dimensional coefficient vector $\boldsymbol{\beta}$ as to comprise p_0 non-zero entries (including the intercept), and we draw each entry of the n-dimensional error vector $\boldsymbol{\varepsilon}$ independently from a univariate Gaussian $N(0, \sigma^2_{\text{SNR}})$. Here $\sigma^2_{\text{SNR}} > 0$ is used to modulate the signal-to-noise ratio $\text{SNR} = \text{var}(\boldsymbol{X}\boldsymbol{\beta})/\sigma^2_{\text{SNR}}$ characterizing each experiment. Next, without loss of generality, we contaminate the first n_0 cases with an MSOM, adding the scalar mean shifts μ_ε and μ_X, respectively, to the errors and each of the $p_0 - 1$ active predictors.

Specific simulation scenarios are defined through the values of the parameters listed above. Here, we present results for $\boldsymbol{\Sigma}_X = \boldsymbol{I}_{p-1}$ (uncorrelated features), $p_0 = 5$ active features with $\beta_j = 2$ (without loss of generality these correspond to $j = 1, \ldots, 5$), $\text{SNR} = 5$, fraction of contamination $n_0/n = 0.1$, mean shifts $\mu_\varepsilon = -10$ and $\mu_X = 10$, increasing sample sizes $n = 50, 100, 150$, and a "low"- and a "high"-dimensional setting with $p = 50, 200$. Results for additional simulation scenarios are provided in Appendix A.2.

Replicating each scenario a certain number of (independent) times, say q, and creating (independent) test data, say $(\boldsymbol{y}^*, \boldsymbol{X}^*)$, from the same generating scheme but without contamination, we compare methods with a variety of criteria, namely: (i) out-of-sample prediction performance, measured by the *root mean squared prediction error* $\text{RMSPE} = \{n^{-1}\sum_{i=1}^{n}(y_i^* - \boldsymbol{x}_i^*\widehat{\boldsymbol{\beta}})^2\}^{1/2}$; (ii) estimation accuracy for $\boldsymbol{\beta}$, measured by the *average mean squared error* $\text{MSE}(\widehat{\boldsymbol{\beta}}) = p^{-1}\sum_{j=1}^{p}\text{MSE}(\widehat{\beta}_j)$, where for each $\widehat{\beta}_j$ we form $\text{MSE}(\widehat{\beta}_j) = q^{-1}\sum_{i=1}^{q}(\widehat{\beta}_{ji} - \beta_j)^2 = (\overline{\beta}_j - \beta_j)^2 + q^{-1}\sum_{i=1}^{q}(\widehat{\beta}_{ji} - \overline{\beta}_j)^2$, decomposed in squared bias and variance (here $\overline{\beta}_j = q^{-1}\sum_{i=1}^{q}\widehat{\beta}_{ji}$); (iii) feature selection accuracy, measured by the *false positive rate* $\text{FPR}(\widehat{\boldsymbol{\beta}}) = |\{j \in \{1,\ldots,p\} : \widehat{\beta}_j \neq 0 \wedge \beta_j = 0\}|/|\{j \in \{1,\ldots,p\} : \beta_j = 0\}|$ and the *false negative rate* $\text{FNR}(\widehat{\boldsymbol{\beta}}) = |\{j \in \{1,\ldots,p\} : \widehat{\beta}_j = 0 \wedge \beta_j \neq 0\}|/|\{j \in \{1,\ldots,p\} : \beta_j \neq 0\}|$, as well as the F_1 score – which is a mixture of the two defined as $F_1(\widehat{\boldsymbol{\beta}}) = (1 - \text{FNR})/\{(1 - \text{FNR}) + (\text{FPR} + \text{FNR})/2\}$; (iv) outlier detection accuracy, which is similarly measured by $\text{FPR}(\widehat{\boldsymbol{\phi}})$, $\text{FNR}(\widehat{\boldsymbol{\phi}})$ and $F_1(\widehat{\boldsymbol{\phi}})$; (v) computational burden, measured as CPU time in seconds (this is used as a rough evaluation, since software implementations of different methods are not entirely comparable).

Using the robust oracle estimator as a benchmark, we compare the following estimators: (a) sparseLTS (Alfons et al., 2013), (b) enetLTS (Kurnaz et al., 2017), and (c) our MIP proposal (see Section 3.3). All methods trim the true number of outliers ($k_n = n_0$) and only their feature sparsity level is tuned. See Appendix A.2 for implementation details.

Table 3.1 provides means and standard deviations (SD) of simulation results over $q = 1000$ replications. Our proposal substantially outperforms competing methods in most criteria. In particular, for the low-dimensional setting ($p = 50$), its RMSPE converges faster to the oracle solution and the variance of its $\widehat{\boldsymbol{\beta}}$ decreases faster as n increases (the bias is essentially non-existent for all methods). Notably, the $\text{FPR}(\widehat{\boldsymbol{\beta}})$ of sparseLTS and enetLTS increases with the sample size, while our approach avoids these type II errors. Even with these sparser solutions, we retain comparable (and at times lower) $\text{FNR}(\widehat{\boldsymbol{\beta}})$. Our method struggles most when $n = 50$, suggesting that additional work for tuning MIP may be beneficial in under-sampled problems. All methods perform very well in terms of $\text{FPR}(\widehat{\boldsymbol{\phi}})$ and $\text{FNR}(\widehat{\boldsymbol{\phi}})$, though enetLTS is

Table 3.1: Mean (SD in parenthesis) of RMSPE, variance and squared bias for $\widehat{\boldsymbol{\beta}}$, FPR and FNR for feature selection and outlier detection (as well as the corresponding F_1 scores), and computing time, based on 1000 simulation replications.

n	p	Method	RMSPE	var($\widehat{\beta}$)	bias($\widehat{\beta}$)^2	FPR($\widehat{\beta}$)	FNR($\widehat{\beta}$)	$F_1(\widehat{\beta})$	FPR($\widehat{\phi}$)	FNR($\widehat{\phi}$)	$F_1(\widehat{\phi})$	Time
50	50	Oracle	1.87(0.27)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	2.53(0.94)	0.05(0.02)	0.03(0.00)	0.18(0.22)	0.02(0.11)	0.91	0.01(0.02)	0.04(0.17)	0.98	12.14(0.78)
		SparseLTS	2.46(0.48)	0.06(0.00)	0.00(0.00)	0.54(0.07)	0.00(0.03)	0.79	0.00(0.01)	0.00(0.06)	1.00	3.50(0.67)
		MIP	2.17(0.76)	0.04(0.01)	0.00(0.00)	0.00(0.01)	0.08(0.16)	0.96	0.00(0.01)	0.01(0.08)	1.00	10.69(18.67)
100	50	Oracle	1.83(0.18)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	2.00(0.23)	0.01(0.00)	0.00(0.00)	0.28(0.21)	0.00(0.00)	0.88	0.00(0.01)	0.00(0.00)	1.00	9.69(0.29)
		SparseLTS	2.12(0.23)	0.03(0.00)	0.00(0.00)	0.66(0.08)	0.00(0.00)	0.75	0.00(0.01)	0.00(0.00)	1.00	4.07(0.71)
		MIP	1.89(0.34)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.01(0.06)	0.99	0.00(0.00)	0.00(0.00)	1.00	36.31(26.75)
150	50	Oracle	1.81(0.15)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	1.93(0.17)	0.01(0.00)	0.00(0.00)	0.40(0.25)	0.00(0.00)	0.84	0.00(0.01)	0.00(0.00)	1.00	10.18(0.33)
		SparseLTS	2.00(0.16)	0.02(0.00)	0.00(0.00)	0.68(0.08)	0.00(0.00)	0.75	0.00(0.01)	0.00(0.00)	1.00	4.23(0.88)
		MIP	1.83(0.22)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.01(0.04)	1.00	0.00(0.00)	0.00(0.00)	1.00	382.74(228.66)
50	200	Oracle	1.88(0.28)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	3.38(1.23)	0.03(0.00)	0.02(0.00)	0.19(0.14)	0.20(0.31)	0.81	0.02(0.03)	0.19(0.31)	0.89	36.26(3.03)
		SparseLTS	2.85(0.85)	0.02(0.00)	0.01(0.00)	0.17(0.02)	0.06(0.19)	0.89	0.01(0.02)	0.06(0.21)	0.96	3.69(0.83)
		MIP	2.44(1.14)	0.02(0.00)	0.00(0.00)	0.00(0.01)	0.13(0.24)	0.93	0.01(0.02)	0.05(0.19)	0.97	24.07(48.81)
100	200	Oracle	1.84(0.19)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	2.79(1.22)	0.02(0.00)	0.01(0.00)	0.24(0.12)	0.10(0.22)	0.84	0.02(0.03)	0.13(0.28)	0.92	46.25(4.45)
		SparseLTS	2.34(0.25)	0.01(0.00)	0.00(0.00)	0.31(0.02)	0.00(0.00)	0.87	0.00(0.01)	0.00(0.00)	1.00	10.02(2.09)
		MIP	1.90(0.35)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.06)	0.99	0.00(0.00)	0.00(0.00)	1.00	334.76(630.38)
150	200	Oracle	1.81(0.14)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	2.47(1.13)	0.02(0.00)	0.00(0.00)	0.23(0.13)	0.06(0.17)	0.87	0.01(0.02)	0.09(0.24)	0.95	48.45(3.82)
		SparseLTS	2.25(0.20)	0.01(0.00)	0.00(0.00)	0.41(0.04)	0.00(0.00)	0.83	0.00(0.01)	0.00(0.00)	1.00	14.01(2.03)
		MIP	1.84(0.22)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.01(0.04)	1.00	0.00(0.00)	0.00(0.00)	1.00	832.13(890.91)

slightly worse for $n = 50$. As expected, the computational burden of our procedure is substantially higher than that of the competing heuristic methods – though we note that averages here are not representative, as there is a marked right skew in the distribution of computing times across replications. For comparison we provide medians and median absolute deviations (MAD) in Table A.1 and find that results are even stronger. For example, the average computing time with $n = 150$ and $p = 200$ is 832.13 minutes compared to a median of 518.92 minutes. Our experience suggests that the growth in computational burden is mainly due to increases in the absolute number of outliers as the sample size increases.

Similar conclusions hold under the high-dimensional scenario with $n < p = 200$. In Appendix A.2 we report results for additional simulation scenarios, e.g., with smaller SNR, collinear features and weaker mean shift parameters, where our method also outperformed others in most settings.

3.5 Connecting childhood obesity and microbiome composition

We now return to the application described in Section 3.1, investigating the relationship between childhood obesity and microbiome composition. All data are publicly available; we accessed microbiome reads and phenotype information from the Sequence Read Archive (SRA, 2017) and database of Genotypes and Phenotypes (dbGaP, 2017) through the National Center for Biotechnology Information (NCBI), respectively. The goal of our analysis is to study which bacterial types (features) may affect children's weight gain accounting for potential outlying cases.

We focused on the *oral* microbiota of children and their mothers, which were found to contain interesting signals in Craig et al. (2018). Based on the pre-processing in Craig et al. (2018), we retained 215 child and 215 maternal oral samples. Correspondingly, we considered the abundances of 67 and 62 bacterial groups, respectively – which the original authors obtained aggregating phylogenetically sparse and correlated abundance data (we further filtered based on those with a MAD of 0 and/or exhibiting 0 counts in half or more of the samples). We also

Table 3.2: Median (MAD in parenthesis) of TMSPE and the number of features selected on the training set on eight train-test splits. Last column: number of features selected on the full data. Robust methods use 20% trimming.

Data	n^{tr}	n^{te}	p	Method	TMSPE	$\widehat{p}_0^{\text{tr}}$	$\widehat{p}_0^{\text{full}}$
Child oral	172	43	68	SparseLTS	0.25(0.03)	54.00(0.26)	52
				EnetLTS	0.12(0.02)	47.00(3.67)	52
				MIP	0.18(0.04)	13.00(0.52)	13
				Lasso	0.19(0.02)	2.00(0.26)	2
Maternal oral	172	43	63	SparseLTS	0.21(0.04)	52.00(1.31)	56
				EnetLTS	0.20(0.03)	52.00(4.46)	62
				MIP	0.15(0.03)	13.50(0.52)	13
				Lasso	0.18(0.02)	1.00(0.00)	1

log-transformed the abundances of each group to mitigate skews. We focused on one among the phenotypes studied in Craig et al. (2018); namely, the *conditional weight gain score* (CWG) – a continuous measure computed from weight gain between birth and six months (a positive CWG indicates an accelerated weight gain) which is commonly used in paediatric research (Savage et al., 2016).

We thus applied our approach along with sparseLTS, enetLTS and classical lasso to two main models; the regressions of children's CWG on log-transformed bacterial groups abundances in oral samples of the children themselves, and of their mothers, respectively. The problem sizes were 215×68 and 215×63 with the inclusion of an intercept term. In addition to applying our approach on the full datasets, we split the data at random into training ($n^{\text{tr}} \approx 0.8n$) and test ($n^{\text{te}} \approx 0.2n$) sets to assess out-of-sample prediction performance of the various procedures. We also considered a different splitting ratio ($n^{\text{tr}} \approx 0.9n$ and $n^{\text{te}} \approx 0.1n$) and obtained similar results (see Table A.4). Since the true contamination level of a given test set is unknown, we calculated a trimmed median squared prediction error (TMSPE) at 50% to be conservative. The trimming level when fitting each robust procedure was set to that found on the full dataset (20% for both children and maternal regressions, which corresponds to 43 cases). We repeated the analysis on 8 different training/test splits for all methods. Table 3.2 provides, for each of the two regressions, medians and MADs of results over the 8 splits – including TMSPE and the number of features selected on the training set ($\widehat{p}_0^{\text{tr}}$). The last column contains the total number of

features selected on the full data ($\widehat{p}_0^{\text{full}}$).

For the regressions on the full datasets, sparseLTS and enetLTS selected a very large number of bacterial groups, hindering interpretation. In contrast, the lasso produced very sparse solutions – so sparse that it only selected the intercept for the maternal regression. For the children regression, lasso selected one bacterial group belonging to the Firmicutes phylum and coinciding with a group selected in Craig et al. (2018). This sparse behavior was consistent across the 8 training/test splits as well. However, MIP outperformed lasso in both regressions based on TMSPE – especially compared to the intercept-only model identified by lasso for the maternal regression. EnetLTS was the most predictive method for the children regression, but MIP again outperformed it in the maternal regression.

In terms of sparsity, our procedure produced solutions much more parsimonious (and thus more interpretable) than those of the other robust methods, but less sparse (and thus more informative) than those of the lasso. MIP selected 13 bacterial groups for both the children and the maternal regression, albeit they were tuned independently. One group among the ones identified in each regression was also found to be related to children's growth curves and rapid infant weight gain in Craig et al. (2018). These were a Bacteriodetes and a Fusobacteria group in the children and maternal oral microbiota, respectively. Interestingly, the Bacteriodetes group contains bacteria from the Porphyromonas genus, which has species capable of producing Butyrate (Vital et al., 2014) – a fatty acid associated with obesity (Liu et al., 2018). Further connections can be found with prior findings reported in the literature. For instance, though our response is CWG, a Firmicutes group selected by our procedure in the maternal oral microbiome consisted of one main genus, namely Streptococcus, which was significantly related to maternal body mass index in Cabrera-Rubio et al. (2012).

Switching to outlier detection, our procedure detected 43 outliers for both the children and the maternal regression. 17 infants with particularly extreme CWG scores in either direction (see Figure A.1) were detected as outliers in both regressions, with extreme (positive or negative) standardized residuals (see Figure A.2). The child with the highest CWG and the largest residual in the children regression

was one of the few infants (15/215) whose mother smoked while pregnant. Notably these 15 children have a significantly higher CWG on average (p-value = 0.042; one-sided t-test), and 40% of them were detected as outliers in either or both of our regressions.

Overall, these results show that our proposal is competitive in terms of predictive power (compared to other robust and non-robust methods), while providing parsimonious, interpretable and informative solutions consistent with literature and effectively detecting outliers. See Appendix A.3 for additional remarks and discussion regarding warm-starts and big-$\mathcal{M}$ bounds used in this analysis, as well as an attempt to further validate our findings exploiting phenotypes studied in previous literature.

3.6 Final remarks

Our proposal provides a general framework to simultaneously perform sparse estimation and outlier detection that can be used for linear models, as well as generalized linear models and several classification and non-parametric methods (Yerlikaya-Özkurt and Taylan, 2020). In our main results, we focus specifically on linear models (as do existing heuristic approaches) – but we directly tackle the original problem and preserve its discrete nature; this facilitates implementation, interpretation, and generalizations. Importantly, we provide optimal guarantees from both optimization and theoretical perspectives, and verify that these hold in numerical experiments.

Our approach relies on L_0 constraints – extending prior work where they were used separately for feature selection or outlier detection. Our simultaneous MIP formulation can handle problems of considerable size, and produces solutions that improve upon existing heuristic methods. Although our formulation provides provably optimal solutions from the optimization perspective, it is crucial to tune its integer constraints. Thus, we also provide computationally efficient, data-driven approaches to induce sparsity in the coefficients and the estimated amount of contamination. Theoretical properties characterizing our proposal include its high breakdown point, the *robustly strong oracle property* – which holds in ultra-high dimensional settings

where the number of predictors grows exponentially with the sample size – and optimality in parameter estimation with respect to the L_2-norm (i.e., optimal prediction error and risk-minimaxity). Our proposal requires weaker assumptions than prior methods in the literature and, unlike such methods, it allows the sparsity level and/or the amount of contamination to grow with the number of predictors and/or the sample size.

In addition to performing numerical experiments, we investigated the relationship between childhood obesity and the human microbiome. Our proposal generally outperformed existing heuristic methods in terms of predictive power, robustness and solution sparsity, and produced results consistent with prior childhood obesity studies.

The work presented here can be expanded in several directions. Even with modern solvers, larger problems and optimal tuning can make the use of MIPs computationally challenging. We are pursuing ways to reduce the computational burden – e.g., efficiently and effectively exploring the graph built by branch & bound algorithms (Gatu et al., 2007), extending the perspective formulation (Frangioni and Gentile, 2006) to the presence of MSOM outliers, and generating high-quality initial solutions for warm-starts and big-$\mathcal{M}$ bounds through continuous methods (Bertsimas and Mazumder, 2014). To improve solution quality, we are further exploring the addition of a ridge-like term, which would naturally benefit from the extension of the perspective formulation, as well as robust versions of whitening methods for feature de-correlation (Kenney et al., 2021). In our future research we also plan to explore the so-called cellwise contamination scheme (Alqallaf et al., 2009), which is a more recent approach for dealing with outliers in high-dimensional settings. Finally, we are particularly interested in the class of generalized linear models and Gaussian graphical models. The use of L_0 constraints for sparse estimation has already been investigated from a theoretical perspective (Shen et al., 2012), but an effective implementation in modern MIP solvers is not trivial and the possible presence of adversarial contamination has not received much attention in the literature.

"Everybody believes in the [normal] law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact"

Henri Poincaré (1912)

Chapter 4

Doubly Robust Feature Selection with Mean and Variance Outlier Detection and Oracle Properties

This chapter is based on: Insolia, L., Chiaromonte, F., Li, R., and Riani, M. (2021a). Doubly robust feature selection with mean and variance outlier detection and oracle properties. *arXiv preprint, arXiv:2106.11941*.

Reproducible and documented code for this chapter is available at: https://github.com/LucaIns/doubly_robust_sparse.

4.1 Introduction

Modern regression problems encompass an ever increasing number of predictor variables, or features – which motivates the use of feature selection techniques. In the real-world, these problems are often also affected by data contamination, e.g., due to recording errors or the presence of different sub-populations. Handling the resulting outliers is critical, as data contamination can hinder classical feature selection and estimation methods. Moreover, outlier detection itself can be a major goal of the analysis, as it often provides valuable domain-specific insights.

Two main contamination mechanisms have been investigated in the literature on linear models (Beckman and Cook, 1983), namely: the *mean-shift outlier model*

(MSOM) and the *variance-inflation outlier model* (VIOM), which are two specific characterizations of the Tukey-Huber contamination model. The MSOM assumes that outlying cases have a shift in mean; *maximum likelihood estimation* (MLE) leads to their removal from the fit – i.e., to the assignment of 0 weights to the cases identified as outliers. While the MSOM was traditionally studied in low-dimensional scenarios (Cook and Weisberg, 1982), it has been recently extended to high-dimensional linear models, where the use of regularization techniques is fundamental (She and Owen, 2011; Alfons et al., 2013; Kurnaz et al., 2017; Insolia et al., 2021d). The VIOM, which is historically considered as an alternative to the MSOM, assumes that contaminated errors have an inflated variance; outliers are retained but down-weighted in the fit. The VIOM was initially investigated by Cook et al. (1982) and Thompson (1985) in the presence of a single outlier, using MLE and *restricted MLE* (REMLE), respectively. More recently, Gumedze (2019) developed hypothesis testing procedures for linear models, considering also the presence of multiple outliers. However, when multiple outliers are present, this approach requires the evaluation of a combinatorial number of outlying-ness tests to avoid masking (undetected outlying cases) and swamping (non-outlying cases flagged as outliers). Insolia et al. (2021b) proposed the use of robust estimation and REMLE to detect and down-weight multiple VIOM outliers, possibly co-occurring with MSOM outliers, in (low-dimensional) linear models.

Here, we investigate high-dimensional linear models affected by the co-occurrence of multiple MSOM and VIOM outliers, where outlying cases can arise both in the response variable and the design matrix. We show that these can be modeled as additional fixed and random components, respectively, and evaluated independently. Specifically, we develop a doubly robust class of nonconcave penalization methods, in which feature selection and MSOM detection rely on a trimmed penalized loss, whereas VIOM detection is based on the penalization of the restricted posterior mode. The resulting procedure: (i) satisfies a robust oracle property for feature selection in the presence of data contamination, which allows the number of features to exponentially increase with the sample size; (ii) detects MSOM and VIOM outliers with asymptotic probability one; (iii) provides optimal units' weights, and thus

achieves an optimal trade-off between high breakdown point and efficiency. To our knowledge, both the theory behind the detection and treatment of multiple VIOM outliers and the study of high-dimensional models affected by VIOMs have not been developed to date. Our proposal bridges this gap and greatly improves estimation of the error variance, which is in turn essential for several statistical learning goals – e.g., in the construction of information criteria for variable selection and in inferential results on regression coefficients (Fan et al., 2012a; Reid et al., 2016). Recently, Insolia et al. (2021b) highlighted the fact that the co-occurrence of MSOM and VIOM can be modeled as an extended mixed-effects linear model, but this formulation was not leveraged in that contribution; on the other hand, such formulation is crucial for our developments in this paper, as it offers a completely different approach to tackle the problem. Moreover, unlike "soft" trimming estimators, which produce a general down-weighting for all points (Loh, 2017; Smucler and Yohai, 2017; Chang et al., 2018; Freue et al., 2019; Amato et al., 2021), our proposal is effective in estimating full weights for non-outlying observations. We also propose a data-driven procedure for VIOM detection. This provides a considerable gain with respect to existing soft trimming estimators that rely on a given nominal efficiency, and thus are not adaptive. Based on our numerical studies (see Section 4.4), being adaptive, our approach guarantees good performance consistently across contamination levels. Notably, our proposal comprises "hard" trimming sparse estimators as a special case – which assume (albeit often implicitly) only the presence of MSOM outliers. However, since we rely on nonconcave penalization methods, our approach satisfies oracle properties under weaker assumptions compared to existing hard trimming methods based on convex penalties (Alfons et al., 2013; Kurnaz et al., 2017). This creates an important bridge between the latter and L_0-constrained formulations that offer optimality guarantees at the expenses of a higher computational burden, and can be effectively leveraged by discrete optimization methods (Insolia et al., 2021d). Finally, we propose effective and computationally lean heuristic procedures that can be used as an alternative.

The reminder of the paper is organized as follows. Section 4.2 reviews relevant background literature. Section 4.3 details our proposal, which is a 3-step procedure,

as well as its heuristic counterpart. Section 4.4 contains numerical studies comparing the empirical properties of different methods both in low- and high-dimensional settings, and Section 4.5 contains real-world applications. Final remarks are given in Section 4.6. Further details, extensions and proofs, as well as the source code to replicate our simulation and application studies, are provided in Appendix B.

4.2 Background

In this section we review two streams of literature that are relevant for our developments; namely, methods for outlier detection in linear models, and approaches for feature selection in high-dimensional mixed-effects linear models.

4.2.1 Outlier detection

Consider a classical linear regression model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ contains observable responses, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times p}$ is the design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ contains unknown fixed effects (possibly sparse), and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ contains unobservable random errors. Classical assumptions specify that such errors are uncorrelated, homoscedastic and Gaussian, so that $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ for $0 < \sigma^2 < \infty$, where $\boldsymbol{I}_n$ is the identity matrix of size n.

The MSOM postulates that for outlying cases $i \in \mathcal{S}_\phi$ (the rationale for this symbol will become clear in Equation 4.2), $\varepsilon_i \sim N(\mu_{\varepsilon_i}, \sigma^2)$ with $\mu_{\varepsilon_i} \neq 0$. Under the assumption that $\mathcal{S}_\phi$ is known and $\text{rank}(\boldsymbol{X}) = p \leq n - |\mathcal{S}_\phi|$ (where $|\cdot|$ denotes the cardinality of a set), the MLE leads to the exclusion of the units in $\mathcal{S}_\phi$ from the fit (Cook and Weisberg, 1982). If there is a single MSOM outlier, this represents the unit with largest absolute Studentized residual, which is a monotone transformation of the deletion residual $t_i = (y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{(i)}) / \{\widehat{\sigma}_{(i)}(1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} \boldsymbol{x}_i)^{1/2}\}$, where the parenthetical subscript indicates the exclusion of unit i from the fit. Importantly, t_i can be computed very cheaply and, for a generic i, follows a Student's t with $n - p - 1$ degrees of freedom under the null – thus, it can be used to test the outlying-ness of each observation. Although this can be easily generalized to the presence of multiple MSOM outliers, it requires the evaluation of a combina-

torial number of fits (i.e., excluding all possible subsets of points of a given size from the fit), which results in a computationally intractable problem. Relatedly, high-breakdown estimators (see Section 4.3.1) aim at limiting the influence of extreme residuals on the fit (Maronna et al., 2006). Although these are traditionally computed using heuristic approaches, the use of *mixed-integer programming* (MIP) techniques has been recently considered to effectively solve the underlying combinatorial problem with optimality guarantees (Zioutas and Avramidis, 2005; Bertsimas and Mazumder, 2014). Importantly, high-breakdown point estimators have also been extended to sparse high-dimensional linear models in combination with penalization methods (Alfons et al., 2013; Smucler and Yohai, 2017; Kurnaz et al., 2017; Loh, 2017; Freue et al., 2019; Chang et al., 2018; Amato et al., 2021; She et al., 2021). Here L_0-constraints, which can be solved through MIP algorithms, provide optimality guarantees and desirable statistical properties for simultaneous feature selection and MSOM detection, with p allowed to increase exponentially with n (Insolia et al., 2021d).

The VIOM postulates that for outlying cases $i \in \mathcal{S}_\gamma$ (also this symbol will become clear in Equation 4.2), $\varepsilon_i \sim N(0, \sigma^2 v_i)$ with $v_i = (1 + \omega_i) \geq 1$. Cook et al. (1982) studied the presence of a single variance-inflated outlier; the MLE estimate of $\boldsymbol{\beta}$ depends on its v_i and results in a *weighted least squares* (WLS) fit $\widehat{\boldsymbol{\beta}}(v_i) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y} = \widetilde{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\widetilde{e}_i[(1-w_i)/\{1-(1-w_i)h_i\}]$, where $\boldsymbol{W}$ is a diagonal matrix containing all ones but $w_i = v_i^{-1}$. The tilde indicates quantities computed from the *ordinary least squares* (OLS) fit, and h_i is the i-th diagonal element of $\boldsymbol{H}_x = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. This highlights the fact that the VIOM is asymptotically equivalent to the MSOM as $v_i \to \infty$. Importantly, in the presence of a single VIOM outlier, the MLE provides a closed-form estimate for v_i, which can be used to estimate $\boldsymbol{\beta}$ and σ^2. Similarly, Thompson (1985) used REMLE in place of MLE to estimate the variance components v_i and σ^2. REMLE relies on $n - p$ linearly independent error contrasts $\boldsymbol{A}^T\boldsymbol{\varepsilon}$, where $\boldsymbol{A} \in \mathbb{R}^{n\times(n-p)}$ is defined such that $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_n$ and $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{P}_x$, with $\boldsymbol{P}_x = \boldsymbol{I}_n - \boldsymbol{H}_x$ (Patterson and Thompson, 1971). Also REMLE provides a closed-form estimate for the single variance-inflation parameter v_i. Notably, the single VIOM outlier position estimated by MLE and REMLE

might differ. A sufficient condition for their agreement is that the unit with maximum absolute OLS residual $\max(|\widetilde{e}_i|)$ also has the largest absolute Studentized residual $\max(|t_i|)$ – the latter estimates the outlier position using REMLE, which is equivalent to the outlier position estimated by MLE under an MSOM (Thompson, 1985). However, differently from the case of a single VIOM outlier (and of multiple MSOM outliers), multiple variance-inflation parameters $\boldsymbol{v}$ cannot be estimated in closed-form even if the outliers are known – thus, iterative procedures are required (Gumedze, 2019). In order to detect multiple VIOM outliers, possibly concurrent with MSOM outliers, Insolia et al. (2021b) proposed the use of robust estimation for outlier detection and of REMLE to estimate units' weights. Nevertheless, to the best of our knowledge, high-dimensional linear models affected by VIOM contamination have not been explored yet.

4.2.2 Feature selection for mixed-effects linear models

Mixed-effects linear models are often used to model data with a natural group structure, such as repeated measurements, measurements in time, and measurements in space (Laird and Ware, 1982). They extend the classical linear model through the inclusion of a random design matrix characterizing the experiment; namely, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}$, where $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t] \in \mathbb{R}^{n \times q}$, and $\boldsymbol{Z}_j \in \mathbb{R}^{n \times q_j}$ indicates the design matrix for the j-th random effect $\boldsymbol{b}_j \in \mathbb{R}^{q_j}$, such that $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_t^T)^T \in \mathbb{R}^q$, and $\sum_j q_j = q$. It is often assumed that $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{\mathcal{B}})$, where $\boldsymbol{\mathcal{B}} = [\boldsymbol{B}_1, \ldots, \boldsymbol{B}_t]$ is a block-diagonal matrix modeling the covariance of each random effect $\boldsymbol{b}_j \sim N(\boldsymbol{0}, \boldsymbol{B}_j)$, with $\text{cov}(\boldsymbol{b}_k, \boldsymbol{b}_l) = 0$ for any $k \neq l$. Moreover, $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$ are assumed to follow independent Gaussian distributions.

Several methods have been developed to simultaneously estimate fixed and random effects. Henderson's mixed-model equations lead to the *best linear unbiased estimator* (BLUE) for the fixed effects $\boldsymbol{\beta}$ and the *best linear unbiased predictor* (BLUP) for the random effects $\boldsymbol{b}$ – which is also known as the *empirical Bayes estimator* as it maximizes the posterior distribution $f(\boldsymbol{b}|\boldsymbol{y})$. However, this approach is unviable to perform feature selection in high-dimensional scenarios (Fan and Li, 2012). For this purpose, hypothesis testing procedures have been developed to select

relevant random effects (Lin, 1997). Different sub-models can be compared through extensions of information criteria, such as the *conditional Akaike information criterion* (CAIC; Liang et al. 2008) and its generalizations. Leveraging penalization methods, other approaches perform sparse estimation of the fixed effects $\boldsymbol{\beta}$. In these, while the dimension p of $\boldsymbol{\beta}$ is allowed to increase with the sample size n, the random component $\boldsymbol{b}$ is often assumed to contain only truly relevant random effects (Schelldorfer et al., 2011). Yet other approaches use penalization methods to select a given number of fixed and random effects (Bondell et al., 2010; Ibrahim et al., 2011; Peng and Lu, 2012). See Müller et al. (2013) and Buscemi and Plaia (2020) for a literature review.

In the following we focus on the class of nonconcave penalization methods introduced by Fan and Li (2012). Importantly, based on REMLE principles, selection of fixed and random effects can be performed independently. Under mild conditions this approach satisfies a weak oracle property for fixed effects estimates and selects truly relevant random effects with asymptotic probability one – where the dimensions p and q of fixed and random effects are allowed to exponentially increase with the sample size.

4.3 Our proposal

We investigate linear models affected by systematic (MSOM) and/or stochastic (VIOM) contaminations. Specifically, we focus on a general *unlabeled* outlier problem (Beckman and Cook, 1983), where the nature (MSOM vs. VIOM) as well as the identity, number and strength of the outliers is unknown. We model the presence of m_V VIOM and m_M MSOM outliers, indexed through the (unknown and non-overlapping) sets $\mathcal{S}_\gamma$ and $\mathcal{S}_\phi$:

$$\varepsilon_i \sim \begin{cases} N(0, \sigma^2 v_i) & \forall\ i \in \mathcal{S}_\gamma \\ N(\mu_{\varepsilon_i}, \sigma^2) & \forall\ i \in \mathcal{S}_\phi \\ N(0, \sigma^2) & \text{otherwise,} \end{cases} \tag{4.1}$$

where $v_i > 1$ and $\mu_{\varepsilon_i} \neq 0$. We exclude overlaps between the two types of contamination because such over-parametrization is equivalent to a MSOM assumption (Cook et al., 1982). Moreover, as customary in the robust statistics literature, we let MSOM outliers also affect the design matrix $\boldsymbol{X}$ (with shifts μ_{x_i}) creating leverage points (Maronna et al., 2006). Note that the MSOM in (4.1) leads to the removal of the largest (squared) residuals. Therefore, it effectively captures outlying points both in $\boldsymbol{y}$ and in $\boldsymbol{X}$ (Cook and Weisberg, 1982, p. 21). Moreover, as we remarked already in the Introduction, detecting and treating VIOM outliers is also very important as they affect the estimated error variance. This, in turn, plays a crucial role in several aspects of an analysis – e.g., information criteria for variable selection and inferential results on regression coefficients (Fan et al., 2012a; Reid et al., 2016).

Notably, the outliers in (4.1) can be equivalently represented adding fixed and random effects to the linear model (Insolia et al., 2021b). In symbols

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}_{\mathcal{S}_\gamma}\boldsymbol{\gamma} + \boldsymbol{D}_{\mathcal{S}_\phi}\boldsymbol{\phi} + \boldsymbol{\epsilon}, \tag{4.2}$$

where $\boldsymbol{D}_{\mathcal{S}_\gamma}$ $(n \times m_V)$ and $\boldsymbol{D}_{\mathcal{S}_\phi}$ $(n \times m_M)$ are matrices composed by dummy column vectors indexing VIOM and MSOM outliers, respectively. The $m_V \times 1$ random vector $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{\Gamma})$ allows one to down-weight VIOM outliers; here $\boldsymbol{\Gamma} = \text{diag}_{m_V}(\boldsymbol{\omega})$ is a diagonal matrix of size m_V. The non-stochastic vector $\boldsymbol{\phi} \in \mathbb{R}^{m_M}$ contains prediction residuals for MSOM outliers (i.e., their residuals based on an estimator which excludes them from the estimation process) and removes their influence from the fit. The associated t-statistics are the deletion residuals $t_{\mathcal{S}_\phi}$. The random error vector is assumed to be $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ and independent from $\boldsymbol{\gamma}$. If the sets of outliers $\mathcal{S}_\phi$ and $\mathcal{S}_\gamma$ are known, and $\text{rank}(\boldsymbol{X}) = p \leq n - m_M$, the formulation in (4.2) allows one to use standard techniques for mixed-effects linear models to estimate variance-inflation parameters $\boldsymbol{v}$ and regression coefficients $\boldsymbol{\beta}$. However, this approach is unfeasible if the outlier identities are unknown and/or if $p > n$. To tackle this problem, we consider the general formulation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{I}_n\boldsymbol{\gamma} + \boldsymbol{I}_n\boldsymbol{\phi} + \boldsymbol{\epsilon} \tag{4.3}$$

and rely on nonconcave penalization methods to select relevant fixed effects $\boldsymbol{\beta}$ – but we also enforce sparsity in $\boldsymbol{\gamma} \in \mathbb{R}^n$, which detects and down-weights VIOM outliers, and $\boldsymbol{\phi} \in \mathbb{R}^n$, which detects and excludes MSOM outliers from the fit. Specifically, we propose a 3-step procedure based on REMLE principles, that extends and combines the approaches in Fan and Li (2012) and Insolia et al. (2021d,b). Operationally, the three steps can be solved iteratively (see Section 4.4), and we first focus on fixed effects estimation, as MSOM outliers can have stronger influence on model estimates and we also address settings with $p > n$.

4.3.1 Step 1: feature selection and MSOM detection

Suppose that $\mathcal{S}_\gamma$ is known. Then, plugging the MLE estimates for $\boldsymbol{\gamma}|\boldsymbol{\beta}$ in the joint density distribution $f(\boldsymbol{y}, \boldsymbol{\gamma})$ leads to the profile log-likelihood:

$$l_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}) \propto \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\phi})^T \boldsymbol{P}_R (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\phi}), \tag{4.4}$$

which produces a WLS estimator as

$$\begin{aligned} \boldsymbol{P}_R &= (\boldsymbol{I}_n - \boldsymbol{B}_\gamma)^T (\boldsymbol{I}_n - \boldsymbol{B}_\gamma) + \boldsymbol{B}_\gamma^T \boldsymbol{D}_{\mathcal{S}_\gamma} \boldsymbol{\Gamma}^{-1} \boldsymbol{D}_{\mathcal{S}_\gamma}^T \boldsymbol{B}_\gamma \\ &= (\boldsymbol{I}_n + \boldsymbol{D}_{\mathcal{S}_\gamma} \boldsymbol{\Gamma} \boldsymbol{D}_{\mathcal{S}_\gamma}^T)^{-1} = \boldsymbol{W}, \end{aligned} \tag{4.5}$$

where $\boldsymbol{B}_\gamma = (\boldsymbol{I}_n + \boldsymbol{D}_{\mathcal{S}_\gamma} \boldsymbol{\Gamma}^{-1} \boldsymbol{D}_{\mathcal{S}_\gamma}^T)^{-1}$. We simultaneously select and estimate fixed effects $\boldsymbol{\beta}$, while detecting and discarding MSOM outliers from the fit, using a feasible and robustly penalized version of (4.4), where an integer constraint and a nonconcave penalty are used for MSOM outlier detection and feature selection, respectively. In symbols

$$\left[\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}\right] = \underset{\boldsymbol{\beta}, \boldsymbol{\phi}}{\arg\min}\ \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\phi})^T \boldsymbol{\mathcal{M}}_R (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\phi}) + (n - k_n) \sum_{j=1}^{p} R_\lambda(|\beta_j|) \tag{4.6}$$

$$\text{s.t. } \|\boldsymbol{\phi}\|_0 = \sum_{i=1}^{n} I(\phi_i \neq 0) \leq k_n, \tag{4.6a}$$

where $I(\cdot)$ is the indicator function, the matrix $\boldsymbol{\mathcal{M}}_R = (\boldsymbol{I}_n + \boldsymbol{\mathcal{M}}_\gamma)^{-1}$ is a proxy for the unknown $\boldsymbol{P}_R$, and $\boldsymbol{\mathcal{M}}_\gamma$ is a proxy for $\boldsymbol{D}_{\mathcal{S}_\gamma} \boldsymbol{\Gamma} \boldsymbol{D}_{\mathcal{S}_\gamma}^T$ (see Appendix B.2 for details). Note that if $\boldsymbol{\mathcal{M}}_R$ is a multiple of the identity matrix, then (4.6) neglects VIOM outliers – i.e., all points receive binary weights; see for instance Insolia et al. (2021d); She et al. (2021).

The penalty function $R_\lambda(\cdot)$ enforces sparsity in $\boldsymbol{\beta}$ estimates and depends on a tuning parameter λ controlling the trade-off between goodness of fit and model complexity. For this task, several penalties have been investigated in the literature. Tibshirani (1996) introduced the *lasso* based on the L_1-penalty, which is very efficient but provides biased estimates. To overcome this limitation, nonconcave penalties have also been used. These include the *smoothly clipped absolute deviation* (SCAD) (Fan and Li, 2001), the *minimax concave penalty* (MCP) (Zhang, 2010), and the *adaptive lasso* (Zou, 2006). Other approaches solve the combinatorial best subset selection problem using an L_0-constraint and MIP algorithms (Bertsimas et al., 2016; Kenney et al., 2021). In this work we focus on penalties satisfying the following conditions.

Conditions List 4.1 (Penalty function)**.** *For any $\lambda > 0$ and $t \in [0, \infty)$, the penalty $R_\lambda(t)$ is: non-decreasing and concave with $R_\lambda(0) = 0$; twice continuously differentiable with first derivative $R'_\lambda(0^+) > 0$; such that $\sup_{t>0} R''_\lambda(t) \to 0$ for $\lambda \to 0$.*

These conditions are fairly common for nonconcave penalization methods (see for instance Fan and Lv 2011), and are used to develop estimators with three desirable properties: unbiasedness, sparsity and continuity (Fan and Li, 2001). We specifically focus on the SCAD penalty $R_\lambda(\cdot)$ in (4.6), but others might be considered as well. The SCAD penalty satisfies $R_\lambda(0) = 0$ and, for $t \in (0, \infty)$, has $R'_\lambda(t) = \lambda I(t \leq \lambda) + [(a\lambda - t)/(a-1)] I(t > \lambda)$, where the constant $a > 2$ controls nonconcavity and is often set to $a = 3.7$. This folded-concave penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$ and singular at 0. Since its derivative is zero outside $[-a\lambda, a\lambda]$, it does not shrink and thus bias large coefficient estimates. Obtaining a global minimum with folded-concave penalties such as SCAD is non-trivial. In the following we focus on the *local linear approximation* (LLA) method (Zou and Li, 2008) to

obtain a local solution which guarantees oracle properties. However, in principle one can achieve the global minimum using MIP techniques (Liu et al., 2016).

The L_0-constraint in (4.6a) is used for MSOM outlier detection, and it is equivalent to a least trimmed squares (LTS) robust loss (Rousseeuw and Leroy, 1987). It depends on an integer tuning parameter $k_n \geq 0$ controlling the trimming level – i.e., the number of points which are identified as MSOMs and excluded from the fit. This guarantees the achievability of high-breakdown estimates (see below). Modern MIP solvers can be used to solve the formulation in (4.6) with optimality guarantees (Insolia et al., 2021d). However, in order to reduce the computational burden, one can also use well-established heuristics (Alfons et al., 2013; Kurnaz et al., 2017) that rely on the FAST-LTS algorithm (Rousseeuw and Van Driessen, 2006). Other choices may also be considered, such as iterative hard-thresholding and progressive iterative quantile-thresholding algorithms (She and Owen, 2011; She et al., 2021), as well as compressive sampling (Needell and Tropp, 2009).

Intuitively, the *breakdown point* (BdP) measures the largest fraction of contamination that an estimator can tolerate before it becomes arbitrarily biased (Donoho and Huber, 1983). The finite-sample replacement BdP is defined as $\varepsilon^*(\widehat{\boldsymbol{\beta}}, \boldsymbol{Z}) = \min(m/n : \sup_{\widetilde{\boldsymbol{Z}}} \|\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{Z}})\|_2 = \infty)$, where $\widetilde{\boldsymbol{Z}}$ denotes the original dataset $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{y})$ after the replacement of m out of n points with arbitrary values. The following result shows that our proposal achieves the highest possible BdP – i.e., a BdP that can be arbitrarily large due to the lack of regression equivariance (Smucler and Yohai, 2017).

Proposition 4.1 (High breakdown-point)**.** *For any $\lambda > 0$ and $a > 2$ the estimator $\widehat{\boldsymbol{\beta}}$ produced by* (4.6) *achieves a breakdown point of $\varepsilon^* = (k_n + 1)/n$.*

Thus, in the presence of MSOM contamination, our proposal breaks down only if $k_n < m_M$. Moreover, this result does not require that the points $(\boldsymbol{x}_i^T, y_i)$ are in general position. This is necessary for low-dimensional estimators to achieve equivariance (Maronna et al., 2006) – something that cannot be achieved by our proposal (Maronna, 2011). Nevertheless, as it was argued also in Alfons et al. (2013), we do not recommend setting $k_n > n/2$ for typical applications since the main goal of robust statistics is to model the "bulk" of the data. This can however

be beneficial in some instances, e.g., if one is interested in detecting heterogeneous populations and p is fairly small.

Notably, lasso estimation can be considered as the first iteration in computing the SCAD penalty based on the LLA method (Zou and Li, 2008). Thus, while SCAD provides stronger theoretical results for feature selection, one can perform MSOM outlier detection with existing robust algorithms based on lasso, e.g., the *sparseLTS* (Alfons et al., 2013) which solves a trimmed loss problem with an L_1-penalty using the FAST-LTS heuristic algorithm. Then, SCAD can be computed on the set of non-outlying cases detected by a robust lasso on the first iteration of LLA; this is the approach followed in our implementation described below. A comparison with MIP-based procedures is also of great interest but out of the scope of the present paper. We also remark that a convex relaxation of (4.6a), albeit computationally very efficient, would lead to the breakdown of the resulting estimates (She and Owen, 2011).

We remark that the notion of breakdown can be misleading for non-equivariant estimators, such as those produced through penalties (Maronna, 2011; Smucler and Yohai, 2017; Insolia et al., 2021d). Hence, we provide additional guarantees in terms of simultaneous MSOM outlier detection and feature selection. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\phi}_0^T)^T \in \mathbb{R}^{p+n}$ be the true parameter vector, and decompose it as $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{\mathcal{S}}^T, \boldsymbol{\theta}_{\mathcal{S}^c}^T)^T = \{(\boldsymbol{\beta}_{\mathcal{S}_\beta}^T, \boldsymbol{\phi}_{\mathcal{S}_\phi}^T), (\boldsymbol{\beta}_{\mathcal{S}_\beta^c}^T, \boldsymbol{\phi}_{\mathcal{S}_\phi^c}^T)\}^T$ where $\boldsymbol{\theta}_{\mathcal{S}}$ contains the p_0 non-zero coefficients belonging to $\mathcal{S}_\beta$, and the m_M outlying cases belonging to $\mathcal{S}_\phi$ (here $(\cdot)^c$ indicates the complement of a set). $\widehat{\boldsymbol{\theta}}_0$ represents a *fixed-effects robust oracle estimator*, behaving as if the true sets of active features and outliers were both known in advance. Let $\|\cdot\|_\infty$ indicate the matrix infinity norm, and $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ the minimum and maximum eigenvalue of a matrix, respectively. We rely on the following conditions to recover $\widehat{\boldsymbol{\theta}}_0$.

Conditions List 4.2 (Fixed-effects robust oracle reconstruction)**.**

A. *Minimum signal strength:* $s_1 n^\tau \{\log(n-m_M)\}^{-3/2} \to \infty$, *where* $s_1 = \min_{j\in\mathcal{S}_\beta} |\beta_{0,j}|$, $\tau \in (0, 1/2)$ *is a given constant, and* $\sup_{t\geq s_1/2} R''_\lambda(t) = o\left((n-m_M)^{-1+2\tau}\right)$.

B. *Design and proxy matrices: for some constants* $\eta \in (2\tau, 1]$ *and* $c_0 > 0$, *the matri-*

ces $(n-m_M)^{-1}(\mathbf{X}^T_{\mathcal{S}^c_\phi,\mathcal{S}_\beta}\mathbf{X}_{\mathcal{S}^c_\phi,\mathcal{S}_\beta})$ *and* $(n-m_M)^{\eta}(\mathbf{X}^T_{\mathcal{S}^c_\phi,\mathcal{S}_\beta}\mathbf{P}_R\mathbf{X}_{\mathcal{S}^c_\phi,\mathcal{S}_\beta})^{-1}$ *have minimum and maximum eigenvalues bounded from below and above by* c_0 *and* c_0^{-1}*, respectively. Moreover*

$$\left\|\left(\frac{1}{n-m_M}\mathbf{X}^T_{\mathcal{S}^c_\phi,\mathcal{S}_\beta}\boldsymbol{\mathcal{M}}_R\mathbf{X}_{\mathcal{S}^c_\phi,\mathcal{S}_\beta}\right)^{-1}\right\|_\infty \leq \frac{\{\log(n-m_M)\}^{3/4}}{(n-m_M)^{\tau}R'_\lambda\left(s_1/2\right)},$$

$$\left\|\mathbf{X}^T_{\mathcal{S}^c_\phi,\mathcal{S}^c_\beta}\boldsymbol{\mathcal{M}}_R\mathbf{X}_{\mathcal{S}^c_\phi,\mathcal{S}_\beta}\left(\mathbf{X}^T_{\mathcal{S}^c_\phi,\mathcal{S}_\beta}\boldsymbol{\mathcal{M}}_R\mathbf{X}_{\mathcal{S}^c_\phi,\mathcal{S}_\beta}\right)^{-1}\right\|_\infty < \frac{R'_\lambda(0+)}{R'_\lambda\left(s_1/2\right)}.$$

C. *Proxy matrix:* $\Lambda_{\min}\left(c_1\boldsymbol{\mathcal{M}}^{\mathcal{S}_\gamma}_\gamma - \boldsymbol{\Gamma}\right) \geq 0$ *and* $\Lambda_{\min}\left(c_1\log(n-m_M)\boldsymbol{\Gamma} - \boldsymbol{\mathcal{M}}^{\mathcal{S}_\gamma}_\gamma\right) \geq 0$ *for some constant* $c_1 > 0$*, and* $\boldsymbol{\mathcal{M}}^{\mathcal{S}^c_\gamma}_\gamma = \mathbf{0}_{n-m_V}$ *(the null square matrix of size* $n-m_V$*). Here* $\boldsymbol{\mathcal{M}}^{\mathcal{S}^c_\gamma}_\gamma$ *and* $\boldsymbol{\mathcal{M}}^{\mathcal{S}_\gamma}_\gamma$ *index rows and columns of the proxy matrix* $\boldsymbol{\mathcal{M}}_\gamma$ *corresponding to non-VIOMs and VIOMs, respectively.*

D. *MSOM strength:* $\Delta_\phi \geq d_\phi\sigma^2\log(n)/n$*, where* $d_\phi > 0$ *is a constant independent of* n *and* p*, and*

$$\Delta_\phi = \min_{\widehat{\boldsymbol{\phi}}_{\widetilde{\mathcal{S}}_\phi},\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}_\beta}} \frac{\|\boldsymbol{X}_{\widetilde{\mathcal{S}}_\beta}\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}_\beta} + \boldsymbol{I}_{n,\widetilde{\mathcal{S}}_\phi}\widehat{\boldsymbol{\phi}}_{\widetilde{\mathcal{S}}_\phi} - \boldsymbol{X}_{\mathcal{S}_\beta}\boldsymbol{\beta}_{\mathcal{S}_\beta} - \boldsymbol{I}_{n,\mathcal{S}_\phi}\boldsymbol{\phi}_{\mathcal{S}_\phi}\|_2^2}{n\max(|\mathcal{S}_\phi\backslash\widetilde{\mathcal{S}}_\phi| + |\mathcal{S}_\beta\backslash\widetilde{\mathcal{S}}_\beta|, 1)}$$

where $\widehat{\boldsymbol{\phi}}_{\widetilde{\mathcal{S}}_\phi}$ *is any estimate such that* $\widetilde{\mathcal{S}}_\phi \neq \mathcal{S}_\phi$ *and* $|\widetilde{\mathcal{S}}_\phi| \leq m_M$*, and* $\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}_\beta}$ *satisfies* $|\widetilde{\mathcal{S}}_\beta| \leq p_0$*.*

Conditions 4.2(A)-(C) are often used in the study of nonconcave penalization methods such as SCAD (Fan and Li, 2012), and they are based only on the set of non-outlying cases indexed by $\mathcal{S}^c_\phi$. Specifically, Condition 4.2(A) controls the minimum signal strength s_n, which is allowed to decay as $n-m_M$ increases. Condition 4.2(B) controls the design and proxy matrices, and ensures that one can limit the correlations between relevant and irrelevant variables. Condition 4.2(C) refers to the choice of the proxy matrix $\boldsymbol{\mathcal{M}}_\gamma$. It also implies that the VIOMs location is known, so that $\boldsymbol{\mathcal{M}}^{\mathcal{S}^c_\gamma}_R = \boldsymbol{I}_{n-m_V}$ and only VIOMs are down-weighted. Operationally, since the set $\mathcal{S}_\gamma$ is typically unknown, one can rely on the methods discussed in Section 4.3.2 for VIOM detection and iterate the process. Condition 4.2(D) is specifically required to detect MSOM outliers based on L_0-constraints (Insolia et al., 2021d). It bounds the difficulty of MSOMs detection based on a *minimal degree of separation* between the

true and a least favorable model. Intuitively it requires that, for models of comparable size, MSOM outliers have larger residuals than non-outlying cases. This relates to the signal-to-noise-ratio and it improves the heuristic argument $n > 5p$ which is often advocated for robust estimation methods (Rousseeuw and Van Zomeren, 1990). The following result ensures that our proposal achieves simultaneous feature selection and MSOM outlier detection consistency.

Proposition 4.2 (Robust weak oracle property)**.** *Under all conditions in lists 4.1 and 4.2,* $B(u_\phi, l_\phi) = \{\boldsymbol{\phi} : \|\boldsymbol{\phi}\|_0 \leq u_\phi, \Delta_\theta \geq l_\phi\}$*, and that* $\log p = o((n - m_M)\lambda^2)$ *and* $\sqrt{n - m_M}\lambda \to \infty$ *as* $(n - m_M) \to \infty$*, then, there exist* k_n *and a strict local minimiser of* (4.6) *such that the resulting robust estimates achieve:*

1. *Sparsity:* $P\left(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{S}}_\beta^c} = \mathbf{0}\right) \to 1$*;*
2. *Bounded* L_∞*-norm:* $P\left(\|\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{S}}_\beta} - \boldsymbol{\beta}_{\mathcal{S}_\beta}\|_\infty < (n - m_M)^{-\tau} \log(n - m_M)\right) \to 1$*;*
3. *MSOM consistency:* $\sup_{\phi_0 \in B(u_\phi, l_\phi)} P\left(\widehat{\mathcal{S}}_\phi \neq \mathcal{S}_\phi\right) \leq \sup_{\phi_0 \in B(u_\phi, l_\phi)} P\left(\widehat{\boldsymbol{\phi}} \neq \widehat{\boldsymbol{\phi}}_0\right) \to 0.$

Here the number of features in $\boldsymbol{\beta}$ is allowed to exponentially increase with the (uncontaminated) sample size $n - m_M$. This is a robust version of the weak oracle property in the sense of Lv and Fan (2009) and Fan and Li (2012). It demonstrates that our proposal asymptotically recovers the sets of truly relevant features and MSOM outliers, and the regression estimates are consistent under the L_∞ loss. Importantly, MSOM detection through L_0-constraints allows the number of non-zero components in $\boldsymbol{\phi}$ to be on the order of $\mathcal{O}(n)$; see for instance Shen et al. (2013); Insolia et al. (2021d). This is in contrast with existing methods that detect MSOMs through continuous penalties and require stronger conditions (Kong et al., 2018) since soft feature selection techniques typically require a sparsity level on the order of $\mathcal{O}(n^\alpha)$ for $\alpha < 1$ (Zhao and Yu, 2006).

We remark that existing robust model selection procedures based on trimming, which implicitly consider only MSOM outliers, can be cast into (4.3). However, differently from (4.6), they do not take into account the random structure of the problem, such as VIOM outliers. Relatedly, our approach can be naturally extended to high-dimensional mixed-effects linear models; however, this is left for future work.

Moreover, regardless the presence of VIOMs, the use of nonconcave penalties in (4.6) provides an important bridge between existing trimming estimators, which promote sparsity in the feature space based on convex penalties (Kurnaz et al., 2017; Alfons et al., 2013), and the optimal approach based on L_0-constraints (Insolia et al., 2021d). Unlike the former, our proposal achieves oracle properties under weaker assumptions, which can be particularly useful for the latter; e.g., to provide better warm-starts and big-$\mathcal{M}$ bounds, and thus accelerate convergence for MIP techniques.

4.3.2 Step 2: VIOM detection

VIOM outlier detection, based on sparse estimation of $\boldsymbol{\gamma}$ in (4.3), differs from sparse estimation of fixed effects ($\boldsymbol{\beta}$ and $\boldsymbol{\phi}$) due to their intrinsic randomness. Indeed, while underfitting $\boldsymbol{\gamma}$, which results in undetected VIOMs, introduces bias in the estimated variance for the fixed effects in $\boldsymbol{\beta}$, the inclusion of irrelevant $\boldsymbol{\gamma}$ components, i.e., wrongly detected VIOMs, decreases the estimator efficiency.

In this section, based on the results from Section 4.3.1, we consider the augmented design matrix $\overline{\boldsymbol{X}} = [\boldsymbol{X}_{\widehat{\mathcal{S}}_\beta}, \boldsymbol{D}_{\widehat{\mathcal{S}}_\phi}]$, where $\boldsymbol{X}_{\widehat{\mathcal{S}}_\beta}$ and $\boldsymbol{D}_{\widehat{\mathcal{S}}_\phi}$ index the estimated k_p active features and k_n MSOM outliers, respectively. We further assume that $n - k_n \geq k_p$, and that $\overline{\boldsymbol{X}}^T\overline{\boldsymbol{X}}$ is an invertible matrix of size $(k_p + k_n)$. The corresponding matrix of error contrasts is denoted as $\overline{\boldsymbol{A}}$, and $\boldsymbol{P}_{\overline{x}}$ is the counterpart of $\boldsymbol{P}_x$ using $\overline{\boldsymbol{X}}$ in place of $\boldsymbol{X}$.

Based on REMLE theory, the conditional distribution $f(\overline{\boldsymbol{A}}^T\boldsymbol{y}|\boldsymbol{\gamma}_{\mathcal{S}_\gamma})$ does not depend on $\boldsymbol{\beta}$, $\boldsymbol{\phi}$ and $\overline{\boldsymbol{A}}$, which leads to the restricted posterior density

$$\begin{aligned} f\left(\boldsymbol{\gamma}_{\mathcal{S}_\gamma}|\overline{\boldsymbol{A}}^T\boldsymbol{y}\right) &= f\left(\overline{\boldsymbol{A}}^T\boldsymbol{y}|\boldsymbol{\gamma}_{\mathcal{S}_\gamma}\right) f(\boldsymbol{\gamma}_{\mathcal{S}_\gamma}) \\ &= (\boldsymbol{y} - \boldsymbol{D}_{\mathcal{S}_\gamma}\boldsymbol{\gamma}_{\mathcal{S}_\gamma})^T\boldsymbol{P}_{\overline{x}}(\boldsymbol{y} - \boldsymbol{D}_{\mathcal{S}_\gamma}\boldsymbol{\gamma}_{\mathcal{S}_\gamma}) + \boldsymbol{\gamma}_{\mathcal{S}_\gamma}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_{\mathcal{S}_\gamma}. \end{aligned} \tag{4.7}$$

However, (4.7) cannot be used to estimate $\boldsymbol{\gamma}$ as it relies on the unknown set of VIOM outliers $\mathcal{S}_\gamma$, as well as their covariance matrix $\boldsymbol{\Gamma}$. We replace (4.7) with the following

objective function

$$\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \ (\boldsymbol{y}-\boldsymbol{\gamma})^T \boldsymbol{P}_{\overline{x}}(\boldsymbol{y}-\boldsymbol{\gamma}) + \boldsymbol{\gamma}^T \boldsymbol{\mathcal{M}}_{\gamma}^{-1}\boldsymbol{\gamma} + (n-k_n)\sum_{i\in\widehat{\mathcal{S}}_{\phi}^{c}} R_{\lambda}(|\gamma_i|) \tag{4.8}$$

where $\boldsymbol{\mathcal{M}}_{\gamma}$ is a proxy for $\boldsymbol{D}_{\mathcal{S}_\gamma}\boldsymbol{\Gamma}\boldsymbol{D}_{\mathcal{S}_\gamma}^T$ (see Appendix B.2 for details). In principle the penalty function $R_\lambda(\cdot)$ might differ from the one in (4.6), but for simplicity we consider nonconcave penalties such as SCAD also here.

In order to control the bias for the oracle-assisted estimator $\gamma_i^2/(n-m_M)$ of $\sigma^2\omega_i$, we condition on the event $\{\min_{i\in\mathcal{S}_\gamma}|\gamma_i| \geq \sqrt{n-m_M}b_0^*\}$, where $b_0^* \in (0, \min_{i\in\mathcal{S}_\gamma}\sigma\sqrt{\omega_i})$ and $\omega_i = \text{var}(\gamma_i)/\sigma^2$. Let $\mathbf{P}_{\overline{x}}^{\mathcal{S}_\gamma}$ comprise the rows and columns of $\mathbf{P}_{\overline{x}}$ belonging to the VIOM outliers in $\mathcal{S}_\gamma$. We rely on the following conditions to detect such outliers.

Conditions List 4.3 (VIOM reconstruction)**.**

A. *Design matrix and VIOM outliers: for some constant $c_3 > 0$, the minimum and maximum eigenvalues of $(n-m_M)^{-1}\mathbf{P}_{\overline{x}}^{\mathcal{S}_\gamma}$ and $\boldsymbol{\Gamma}$ are bounded from below and above, respectively, by c_3 and c_3^{-1}. Moreover, there exists $\delta \in (0, 1/2)$ such that*

$$\left\|(\mathbf{P}_{\overline{x}}^{\mathcal{S}_\gamma} + \boldsymbol{\Gamma}^{-1})^{-1}\right\|_{\infty} \leq \frac{(n-m_M)^{-(1+\delta)/2}}{R'_\lambda\left(\sqrt{n-m_M}b_0^*/2\right)},$$

$$\max_{i\in\mathcal{S}_\gamma^c\cap\mathcal{S}_\phi^c}\left\|\mathbf{P}_{\overline{x},i}\boldsymbol{D}_{\mathcal{S}_\gamma}(\mathbf{P}_{\overline{x}}^{\mathcal{S}_\gamma} + \boldsymbol{\Gamma}^{-1})^{-1}\right\|_2 < \frac{R'_\lambda(0+)}{R'_\lambda\left(\sqrt{n-m_M}b_0^*/2\right)}.$$

B. *VIOM strength:* $\sup_{\{t\geq\sqrt{n-m_M}b_0^*/2\}} R''_\lambda(t) = o\left((n-m_M)^{-1}\right)$.

C. *Proxy matrix:* $\Lambda_{\min}\left(\boldsymbol{\mathcal{M}}_{\gamma}^{\mathcal{S}_\gamma^c}\right) \geq 0$ *and* $\Lambda_{\min}\left(\boldsymbol{\mathcal{M}}_{\gamma}^{\mathcal{S}_\gamma} - \boldsymbol{\Gamma}\right) \geq 0$.

Similar conditions can be found in Fan and Li (2012) to perform feature selection on random effects using nonconcave penalties. The first part of Condition 4.3(A), which is not particularly restrictive, ensures that the reduced design matrix $\overline{\boldsymbol{X}}$ and the covariance matrix of the VIOM outliers $\boldsymbol{\Gamma}$ are well-behaved. The second part of the condition, controls the correlation between active and non-active features, as well as the effect of VIOMs, and depends on the penalty in use. Condition 4.3(B) is used to detect VIOMs with sizeable residuals, and is easily satisfied by nonconcave penalties such as SCAD for a suitable tuning parameter λ. Condition 4.3(C) limits

the choice of the proxy matrix $\boldsymbol{\mathcal{M}}_\gamma$ which is used to detect VIOM outliers; see again Appendix B.2 for details. The following result shows that our proposal is effective in dealing with VIOM outliers.

Proposition 4.3 (VIOM treatment)**.** *Under all conditions in lists 4.1 and 4.3, and assuming that* $b_0^*(n-m_M)^{\delta-1/2} \to \infty$ *as* $(n-m_M) \to \infty$*, there exists* λ *such that a strict local minimizer of* (4.8) *satisfies:*

1. *VIOM detection:* $P\left(\widehat{\mathcal{S}}_\gamma = \mathcal{S}_\gamma\right) \to 1$*;*
2. *VIOM down-weighting:* $\max_{i\in\mathcal{S}_\gamma}|\widehat{\gamma}_i - \gamma_i| \leq (n-m_M)^{-\delta}$.

Proposition 4.3 ensures that our proposal detects VIOM outliers with asymptotic probability one, and effectively down-weights them – meaning that the estimated weights converge to population weights.

4.3.3 Step 3: weights estimation

Steps 1 and 2 described above might induce non-negligible biases, especially in a finite-sample setting. To mitigate such biases, we propose an *ex-post* update for the VIOM outlier weights and other regression parameters depending on them. This is similar in spirit to post-selection updates implemented with feature selection methods; e.g., lasso followed by an OLS fit restricted to the set of active features (Liu and Yu, 2013).

Specifically, we consider a feasible counterpart of the mixed-effects linear model in (4.2), which is based on the estimated sets $\widehat{\mathcal{S}}_\phi$ and $\widehat{\mathcal{S}}_\gamma$ (MSOM and VIOM outliers), and $\widehat{\mathcal{S}}_\beta$ (active features). We first remove the units belonging to $\widehat{\mathcal{S}}_\phi$ from the fit, and apply REMLE to estimate weights for the units in $\widehat{\mathcal{S}}_\gamma$ conditionally on the features in $\widehat{\mathcal{S}}_\beta$. Next, we use these weights to update the estimates of $\boldsymbol{\beta}_{\widehat{\mathcal{S}}_\beta}$. This approach guarantees that, if Steps 1 and 2 identify the true model in terms of features ($\mathcal{S}_\beta$) as well as outliers ($\mathcal{S}_\phi$ and $\mathcal{S}_\gamma$), then our proposal reaches an optimal trade-off between breakdown point and efficiency.

The following definition extends the robustly strong oracle property in the sense of Insolia et al. (2021d) to the concurrent presence of MSOM and VIOM outliers – hence, we refer to it as the *doubly robust strong oracle property*.

Definition 4.1 (Doubly robust strong oracle property)**.** *Let* $\mathcal{S} = \{\mathcal{S}_\beta, \mathcal{S}_\phi, \mathcal{S}_\gamma\}$*, and define the doubly robust oracle estimator* $\widehat{\boldsymbol{\beta}}_S = \widehat{\boldsymbol{\beta}}|\mathcal{S}$ *as the solution for* $\boldsymbol{\beta}$ *in* (4.2)*. An estimator* $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{S}}}$ *satisfies the doubly robust strong oracle property if (asymptotically) there exist tuning parameters which ensure* $P(\widehat{\mathcal{S}} = \mathcal{S}) \geq P(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{S}}} = \widehat{\boldsymbol{\beta}}_S) \to 1$ *in the presence of MSOM and VIOM outliers.*

The following result refines Propositions 4.2 and 4.3, and ensures that our proposal achieves the doubly robust strong oracle property – allowing us to rely on large sample inference.

Theorem 4.1 (Doubly robust strong oracle property)**.** *Under all conditions in lists 4.1-4.3, as* $(n - m_M) \to \infty$ *there exist tuning parameters* k_n *and* λ*'s in* (4.6) *and* (4.8) *such that the resulting estimator plugging* $\widehat{\mathcal{S}}$ *in* (4.2) *achieves:*

1. *Asymptotic unbiasedness:*

$$\|E\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \leq 2P(\widehat{\mathcal{S}} \neq \mathcal{S}) \left\{ \|\boldsymbol{\beta}_0\|_2^2 + \lambda_M \left(\|\boldsymbol{W}^{1/2}\boldsymbol{X}\boldsymbol{\beta}_0\|_2^2 + \sigma^2 \operatorname{tr}(\boldsymbol{W}) \right) \right\} \to 0$$

   where $\operatorname{tr}(\cdot)$ *is the matrix trace,* $\lambda_M = \Lambda_{\max}\{(\boldsymbol{X}^T_{\widetilde{\mathcal{S}}_\beta}\boldsymbol{W}\boldsymbol{X}_{\widetilde{\mathcal{S}}_\beta})^+\} > 0$ *and* $\{\widetilde{\mathcal{S}}_\beta : \widehat{\mathcal{S}}_\beta \neq \mathcal{S}_\beta\}$*.*

2. *Optimal mean squared error (MSE):*

$$\begin{aligned} E\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \leq{}& \sigma^2 \operatorname{tr}(\boldsymbol{\Sigma}_X^{-1}) / \operatorname{tr}(\boldsymbol{W}) \\ &+ 2P(\widehat{\mathcal{S}} \neq \mathcal{S}) \left\{ (\lambda_M + \lambda_{M_s}) \left(\|\boldsymbol{W}^{1/2}\boldsymbol{X}\boldsymbol{\beta}_0\|_2^2 + \sigma^2 \operatorname{tr}(\boldsymbol{W}) \right) \right\} \end{aligned}$$

   where $\lambda_{M_s} = \Lambda_{\max}\{\boldsymbol{\Sigma}_X^{-1}\}$ *and* $\boldsymbol{\Sigma}_X = (\mathbf{X}^T_{\mathcal{S}_\beta}\boldsymbol{W}\mathbf{X}_{\mathcal{S}_\beta})$*.*

3. *Asymptotic normality:* $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to^d N(\mathbf{0}, \sigma^2(\boldsymbol{\Sigma}_X/n)^{-1})$*.*

Theorem 4.1 demonstrates that our proposal asymptotically behaves as if the sets of truly relevant features, MSOMs and VIOMs were jointly known in advance. This guarantees asymptotic unbiasedness and normality of the resulting regression estimates. Indeed, since the estimated unit weights recover the ones in (4.1), where only MSOMs are excluded from the fit and VIOMs are down-weighted, our proposal

provides an optimal trade-off between BdP and efficiency. Importantly, Theorem 4.1 provides also some intuition on the estimator's behavior when it does not retrieve the doubly robust oracle solution, as well as in finite-sample settings. Indeed, points 1 and 2 in Theorem 4.1 depend on the probability of not recovering the true model, in terms of active features and/or outlying cases – which increases estimation biases and MSE.

Finally, weights estimates obtained in Step 3 can be used to update the proxy matrices used in Sections 4.3.1 and 4.3.2, suggesting an iterative strategy whereby the process in Steps 1-3 is repeated improving model selection and estimation results (see Section 4.4). A similar approach was proposed in Fan and Li (2012) to select and estimate fixed and random effects; here our iteration includes an additional third step to update the weights. We remark that Steps 1 and 2 of our procedure require a careful tuning process, which is critical to estimate the set of active features, as well as weights in a data-driven fashion to guarantee their "adaptiveness" (i.e., the breakdown point and the efficiency of the corresponding $\boldsymbol{\beta}$ estimates). In Appendix B.2.3 we describe the robust *Bayesian information criteria* (BIC) proposed for this tuning.

4.3.4 A heuristic procedure

Here we present a computationally lean heuristic procedure similar to two-stage regression for mixed-models (Fahrmeir and Tutz, 1994), which is inspired by our main proposal; namely:

1. Solve (4.6) using the proxy matrix $\boldsymbol{\mathcal{M}}_R = \boldsymbol{I}_n$. Let $\boldsymbol{y}^* = \boldsymbol{y}_{\widehat{\mathcal{S}}_\phi^c}$ and $\boldsymbol{X}^* = \boldsymbol{X}_{\widehat{\mathcal{S}}_\phi^c, \widehat{\mathcal{S}}_\beta}$ comprise response and predictor values restricted to the selected relevant features and non-outlying cases.

2. Consider again (4.6) using $\boldsymbol{y}^*$, $\boldsymbol{X}^*$ and $\boldsymbol{\gamma}_{\widehat{\mathcal{S}}_\phi^c}$ in place of $\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{\phi}$, respectively. Using $\boldsymbol{\mathcal{M}}_R = \boldsymbol{I}_{n-k_n}$ and leaving the estimation of $\boldsymbol{\beta}$ unpenalized, solve the model relaxing the L_0-constraint (e.g., using SCAD or lasso). Let $\widehat{\boldsymbol{\gamma}}_{\widehat{\mathcal{S}}_\gamma}$ indicate the resulting sparse estimates.

3. Consider $\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and, similar to Section 4.3.3, estimate weights for the units $i \in \widehat{\mathcal{S}}_\gamma$ using REMLE and use WLS to update the estimation of $\boldsymbol{\beta}$.

Step 1 can be efficiently tackled using sparse high-breakdown point estimators based on heuristics. It detects MSOMs (i.e., it estimates non-zero entries in $\boldsymbol{\phi}$) and selects active features in $\boldsymbol{\beta}$. Step 2, which is related to ridge regression, is used to detect VIOMs. This is equivalent to assuming a MSOM if the active $\boldsymbol{\gamma}$ coefficients are not shrunk (e.g., using L_0-constraints these units receive zero weights). Otherwise units are down-weighted or left with their full weights; we follow this approach as MSOMs are detected in Step 1. Step 3, which might be skipped if one is only interested in $\boldsymbol{\beta}$, is useful to reduce possible biases introduced in Steps 1-2, and in principle might be combined with Step 2.

In Appendix B.2 we describe tuning strategies for Steps 1 and 2 of our heuristic procedure, and discuss its connections with ridge and M-estimation.

4.4 Simulation study

In this section we compare our proposal with state-of-the-art methods through numerical simulations. The data is generated as follows. Each row of the $n \times p$ design matrix $\boldsymbol{X}$ contains a 1 (for the intercept), and then entries drawn independently from a $N(\boldsymbol{0}, \boldsymbol{\Sigma}_{p-1})$. The p-dimensional coefficient vector $\boldsymbol{\beta}$ contains p_0 non-zero entries (including the intercept), and the errors ε_i are drawn independently from a $N(0, \sigma^2_{\text{SNR}})$. σ^2_{SNR} depends on the signal-to-noise-ratio $\text{SNR} = \text{var}(\boldsymbol{X}\boldsymbol{\beta})/\sigma^2_{\text{SNR}}$ and controls the difficulty of the problem. Then, m_V and m_M points out of n are contaminated as in (4.1). Mean shifts affect error and active predictors in the design matrix, with strengths μ_ε and μ_x, respectively. Variance inflation affects only the error, with a common parameter v. Each simulation scenario is replicated t times and results are averaged.

We consider the following performance metrics: **(i)** MSE of $\widehat{\boldsymbol{\beta}}$ partitioned into variance and squared bias. For each estimated coefficient

$$\text{MSE}(\widehat{\beta}_j) = \frac{1}{t}\sum_{i=1}^{t}(\widehat{\beta}_{ij} - \beta_j)^2 = \frac{1}{t}\sum_{i=1}^{t}(\widehat{\beta}_{ij} - \overline{\beta}_j)^2 + (\overline{\beta}_j - \beta_j)^2, \tag{4.9}$$

where $\overline{\beta}_j = \frac{1}{t}\sum_{i=1}^{t}\widehat{\beta}_{ij}$, and we average the MSE across coefficients to produce $\text{MSE}(\widehat{\boldsymbol{\beta}}) = \frac{1}{p}\sum_{j=1}^{p}\text{MSE}(\widehat{\beta}_j)$. **(ii)** We also consider the MSE of a weighted estimate of the error variance

$$\widehat{s}^2 = \frac{\sum_{i=1}^{n}\widehat{w}_i e_i^2}{n - \widehat{p}_0},$$

where the e_i's are raw residuals, the $\widehat{w}_i$'s are estimated weights, and $\widehat{p}_0$ is the number of selected features. This takes into account weight estimates regardless of whether some units are in fact contaminated. The MSE decomposition for $\widehat{s}^2$ is computed as in (4.9), with $\sigma^2_{\text{SNR}} - \widehat{s}^2$ and 0 replacing $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, respectively, since σ^2_{SNR} varies at each iteration. **(iii)** Let the non-zero entries of $\boldsymbol{\tau} = \boldsymbol{\phi} + \boldsymbol{\gamma}$ indicate MSOMs and/or VIOMs. Outlier detection accuracy is measured in terms of false positive and false negative rates

$$\text{FPR}(\widehat{\boldsymbol{\tau}}) = \frac{|\{i \in \{1,\dots,n\} : \widehat{\tau}_i \neq 0 \wedge \tau_i = 0\}|}{|\{i \in \{1,\dots,n\} : \tau_i = 0\}|}, \tag{4.10}$$

$$\text{FNR}(\widehat{\boldsymbol{\tau}}) = \frac{|\{i \in \{1,\dots,n\} : \widehat{\tau}_i = 0 \wedge \tau_i \neq 0\}|}{|\{i \in \{1,\dots,n\} : \tau_i \neq 0\}|}. \tag{4.11}$$

These indicate the proportion of uncontaminated units wrongly detected as outliers, and of undetected contaminated units, respectively. **(iv)** For sparse settings, we also consider feature selection accuracy – which is measured in terms of FPR and FNR as in (4.10) and (4.11), using β_j and $\widehat{\beta}_j$ (for $j = 1,\dots,p$) in place of τ_i and $\widehat{\tau}_i$, respectively.

4.4.1 Scenario 1: low-dimensional VIOMs

Here we set $p = p_0 = 2$, with $\boldsymbol{\beta} = (2,2)^T$ and SNR = 3. The proportion of VIOM outliers is $m_V/n = 0.25$ and $v = 10$. The sample size n increases from 50 to 500 with 10 equispaced values. Data for each setting are replicated $t = 100$ times.

We consider the *oracle benchmark* (Opt), i.e., a WLS fit based on the true population weights $\boldsymbol{w}$, along with: **(a)** OLS, the ordinary least squares estimator **(b)** LTS, the least trimmed squares estimator with trimming set to the true m_V/n (Rousseeuw and Leroy, 1987); **(c)** MM85, an MM-estimator using a preliminary S-estimator with 50% BdP and Tukey's bisquare loss function, with tuning constant

set to achieve 85% nominal efficiency (Maronna et al., 2006); **(d)** MM95, as in (c), with 95% nominal efficiency; **(e)** FSRws, which uses a variant of forward search and single REMLE weights as described in Insolia et al. (2021b); **(f)** Heur, our heuristic procedure (Section 4.3.4), where in Step 2 $\boldsymbol{\gamma}$ is estimated through a SCAD penalty, and in Step 3 each weight is estimated independently using REMLE as in FSRws; **(g)** SCADws, our main proposal (Section 4.3), where in Step 3 weights are estimated by a REMLE fit on the active random components of $\boldsymbol{\gamma}$ detected by SCAD – as in FSRws and Heur, these weights are estimated independently. See Appendix B.3 for further details on algorithmic implementations.

Figure 4.1 shows the MSE for $\widehat{\boldsymbol{\beta}}$; SCADws, Heur, MM85 and MM95 perform comparably and outperform other methods, FSRws improves on LTS and OLS (which perform poorly across sample sizes). Figure 4.2 shows the MSE for $\widehat{s}^2$. SCADws outperforms other methods in terms of bias, and its MSE is lowest as n increases. Unlike the oracle estimator, SCADws is capable of estimating full weights for VIOM outliers that do not carry sizeable residuals. Relatedly, non-outlying cases with large residuals by chance are given full weight by the oracle estimator, but not necessarily by SCADws (see circled dots on the right panel of Figure 4.3). We note also that in a very few instances, especially the ones with smaller sample sizes, SCADws exhibits larger variability. For comparison we compute a trimmed version of the MSE in Figure B.2 of Appendix B.3; eliminating these extreme instances, SCADws shows much stronger performance in terms of trimmed MSE. Heur performs comparably to SCADws, although its estimates have larger variability, and it outperforms LTS and OLS, which provide strongly biased estimates because each point receives a binary or full weight. The performance of FSRws decreases for smaller sample sizes, where outliers are more often undetected. MM85 outperforms MM95 at this contamination level, but they both have strong biases across sample sizes, highlighting the drawbacks of M-estimators with pre-specified efficiency values.

The two left panels of Figure 4.3 show FPR and FNR for VIOM detection across methods, respectively. Overall, SCADws outperforms other methods; its decrease in terms of FPR along sample sizes is partially compensated by an increase in FNR. FSRws is close to SCADws for larger sample sizes, but for smaller ones it fails to

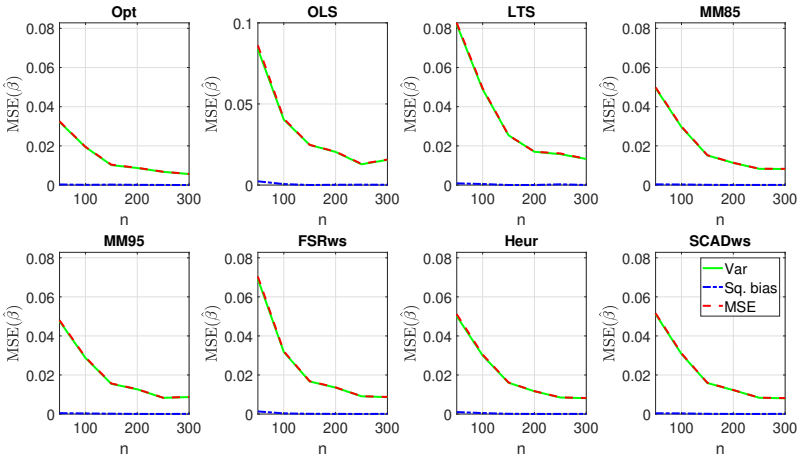


Figure 4.1: Scenario 1. MSE($\widehat{\boldsymbol{\beta}}$) comparisons across procedures and sample sizes.

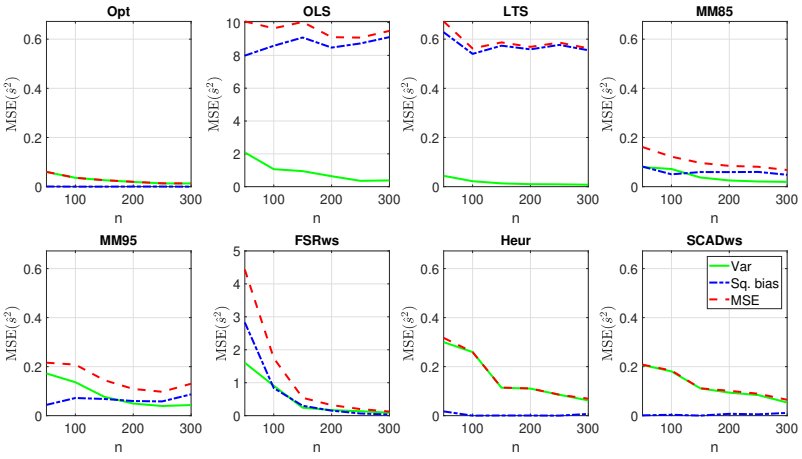


Figure 4.2: Scenario 1. MSE($\widehat{s}^2$) comparisons across procedures and sample sizes.

detect some outliers (low FPR and high FNR). Heur performs similarly to SCADws, and MM-estimators perform poorly in these metrics due to a general down-weighting of all units. These trends demonstrate the ability of SCADws to detect truly outlying cases as the sample size increases. On the other hand, while FSRws tends to be more conservative across sample sizes, LTS has a more aggressive behavior resulting in larger FPR and lower FNR. The right panel of Figure 4.3 shows a scatterplot summarizing results for a typical simulation ($n = 500$). True VIOM outliers, as well as the ones detected by SCADws, are highlighted.

The box-plots in Figure 4.4 show estimation accuracy across different methods in terms of $\frac{1}{p}\sum_{i=1}^{p}(\widehat{\beta}_j - \beta_j)^2$ (top panels) and $\sigma^2_{\text{SNR}} - \widehat{s}^2$ (bottom panels) as we fix $n = 100$ and increase the contamination level m_V/n from 0 to 0.3 with a step size

of 0.1 (from left to right). Here we performed $t = 200$ independent replications for each setting. All methods perform comparably in terms of estimation accuracy for $\widehat{\boldsymbol{\beta}}$, with SCADws, Heur and MM-estimators reporting a lower variability in contaminated settings. Focusing on $\widehat{s}^2$, SCADws outperforms other methods in all settings with VIOMs in terms of bias, as its median is always very close to 0 (dashed red line) and has a moderate variability. This highlights the effectiveness of its adaptive weights, which can accommodate various levels of contamination in a data-driven fashion. Importantly, also in the absence of VIOMs, SCADws performs comparably to non-robust methods (although it has slightly higher bias and dispersion). Heur performs comparably to SCADws in most settings, but has larger biases and variability, and other methods suffer as the fraction of contamination varies. In particular, MM85 and MM95 report a small variability across different settings, but they have very large biases in most instances – they achieve small biases only at 30 and 20% contamination, respectively. This highlights again the drawbacks of M-estimators with pre-specified nominal efficiency.

Additional simulation results with different SNR regimes are reported in Appendix B.3, and they are consistent with the ones discussed above.

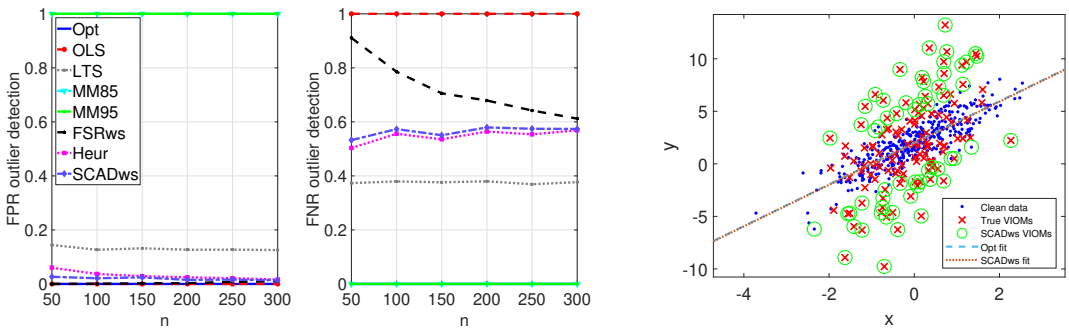


Figure 4.3: Scenario 1. Left: comparisons of FPR and FNR for outlier detection across procedures and sample sizes. Right: scatterplot summarizing results for a typical simulation with $n = 500$ – true VIOMs and VIOMs detected by SCADws are highlighted.

4.4.2 Scenario 2: high-dimensional VIOMs and MSOMs

Here we mimic Scenario 1, but we use sparse fixed effects in $\boldsymbol{\beta}$ and introduce MSOM outliers. Specifically, we set $p = 150$ with $\boldsymbol{\Sigma}_{p-1} = \boldsymbol{I}_{p-1}$, $p_0 = 4$, and $n = 100, 150, 200$. The proportions of VIOM and MSOM outliers are set to $m_V/n = 0.25$

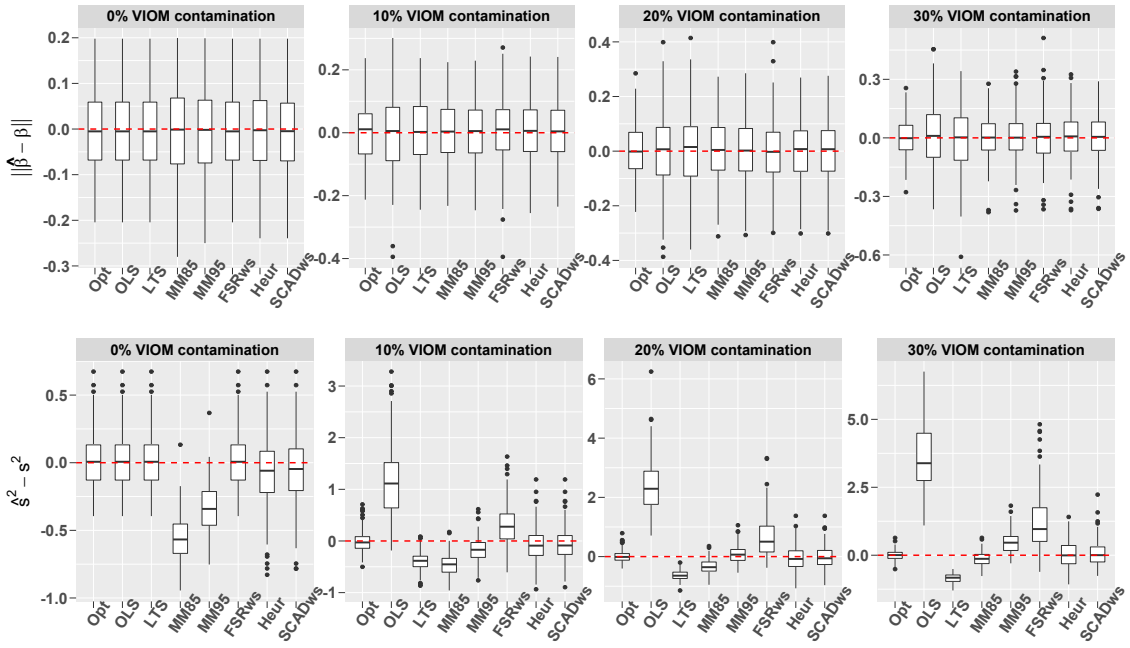


Figure 4.4: Scenario 1. Estimation accuracy across different methods for $\boldsymbol{\beta}$ (top panels) and σ^2_{SNR} (bottom panels) as the contamination level m_V/n increases (from left to right).

and $m_M/n = 0.05$, respectively, with mean shifts $\mu_\varepsilon = -10$ and $\mu_x = 10$ in order to create bad leverage points. Data for each setting are again replicated $t = 100$ times.

The oracle benchmark (Opt) is computed using population weights and the active feature set. In addition to it, we consider: **(a)** lasso (Tibshirani, 1996); **(b)** sparseLTS with 5% trimming (Alfons et al., 2013); **(c)** TaL, adaptive lasso with Tukey's bisquare loss, a preliminary sparseLTS fit, and tuning constant fixed to achieve 85% nominal efficiency (Chang et al., 2018); **(d)** Heur, as in Scenario 1, but with a preliminary fixed-effects selection and MSOM detection using robust SCAD. **(f)** SCADws, as in Scenario 1, but with a preliminary fixed-effects selection and MSOM detection based on (4.6); **(g)** SCAD2s, two iterations of SCADws where weights estimated in the first iteration are used to update the proxy matrices and re-run our 3-step procedure; **(h)** SCADopt, similar to SCADws, but with proxy matrices built with VIOM population weights. For simplicity, robust methods all use the true trimming level m_M/n. Further details about algorithmic implementation and the choice of tuning parameters are provided in Appendix B.3.

Table 4.1 provides MSE decompositions for $\widehat{\boldsymbol{\beta}}$ and $\widehat{s}^2$, and mean (with standard deviations in parenthesis) FPR and FNR for feature selection and outlier detection.

Table 4.1: Scenario 2. MSE for $\widehat{\boldsymbol{\beta}}$ and $\widehat{s}^2$ (decomposed into squared bias and variance), and mean (SD in parenthesis) FPR and FNR for feature selection and outlier detection, based on 100 simulation replications.

n	p	Method	bias$(\widehat{\beta})^2$	var$(\widehat{\beta})$	bias$(\widehat{s}^2)^2$	var$(\widehat{s}^2)$	FPR$(\widehat{\beta})$	FNR$(\widehat{\beta})$	FPR$(\widehat{\tau})$	FNR$(\widehat{\tau})$
100	150	Opt	0.00000	0.00081	0.01	0.11	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.08391	0.00497	433.78	18.73	0.03(0.05)	0.57(0.20)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.00216	0.02984	47.99	18.25	0.41(0.04)	0.00(0.05)	0.00(0.00)	0.83(0.01)
		TaL	0.00100	0.00545	2.39	2.24	0.02(0.04)	0.00(0.05)	1.00(0.00)	0.00(0.00)
		Heur	0.00019	0.01780	0.92	2.04	0.06(0.04)	0.00(0.05)	0.01(0.02)	0.69(0.10)
		SCADopt	0.00001	0.00194	0.00	0.49	0.00(0.00)	0.01(0.08)	0.01(0.05)	0.45(0.11)
		SCADws	0.00014	0.00565	0.06	1.81	0.00(0.01)	0.02(0.08)	0.03(0.04)	0.48(0.13)
		SCAD2s	0.00001	0.00220	0.00	0.62	0.00(0.00)	0.01(0.08)	0.03(0.06)	0.47(0.12)
150	150	Opt	0.00000	0.00062	0.02	0.07	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.08497	0.00312	423.81	13.33	0.03(0.04)	0.52(0.16)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.00085	0.01909	35.73	6.13	0.47(0.05)	0.00(0.00)	0.00(0.00)	0.84(0.00)
		TaL	0.00028	0.00173	1.58	0.90	0.01(0.01)	0.00(0.00)	1.00(0.00)	0.00(0.00)
		Heur	0.00009	0.00798	0.35	0.88	0.04(0.04)	0.00(0.00)	0.01(0.02)	0.61(0.12)
		SCADopt	0.00000	0.00078	0.02	0.23	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.47(0.09)
		SCADws	0.00000	0.00137	0.02	0.32	0.00(0.00)	0.00(0.02)	0.03(0.03)	0.46(0.09)
		SCAD2s	0.00001	0.00092	0.03	0.30	0.00(0.00)	0.00(0.00)	0.03(0.03)	0.46(0.09)
200	150	Opt	0.00000	0.00045	0.02	0.07	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.08608	0.00240	448.48	11.23	0.03(0.04)	0.43(0.17)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.00063	0.01536	32.44	4.44	0.51(0.05)	0.00(0.00)	0.00(0.00)	0.83(0.00)
		TaL	0.00019	0.00107	1.26	0.72	0.01(0.01)	0.00(0.00)	1.00(0.00)	0.00(0.00)
		Heur	0.00005	0.00515	0.14	0.80	0.04(0.04)	0.00(0.00)	0.01(0.02)	0.56(0.11)
		SCADopt	0.00000	0.00059	0.02	0.18	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.47(0.07)
		SCADws	0.00000	0.00074	0.00	0.22	0.00(0.00)	0.00(0.00)	0.02(0.02)	0.47(0.08)
		SCAD2s	0.00000	0.00067	0.00	0.24	0.00(0.00)	0.00(0.00)	0.02(0.02)	0.48(0.08)

In terms of MSE for $\widehat{\boldsymbol{\beta}}$, as expected, SCADopt resembles very closely the oracle estimator. SCAD2s, which improves upon SCADws, outperforms other feasible estimation methods. TaL has higher biases and variances, and Heur improves upon sparseLTS. Lasso breaks down due to the presence of MSOM outliers. In terms of MSE for $\widehat{s}^2$, SCADopt converges faster to the oracle solution, SCAD2s improves again upon SCADws, and they outperform competing methods. Heur outperforms TaL, but they have higher biases and variances compared to SCADws and SCAD2s, which are even larger for sparseLTS. Lasso provides very poor estimates due to the presence of outliers.

In terms of FPR and FNR for feature selection, SCADopt behaves as the oracle estimator, SCAD2s, which improves upon SCADws, generally outperforms other methods. TaL produces higher FPR across sample sizes, and Heur provides denser solutions – but still sparser than sparseLTS. Lasso performs poorly also here (high number of false negatives), since it breaks down. We note that most robust methods are at times affected by MSOMs for smaller sample size (larger FNR and MSE) where their detection is harder.

Focusing on the FPR and FNR for outlier detection, unlike the oracle estimator, SCADopt is capable of estimating full weights for VIOMs with negligible residuals (higher FNR), and it is not prone to detecting non-outliers with large residuals by chance (very low FPR). Notably though, although weights need to be estimated, also SCADws and SCAD2s perform well in both these metrics. For smaller sample sizes, SCAD2s reduces FNR compared to SCADws, which results is an overall performance increase for the iterative approach. Heur provides larger FNR and smaller FPR. SparseLTS has FPR equal to 0 and large FNR, as it detects only extreme MSOM outliers. TaL performs poorly due to a general down-weighting of all points, and lasso performs poorly because it assigns full weight to each observation.

Additional simulation results with larger sparsity level, weaker SNR and collinear features, where our proposal generally outperformed others in most settings, are provided in Appendix B.3.

4.5 Real-data examples

4.5.1 An Application to Boston housing data

The Boston housing dataset (http://lib.stat.cmu.edu/datasets/boston) contains $n = 506$ housing location and 13 predictors; namely: 1. *crim* (the per capita crime rate), 2. *zn* (the proportion of residential land zoned for lots over 25,000 sq.ft), 3. *indus* (the proportion of non-retail business acres), 4. *chas* (a "Charles River" dummy), 5. *nox* (the nitrogen oxides concentration in parts per 10 million), 6. *rm* (the average number of rooms per dwelling), 7. *age* (the proportion of owner-occupied units built prior to 1940), 8. *dis* (a weighted mean distance to five Boston employment centers), 9. *rad* (an index of accessibility to radial highways), 10. *tax* (the full-value property-tax rate per $10,000), 11. *ptratio* (the pupil-teacher ratio), 12. *black* (1000(B_k - 0.63)2, where B_k is the proportion of African-American residents), and 13. *lstat* (the percentage of the population in lower socioeconomic status). These are used to explain *medv*, the median value of owner-occupied homes in thousand dollars.

Using all predictors plus an intercept, we applied the LTS estimator with decreasing trimming and computed the robust BIC in Riani et al. (2022). This helps identify a reasonable trimming level to use across different methods. The left panel of Figure 4.5 shows that the curve flattens for low levels, with a noticeable drop only for very small amounts of trimming. Thus, using a 2% trimming, we used SCAD2s to select the relevant features on the full dataset. These are the predictors number 6, 8, 10, 11, 12, 13 (plus the intercept). The central panel of Figure 4.5 shows the robust BIC recomputed on these features alone. Compared to the left panel, it shows stronger evidence of both MSOM outliers (the curve achieves a maximum around 2% trimming) and VIOM outliers (the curve flattens starting from 10-5%). The right panel of Figure 4.5 shows robust residuals obtained by SCAD2s on the full dataset, where cases detected as MSOM and VIOM outliers are highlighted; they are 10 and 26, respectively. See Table B.7 in Appendix B.4 for a comparison with other methods.

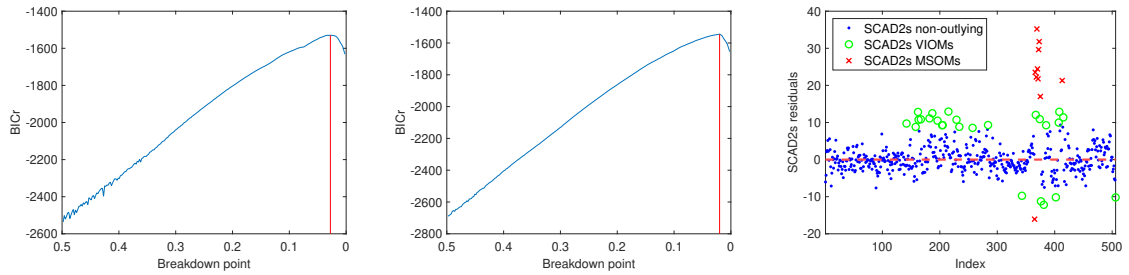


Figure 4.5: Boston housing data results. Left: robust BIC computed with the LTS on all points and features. Center: robust BIC computed on all points and only the features selected using SCAD2s. Right: SCAD2s residuals labeled as non-outlying (blue), MSOM (red), and VIOM (green).

Next, we extended the analysis along lines similar to Chang et al. (2018). We considered 50 random splits of the data in training and testing sets (300 and 206 units, respectively). Based on the observations above we used again 2% trimming across hard trimming methods. The left panel of Figure 4.6 shows box-plots of the sparsity levels, i.e., the number of features retained by different methods, across the 50 random training sets. Some methods do not provide sparse estimates by definition, but also lasso and our heuristic proposal provide very dense solutions. TaL and sparseLTS provide denser solutions compared to SCAD2s and SCADws,

and SCAD2s reports higher variability than SCADws. The right panel of Figure 4.6 shows the distribution of the selected features across the 50 random training sets. The solution for SCAD2s is in line with our prior analyses on the full data – predictors number 6, 8, 10, 11, 12, 13 are selected most of the times. TaL provides similar results, but supports the relevance of predictors number 1 and 4, and selects 10 on a very few replications.

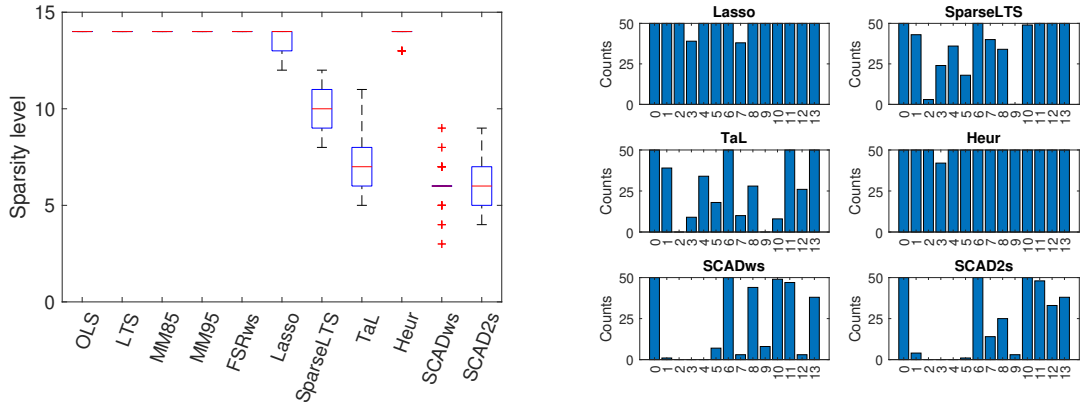


Figure 4.6: Box-plots of the estimated sparsity levels (left) and distribution of the selected features for sparse methods (right) across 50 random training sets for different methods on Boston housing data.

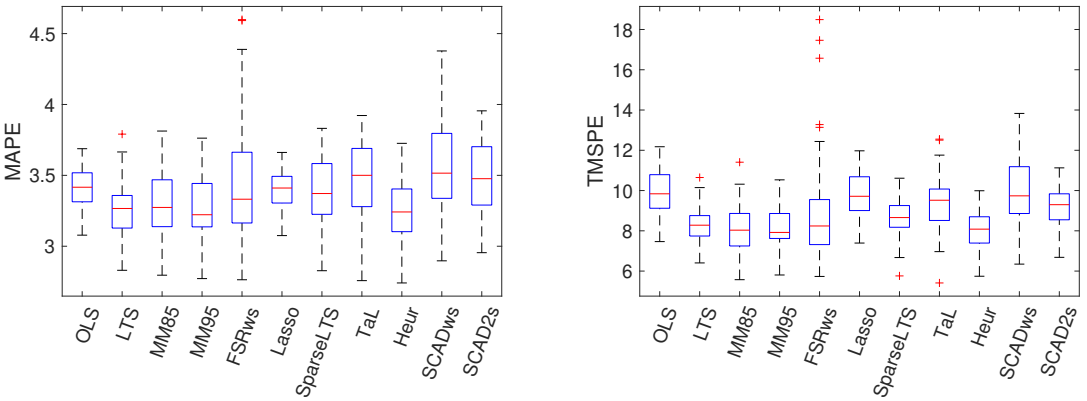


Figure 4.7: Box-plots of MAPE (left) and TMSPE (right) across 50 random training/testing splits for different methods on Boston housing data.

Figure 4.7 compares the prediction accuracy of different methods across the 50 random training/testing splits based on the mean absolute (MAPE, left panel) and trimmed mean squared (TMSPE, right panel) prediction errors, with an upper 10% trimming. SCAD2s improves upon SCADws and provides a good trade-off between model parsimony and prediction accuracy. Both in terms of MAPE and TMSPE,

Table 4.2: Comparison across different methods for glioblastoma gene expression data in terms of sparsity level, and MAPE and TMSPE computed on testing data.

Method	$\widehat{p}_0$	MAPE	TMSPE
Lasso	2	0.7405	0.5111
SparseLTS	25	0.6682	0.3959
TaL	4	0.6813	0.4309
Heur	7	0.7947	0.5641
SCADws	4	0.6323	0.3442
SCAD2s	3	0.6197	0.3360

SCAD2s performs comparably to other methods with denser solutions, and at times even better, but it provides more interpretable results. Our heuristic procedure performs very well in terms of prediction – often better than non-sparse robust estimators – but it has very dense solutions.

4.5.2 An application to glioblastoma gene expression data

Glioblastoma is the most frequent primary malignant brain tumor in the adult population, and one of the most lethal forms of cancer. Glioblastoma microarray gene expression data (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/ASPMgene) were collected through high-density Affymetrix arrays by Horvath et al. (2006). They include two different sets of clinical tumor samples, where the number of patients is 55 and 65 and, for each sample, recordings of 3,600 genes expression values.

Wang et al. (2011) and Chang et al. (2018) used feature selection techniques and their robust counterparts to analyze these data. Following their modeling strategy, we used the set with $n_{\text{tr}} = 55$ samples as training set, the one with $n_{\text{te}} = 61$ samples as test set, and took the logarithm of time to death as response variable. Censored observations (i.e., patients that were alive at the last followup) were removed at the outset, bringing the sample sizes to $n_{\text{tr}} = 50$ and $n_{\text{te}} = 55$. We further log-transformed all 3,600 predictors (gene expression values) to mitigate skews, and if the absolute pairwise Pearson correlation between two predictors in the training set exceeded the cut-off of 0.75, then the variable with higher mean absolute correlation across all remaining features was removed from the modeling exercise at the outset – bringing the number of predictors down to $p = 570$ (plus the intercept term).

We applied SCADws, SCAD2s and Heur along with sparseLTS, TaL, and classical lasso on the training set, and evaluated their predictive power on the test set – hard trimming approaches used a conservative 50% trimming. Table 4.2 compares these methods in terms of sparsity, MAPE and TMSPE (also here with an upper 10% trimming). Focusing on out-of-sample prediction accuracy, as measured by MAPE and TMSPE, SCAD2s improves upon SCADws and outperforms other methods. SparseLTS and TaL perform comparably, outperforming our heuristic procedure and lasso. Moreover, while SCAD2s selects only 3 features, SCADws and TaL select 4 predictors, and their solutions are much sparser and thus more interpretable than the ones for sparseLTS and Heur. Lasso produces a very sparse solution but performs poorly in terms of predictive power, which is likely due to the presence of outlying cases. Overall, our analysis is consistent with previous studies in the literature, and also here SCAD2s provides a very good trade-off between sparsity and prediction accuracy.

4.6 Final remarks

We combine different contamination schemes with sparse estimation methods for linear regression settings. This extends robust, sparse estimators based on hard trimming, which explicitly assume only MSOM outliers, to the co-occurrence of VIOM outliers. Importantly, as we rely on nonconcave penalties, our approach bridges the gap between robust estimation methods enforcing sparsity based on convex penalties, and the use of optimal L_0-constraints. Moreover, unlike methods which provide a general down-weighting for all points based on M-estimation, our proposal effectively estimates the weight for each data point. Indeed, asymptotically, non-outlying cases receive full weights, MSOMs are excluded from the fit, and only VIOMs are down-weighted.

The theoretical results characterizing our proposal include its high breakdown point, a robust oracle property – which allows the number of feature to increase exponentially with the sample size – and the accurate detection of each type of outliers with probability tending to one. Moreover, including a computationally

cheap extra step, our proposal achieves a doubly robust strong oracle property. This provides optimal unit weights and thus an optimal trade-off between high-breakdown point and efficiency.

Our work can be extended in several directions. We plan to investigate scenarios with correlated errors, generalizing our approach for VIOM outlier detection to non-diagonal covariance matrices. More generally, we are studying high-dimensional mixed-effects linear models affected by data contamination, which allow one to effectively model data with a natural group structure (e.g., spatio/temporal relations). In this setting, VIOM outliers might also arise in the random effects. This has been investigated in Gumedze et al. (2010) for a single outlier in a known position, but we plan to extend it to the case of multiple MSOM and VIOM outliers in unknown positions.

Moreover, as our theoretical results critically rely on tuning parameters controlling the trade-off between sparsity and efficiency, we are interested in the development of suitable information criteria for sparse models affected by different sources of contamination, extending the robust BIC introduced in this work. We are also developing more effective ways to build proxy matrices used in our procedure, as well as iterative approaches. Relatedly, VIOMs can be studied in the context of heteroscedatic linear models, where some robust and sparse approaches have been recently investigated in Gijbels and Vrinssen (2019); Wang and Loh (2020). Finally, we are exploring how to include into our framework cellwise contamination (Alqallaf et al., 2009; Filzmoser et al., 2020), which is recently receiving a lot of attention for high-dimensional settings (Su et al., 2021; Bottmer et al., 2022).

"Strange events permit themselves the luxury of occurring"

Charlie Chan (1928)

Chapter 5

Robust Variable Selection with Optimality Guarantees for High-Dimensional Logistic Regression

This chapter is based on: Insolia, L., Kenney, A., Calovi, M., and Chiaromonte, F. (2021c). Robust variable selection with optimality guarantees for high-dimensional logistic regression. *Stats*, 4(3):665–681.

Reproducible and documented code for this chapter is available at: `https://github.com/LucaIns/SFSOD_logreg`.

5.1 Introduction

Logistic regression is widely used to solve classification tasks and provides a probabilistic relation between a set of covariates (i.e., features, variables or predictors) and a binary or multi-class response (Cox and Snell, 1989; McCullagh and Nelder, 1989). The use of the logistic function can be traced back to the early 19th century, when it was employed to describe population growth (Cramer, 2002). However, despite its popularity, the classical logistic regression framework based on maximum likelihood (ML) estimation can suffer from several drawbacks. In this work, we specifically

focus on two key challenges: high dimensionality and data contamination. The large dimensionality might lead to overfitting or even singularity of the estimates if the sample size is smaller than the number of features, and this motivates the use of penalized estimation techniques. Importantly, penalized methods can also promote sparsity of the estimates in order to improve the interpretability of the model (Friedman et al., 2010). On the other hand, the presence of outliers might disrupt classical and non-robust estimation methods leading to biased estimates and poor predictions. In particular, since the log-odds ratio depends linearly on the set of covariates included in the model, an *adversarial contamination* of the latter might create *bad leverage values* that break down ML-based approaches (Maronna et al., 2006). This motivates the development of robust estimation techniques. Notably, penalized estimation and robustness with respect to the presence of outliers are very closely related topics (She and Owen, 2011; Insolia et al., 2021d,a), and they have recently also been combined for logistic regression settings (Tibshirani and Manning, 2013; Kurnaz et al., 2017).

In this work, we provide a provably optimal approach to perform simultaneous feature selection and estimation, as well as outlier detection and exclusion for logistic regression problems. Here optimality refers to the fact the the global optimum of the underlying "double" combinatorial problem is indeed achievable and, even if the algorithm is stopped before convergence, one can obtain optimality guarantees by monitoring the gap between the best feasible solution and the problem relaxation (Bertsimas et al., 2016; Schrijver, 1986). Specifically, we consider an L_0 sparsity assumption on the coefficients (Zhang and Zhang, 2012) and a logistic slippage model for the outlying observations (Bedrick and Hill, 1990). We further build upon the work in Insolia et al. (2021d) and rely on L_0-constraints to detect outlying cases and select relevant features. This requires us to solve a double combinatorial problem, across both the units and the covariates. Importantly, the underlying optimization can be effectively tackled with state-of-the-art *mixed-integer conic programming* solvers. These target a global optimum and, unlike existing heuristic methods, provide optimality guarantees even if the algorithm is stopped before convergence.

We use our proposal to investigate the main drivers of honey bee (*Apis mel-*

lifera) loss during winter (overwintering), which represents the most critical part of the year in several areas (Seeley and Visscher, 1985; Döke et al., 2015; Beyer et al., 2018). In particular, we use survey data collected by the Pennsylvania State Beekeepers Association, which include information related to honey bee survival, stressors and management practices, as well as bio-climatic indexes, topography and land use information (Calovi et al., 2021). Previous studies mainly focused on predictive performance and relied on statistical learning tools such as random forest, which capture relevance but not effect signs for each feature, and do not account for the possible impact of outlying cases – making results harder to interpret and potentially less robust. In our analysis, based on a logistic regression model, we are able to exclude redundant features from the fit while accounting for potential data contamination through an estimation approach that simultaneously addresses sparsity and statistical robustness. This provides important insights on the main drivers of honey bee loss during overwintering – such as the exposure to pesticides, as well as the average temperature of the driest quarter and the precipitation level during the warmest quarter. We also show that the data set does indeed contain outlying observations.

The remainder of the paper is organized as follows. Section 5.2 provides some background on existing penalized and robust estimation methods. Section 5.3 details our proposal and its algorithmic implementation. This is compared with existing methods through numerical simulations in Section 5.4. Our analysis of the drivers of honey bee loss is presented in Section 5.5. Final remarks are provided in Section 5.6.

5.2 Background

Let $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})^T \in \mathbb{R}^{n \times p}$ be an observed design matrix, and $\boldsymbol{y} \in \{-1, 1\}^n$ the corresponding set of binary response classes. The *two-class logistic regression model* assumes that the log-odds ratio is a linear function of the covariates

$$\log\left(\frac{P(y_i = 1|\boldsymbol{x}_i)}{1 - P(y_i = 1|\boldsymbol{x}_i)}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}, \tag{5.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ are the unknown regression parameters (possibly sparse). We also assume the presence of an intercept term, so that $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_{p-1}\}$ and $\boldsymbol{X}$ contains only 1's in the first column. Thus, for any $\boldsymbol{x}_i \in \mathbb{R}^p$, it follows from (5.1) that

$$P(y_i = 1|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} = \frac{1}{1 + \exp(-\boldsymbol{x}_i^T\boldsymbol{\beta})}$$

and

$$P(y_i = -1|\boldsymbol{x}_i) = 1 - P(y_i = 1|\boldsymbol{x}_i) = \frac{1}{1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}.$$

Hence, in full generality, the logistic model can be expressed as

$$P(y_i|\boldsymbol{x}_i) = \frac{1}{1 + \exp(-y_i\boldsymbol{x}_i^T\boldsymbol{\beta})}. \tag{5.2}$$

Assuming that $y_i|\boldsymbol{x}_i$, for $i = 1, \ldots, n$, follow independent Bernoulli distributions, the likelihood function associated to (5.2) is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(y_i = 1|\boldsymbol{x}_i)^{(1+y_i)/2} P(y_i = -1|\boldsymbol{x}_i)^{(1-y_i)/2},$$

which provides the ML estimator

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} d(\boldsymbol{x}_i^T\boldsymbol{\beta}, y_i) \tag{5.3}$$

where the deviance is defined as: $d(\boldsymbol{x}_i^T\boldsymbol{\beta}, y_i) = \log\left(1 + \exp(-y_i\boldsymbol{x}_i^T\boldsymbol{\beta})\right)$. The optimization problem in (5.3) is convex, and it admits a unique and finite solution if and only if the points belonging to each class "overlap" to some degree (i.e., the two classes are not linearly separable based on predictors information; Albert and Anderson 1984; Santner and Duffy 1986). Otherwise, there exist infinitely many hyperplanes perfectly separating the data, and the ML estimator is undetermined. Importantly, in this setting, the ML estimator is consistent and asymptotically normal as $n \to \infty$ under weak assumptions (Fahrmeir and Kaufmann, 1985). However, unlike ML estimation for linear regression problems, there is no closed-form solution for (5.3), and iterative methods such as the Newton–Raphson algorithm are commonly employed, which can be solved through iteratively reweighted least

squares (McCullagh and Nelder, 1989; Hastie et al., 2009).

5.2.1 Penalized logistic regression

The ML estimator in (5.3) does not exist if $p > n$. Moreover, in the presence of strong collinearities in the predictor space, even if $p < n$, the ML estimator might provide unstable estimates or lead to overfitting (i.e., to estimates with low bias and high variance and thus poor predictive power). In order to overcome these limitations, penalized estimation methods based on the L_2-penalty have been considered (Duffy and Santner, 1989; Le Cessie and Van Houwelingen, 1992). To promote sparse estimates and improve interpretability, several authors also studied the use of the L_1-norm (Koh et al., 2007; Friedman et al., 2010). Although this class of "soft" penalization methods is computationally very efficient due to convexity, it provides biased estimates. Further approaches combine the L_1 and L_2-norms in what is known as the *elastic net* penalty (Zou and Hastie, 2005) – coupled with an adaptive weighting strategy to regularize the coefficients (Algamal and Lee, 2015). Importantly, under suitable assumptions, this guarantees that the resulting estimator satisfies the so-called oracle property, meaning that the probability of selecting the truly active set of covariates (i.e., the ones corresponding to nonzero coefficients) converges to one, and at the same time the coefficient estimates are asymptotically normal with the same means and variance structure as if the set of active features was known a priori (Fan and Li, 2001).

Best subset selection is a traditional "hard" penalization method that approaches the feature selection problem combinatorially (Miller, 2002). Ideally, one should compare all possible fits of a given size, for all possible sizes – say $1 \leq k_p \leq \min(n, p)$. This was long considered unfeasible for problems of realistic size p even in the linear regression setting (Hastie et al., 2009). Nevertheless, leveraging recent developments in hardware and *mixed-integer programming* solvers, Bertsimas et al. (2017) proposed the use of L_0-constraints on $\boldsymbol{\beta}$ to efficiently and effectively solve the underlying best subset logistic regression problem using *mixed-integer nonlinear programming* techniques. This extends the approach in Bertsimas et al. (2016) for linear regression and relies on the L_0 pseudo-norm, which is defined as $\|\boldsymbol{\beta}\|_0 = \sum_j I(\beta_j \neq 0)$,

where $I(\cdot)$ is the indicator function. Notably, oracle properties can be established in this setting under weaker assumptions than other proposals (Shen et al., 2012).

5.2.2 Robust logistic regression

Outliers may influence the fit, hindering the performance of ML-based estimators and leading to estimation bias and weaker inference (Copas, 1988). Multiple outliers are particularly problematic and difficult to detect since they can create masking (false negative) and swamping (false positive) effects (Imon and Hadi, 2008). Here, as in linear regression, raw (deviance) residuals can be used to build several regression diagnostics (Imon and Hadi, 2008; Landwehr et al., 1984; Pregibon, 1981). Different approaches have been introduced to overcome the limitations of classical ML estimation in low-dimensional settings (Maronna et al., 2006). For instance, a weighted counterpart of ML estimation was proposed in Carroll and Pederson (1993) (see also Rousseeuw and Christmann 2003), robust M-estimators were developed in Pregibon (1981), and Bianco and Yohai (1996) introduced an additional correction term that provides a robust class of Fisher-consistent M-estimators – see also Künsch et al. (1989); Croux and Haesbroeck (2003) for bounded influence estimators. Furthermore, an adaptive weighted maximum likelihood where the estimator efficiency is calibrated in a data-driven way was considered in Gervini (2005). A distributionally robust approach was proposed in Shafieezadeh Abadeh et al. (2015), which is similar in spirit to the use of robust optimization in Bertsimas et al. (2017) where uncertainty sets have to be taken into account.

The *logistic slippage model*, which closely resembles the mean-shift outlier model for linear regression problems (Beckman and Cook, 1983), was explicitly considered in Bedrick and Hill (1990) and leads to the removal of outliers from the fit. However, since the number and position of outlying cases are generally unknown, one should in principle compare the exclusion of $0 \leq k_n \leq n/2$ points from the fit (if one is willing to assume that less than half of the data are in fact contaminated). Building upon high breakdown point estimators and deletion diagnostics, a forward search procedure based on graphical diagnostic tools that is effective in detecting masked multiple outliers and highlights the influence of individual observations on

the fit was developed in Atkinson and Riani (2000, 2001). This approach is robust, computationally cheap and provides a natural order for the observations according to their agreement with the model.

For high-dimensional settings, Tibshirani and Manning (2013) focused on the possible contamination of the $\boldsymbol{y}$ labeling and proposed L_1 penalization methods for reducing the influence of outliers and performing feature selection. However, this provides a sub-optimal strategy both for sparse estimation (Shen et al., 2012) and outlier detection (She and Owen, 2011). More recently, the elastic net penalty has been combined with a trimmed loss function which excludes the k_n most influential observations from the fit (Kurnaz et al., 2017). This mimics the *least trimmed squares* (LTS) estimator for linear regression (Rousseeuw and Leroy, 1987), and is equivalent to assuming a logistic slippage model. On the other hand, the trimmed loss function is solved through heuristic methods based on resampling, and the elastic net penalty in use is sub-optimal in terms of feature selection.

5.3 MIProb: robust variable selection under the logistic slippage model

We consider a two-class logistic regression model affected by data contamination (i.e., outliers) and comprising irrelevant covariates. Specifically, we focus on the logistic slippage model, where the number, position and strength of the outliers are unknown (Beckman and Cook, 1983; Bedrick and Hill, 1990). The main idea is to enforce integer constraints on the number of outlying cases and relevant features in order to improve the interpretability of the model and its robustness. Now, we introduce a general formulation that in addition to simultaneous feature selection and outlier detection encompasses an optional ridge penalty, which can be useful to tackle strong collinearity structures (Zou and Hastie, 2005; Bertsimas et al., 2017), low signal-to-noise ratio regimes (Hastie et al., 2020) and data perturbations (Breiman, 1995). Thus, we propose to solve the following discrete optimization problem:

$$\left[\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}\right] = \underset{\boldsymbol{\beta},\boldsymbol{\phi}}{\arg\min} \sum_{i=1}^{n} d(\boldsymbol{x}_i^T\boldsymbol{\beta} + \phi_i, y_i) \tag{5.4}$$

$$\text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k_p \tag{5.4a}$$

$$\|\boldsymbol{\phi}\|_0 \leq k_n \tag{5.4b}$$

$$\|\boldsymbol{\beta}\|_2 \leq l. \tag{5.4c}$$

Due to the (double) combinatorial nature of the problem, the formulation in (5.4) is computationally daunting (Bernholt, 2006). Nevertheless, nowadays it can be solved effectively and at times also efficiently with specialized solvers. Importantly, it relates to the use of a trimmed loss function as in Kurnaz et al. (2017), and it extends the work in Insolia et al. (2021d) for sparse linear regression models affected by data contamination in the form of mean-shift outliers. However, here the use of a nonlinear and nonquadratic objective function complicates the matter and requires special attention.

We also note that (5.4) can be easily extended to model structured data, such as hierarchical or group structures. For instance, in Section 5.5 we enforce the so-called *group sparsity constraints* (Yuan and Lin, 2006) to model categorical features. Moreover, it can be naturally extended to multinomial logistic regression models along lines similar to those in Bianco and Yohai (1996).

5.3.1 Algorithmic implementation

The optimization problem in (5.4) can be formulated as a *mixed-integer conic program*. For simplicity, we first consider only the objective function and the L_2 ridge-like penalty. Specifically, including auxiliary variables $t_1, \cdots, t_n$ and r, the objective (5.4) and the constraint (5.4c) can be equivalently reformulated as

$$\min_{\boldsymbol{t},r,\boldsymbol{\beta}} \sum_{i=1}^{n} t_i + \lambda r \tag{5.5}$$

$$\text{s.t.} \quad t_i \geq \log\left(1 + \exp\left(-y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \phi_i)\right)\right) \tag{5.5a}$$

$$r \geq \|\boldsymbol{\beta}\|_2. \tag{5.5b}$$

The constraints in (5.5a) can be expressed using the exponential cone

$$K_{\text{exp}} = \left\{(x, y, z) \in \mathbb{R}^3 : y \exp(x/y) \leq z\right\},$$

and provide

$$\exp(-t_i) + \exp(u_i - t_i) \leq 1$$

where $u_i = -y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \phi_i)$. Including auxiliary variables z_{i1} and z_{i2} such that $z_{i1} \geq \exp(u_i - t_i)$ and $z_{i2} \geq \exp(-t_i)$, it follows that (5.5a) is equivalent to

$$\begin{cases} (u_i - t_i, 1, z_{i1}) \in K_{\text{exp}} \\ (-t_i, 1, z_{i2}) \in K_{\text{exp}} \\ z_{i1} + z_{i2} \leq 1. \end{cases}$$

Thus, the proposed mixed-integer conic programming formulation for logistic regression in (5.4) – denoted *MIProb* for simplicity – which provides sparse estimates for $\boldsymbol{\beta}$ and removes outliers through $\boldsymbol{\phi}$, is

$$\min_{\boldsymbol{t},\boldsymbol{z},r,\boldsymbol{\beta},\boldsymbol{z}^\beta,\boldsymbol{\phi},\boldsymbol{z}^\phi} \quad \sum_{i=1}^{n} t_i + \lambda r \tag{5.6}$$

$$\text{s.t.} \quad -\mathcal{M}_j^\beta z_j^\beta \leq \beta_j \leq \mathcal{M}_j^\beta z_j^\beta \tag{5.6a}$$

$$-\mathcal{M}_i^\phi z_i^\phi \leq \phi_i \leq \mathcal{M}_i^\phi z_i^\phi \tag{5.6b}$$

$$\sum_{j=1}^{p} z_j^\beta \leq k_p \tag{5.6c}$$

$$\sum_{i=1}^{n} z_i^\phi \leq k_n \tag{5.6d}$$

$$(u_i - t_i, 1, z_{i1}) \in K_{\text{exp}}$$

$$(-t_i, 1, z_{i2}) \in K_{\text{exp}}$$

$$z_{i1} + z_{i2} \leq 1$$

$$r \geq \|\boldsymbol{\beta}\|_2$$

$$z_j^\beta \in \{0, 1\}, \quad \beta_j \in \mathbb{R}, \quad j = 1, \ldots, p$$

$$z_i^\phi \in \{0, 1\}, \quad \phi_i \in \mathbb{R}, \quad i = 1, \ldots, n.$$

The big-$\mathcal{M}$ bounds $\boldsymbol{\mathcal{M}}^{\beta}$ and $\boldsymbol{\mathcal{M}}^{\phi}$ in constraints (5.6a) and (5.6b) have p and n entries, respectively, which can be tailored for each β_j and ϕ_i. These should be wide enough to include the true regression coefficients and zero-out the effects of the true outliers, but not so wide as to substantially increase the computational burden. For instance, an ensemble method based on existing heuristic and robust procedures to create suitable big-$\mathcal{M}$ bounds was considered in Insolia et al. (2021d). However, a similar approach is challenging in this framework given a "pool" of openly available robust algorithms is not available for logistic regression models – unlike in linear regression. Here, we simply set large, more conservative bounds to maintain accuracy at the cost of computing time. Extensions of additional heuristics to strengthen these bounds are worth further investigation, but beyond the scope of this work. The L_0-norm constraints (5.6c) and (5.6d) depend on positive integers k_p and k_n, which control the sparsity level for feature selection and the trimming level for outlier detection, respectively. As with any selection procedure, these tuning parameters are key to retain selection and detection accuracy. However, k_p and k_n can be treated differently. For the former, any deviation from the true sparsity level will result in false negatives/positives. For the latter, a common approach (Kurnaz et al., 2017; Atkinson and Riani, 2001) is to select an inflated trimming amount (i.e., higher than the true level) to avoid masking and swamping effects, and then refine the solution to recover efficiency.

Importantly, in this work we use existing specialized solvers (see Section 5.4) but the development of a tailored approach could be beneficial. For instance, outer approximation techniques in mixed-integer nonlinear programming with dynamic constraint generation were combined in Bertsimas et al. (2017), as well as the use of first-order methods, which reduce the computational burden compared to general-purpose solvers. Extensions of such approaches to this setting are left for future work.

5.3.2 Additional details

In order to achieve good estimates it is essential to tune the sparsity level k_p and the trimming level k_n, as well as the ridge-like tuning parameter λ if present, in a data-

driven fashion. For instance, one might consider robust counterparts of information criteria or cross-validation. In our simulation study, we do not include the L_2-constraint and, for a given trimming level k_n, we use a robust version of the *Bayesian information criterion* (BIC) similarly to Insolia et al. (2021d). In symbols, this is $\text{BIC} = \sum_{i=1}^{n} d(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}, y_i) + k_p \ln(n - k_n)$, where $d(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}, y_i)$ are the final deviances for a given estimator – recall that deviances corresponding to trimmed points are equal to 0. If an intercept term is included in the model, we force its selection as an active feature. Other tuning procedures such as cross-validation benefit from the use of effective warm-starts to accelerate convergence of the algorithm when solving over several training and testing sets splits – see Insolia et al. (2021d) for additional details.

The *breakdown point* (BdP) is the largest fraction of contamination that an estimator can tolerate before it might provide completely unreliable estimates (Donoho and Huber, 1983). It can be formalized either by replacing good units with outliers or adding outliers to an uncontaminated dataset. Using a unit-replacement approach, it has been shown that one can break down (unpenalized) ML estimation by simply removing units belonging to the overlaps among classes (Christmann, 1994; Künsch et al., 1989). Using unit-addition, Croux et al. (2002) showed that when severe outliers are added to a non-separable dataset, ML estimates do not break down due to "explosion" (to infinity), but they can break down due to "implosion" (to zero). Specifically, the BdP for the ML estimator is equal to $\varepsilon^*_{\text{ML}} = 2(p-1)/\{n+2(p-1)\}$ (which is 0% asymptotically), since the estimates can implode to zero, adding $2(p-1)$ appropriately chosen outliers. Thus, unlike in linear regression, here one has to take into account not only the explosion of the estimates, but also their implosion, which is often more difficult to detect.

We leave theoretical derivations concerning our MIProb proposal in (5.6) to future work. However, we note that MIProb clearly represents a trimmed likelihood estimator as a special case, so in this special case it inherits properties such as the high breakdown point (Müller and Neykov, 2003; Hadi and Luceño, 1997). Moreover, these results might be combined with the oracle properties for feature selection described in Shen et al. (2012) in order to obtain a logistic version of the robustly

strong oracle property introduced in Insolia et al. (2021d).

5.4 Simulation study

In this section, we use a simulation study to compare the performance of our proposal with state-of-the-art methods. The simulated data is generated as follows. The first column of the $n \times p$ design matrix $\boldsymbol{X}$ comprises all 1's (for the model intercept) and we draw the remaining entries of each row independently from a standard $(p-1)$-variate normal distribution $N(\boldsymbol{0}, \boldsymbol{I}_{p-1})$. The values of the p-dimensional coefficient vector $\boldsymbol{\beta}$ comprise p_0 non-zero entries (including the intercept) and $p - p_0$ zeros. The response labels $y_i \in \{1, -1\}$, for $i = 1, \ldots, n$, are generated from Bernoulli distributions with probabilities $1/(1 + e^{-\boldsymbol{x}_i^T \boldsymbol{\beta}})$. Next, without loss of generality, we contaminate the first n_0 cases with a logistic slippage model, adding the scalar mean shifts μ_X to the active predictors only (excluding the intercept). In order to generate *bad leverage points*, we also assign opposite signs to the labels of each contaminated unit: $\text{sign}(y_i) = -\text{sign}(\boldsymbol{x}_i^T \boldsymbol{\beta})$.

The simulation scenarios are defined according to the values of the parameters discussed above. Here, we present results for $p_0 = 4$ active predictors with $\beta_j = 3$ (without loss of generality, these correspond to the intercept and the last 3 features), sample size $n = 100$, increasing dimension $p = 20, 50$ (low) and 150 (high), $n_0 = 5$ contaminated units (i.e., 5% contamination), and mean shifts $\mu_X = 10$. Each simulation scenario is replicated q independent times, and random test data, say $(\boldsymbol{y}^*, \boldsymbol{X}^*)$, are generated from the same simulation scheme, but without any form of contamination.

Different estimators are compared based on: (i) the *mean of the negative log-likelihoods* $\text{MNLL}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} d(\boldsymbol{x}_i^{*\boldsymbol{T}} \widehat{\boldsymbol{\beta}}, y_i^*)$, i.e., the average of deviances computed on the uncontaminated test set; (ii) the outlier *misclassification rate* $\text{MR}(\widehat{\boldsymbol{\beta}}) = c/n$, where c counts the number of misclassified observations on the uncontaminated test set; (iii) estimation accuracy in terms of *average mean squared error* $\text{MSE}(\widehat{\boldsymbol{\beta}}) = \frac{1}{p} \sum_{j=1}^{p} \text{MSE}(\widehat{\beta}_j)$, where for each $\widehat{\beta}_j$ we decompose $\text{MSE}(\widehat{\beta}_j) = \frac{1}{q} \sum_{i=1}^{q} (\widehat{\beta}_{ji} - \beta_j)^2 = (\overline{\beta}_j - \beta_j)^2 + \frac{1}{q} \sum_{i=1}^{q} (\widehat{\beta}_{ji} - \overline{\beta}_j)^2$ in squared bias and variance (here $\overline{\beta}_j = \frac{1}{q} \sum_{i=1}^{q} \widehat{\beta}_{ji}$)

(iv) feature selection accuracy, measured by the *false positive rate* $\text{FPR}(\widehat{\boldsymbol{\beta}}) = |\{j \in \{1,\dots,p\} : \widehat{\beta}_j \neq 0 \wedge \beta_j = 0\}|/|\{j \in \{1,\dots,p\} : \beta_j = 0\}|$ and the *false negative rate* $\text{FNR}(\widehat{\boldsymbol{\beta}}) = |\{j \in \{1,\dots,p\} : \widehat{\beta}_j = 0 \wedge \beta_j \neq 0\}|/|\{j \in \{1,\dots,p\} : \beta_j \neq 0\}|$; (v) outlier detection accuracy, which is similarly measured by $\text{FPR}(\widehat{\boldsymbol{\phi}})$ and $\text{FNR}(\widehat{\boldsymbol{\phi}})$.

We use the *robust oracle* estimator as a benchmark, which is a logistic fit computed only for the active set of features and only on the uncontaminated units (we used our MIP formulation to compute the robust oracle). The following estimators are compared: (a) *enetLTS* with $\alpha = 1$ (i.e., robust lasso) (Kurnaz et al., 2017); (b) *MIProb*, our robust MIP proposal without a ridge-like constraint (see Section 5.3); (c) *MIP*, the non-robust MIP implementation performing only feature selection (i.e., as MIProb but using $k_n = 0$); (d) *Lasso*, the non-robust L_1-penalized loss computed through the `glmnet` package in `R` (Friedman et al., 2010). Robust methods trim the true number of outliers ($k_n = n_0$), though this does not guarantee exact outlier detection, and only the sparsity level in the feature space is tuned for each method based on (robust) information criteria or cross-validation. However, since enetLTS is a heuristic method that relies on resampling rather than exact trimming, we inflate the trimming proportion to 20% and then take the re-weighted estimates in order to improve its outlier detection performance.

Table 5.1 provides medians and median absolute deviations (MAD) of simulation results over $q = 30$ replications. Our proposal substantially outperforms competing methods in most criteria. In both low ($p = 20, 50$) and high ($p = 150$) dimensional settings, the MNLL and MR of MIProb are closest to values produced by the oracle. In terms of estimation accuracy, MIProb has the lowest bias, but the non-robust lasso has distinctly lower variance than all procedures aside from enetLTS when $p = 150$. MIProb has very strong feature selection accuracy with $\text{FPR}(\widehat{\beta})$ and $\text{FNR}(\widehat{\beta})$ equal to 0 in the low-dimensional settings ($p = 20, 50$). In the high-dimensional setting, it maintains the lowest false positive rate, but has a higher false negative rate than enetLTS (though still lower than the non-robust methods). This motivates the development of more effective tuning strategies as p increases. On the other hand, enetLTS tends to overselect, since it has $\text{FNR}(\widehat{\beta}) = 0$ in all settings, but the highest $\text{FPR}(\widehat{\beta})$ across methods. Similar results were found in Insolia et al. (2021d).

Regarding outlier detection, enetLTS and MIProb produce very similar solutions with FPR and FNR almost always 0. Thus, both methods are highly effective at detecting contaminated units.

5.4.1 Computational details

In this section, we discuss further computational details and the tuning approaches for each procedure. Our proposal, MIProb is computationally more demanding than the other methods under comparison, including the non-robust MIP. This is natural, given methods like enetLTS are heuristics and avoid directly solving the full combinatorial problem. As discussed in more detail in Kenney et al. (2021); Bertsimas et al. (2016), a common challenge with MIP formulations is the weak lower bound produced by the relaxed version of the problem. Thus, while the optimal solution may have already been found, the majority of computing time may be used to verify its optimality. For our settings, we set generous stopping criteria where the algorithm ends when either a maximum computing time of 40 min (this can be as low as 3 min in other literature; Hastie et al. 2020) or an optimality gap of 2.5% (i.e., the relative difference between the upper and lower bounds) is met. While this maximum time may be hit, especially under the most challenging scenarios with $p = 150$, the consistent quality of solutions close to the oracle (see Table 5.1) further supports this observation of weak lower bounds. However, for comparison, enetLTS only takes an average of 14 s. Thus, the use of other warm-starts, heuristics, etc., to improve lower bounds would be very beneficial for MIP-based feature selection and outlier detection approaches.

We also found that the computational burden of MIProb varies vastly based on the tuning parameter k_p. In our numerical experiments, computing time decreases as more features are selected, especially for $k_p > p_0$. For instance, we considered other simulation scenarios not reported here, including one with a lower sample size $n = 50$ and thus a higher contamination percentage. We observed the pattern in Figure 5.1 where the average computing time is much higher for lower values of k_p, but rapidly decreasing after the "elbow" occurring around $k_p = p_0$. This could be due to the outlier detection portion of the problem being more difficult when some

Table 5.1: Median (MAD in parenthesis) of MNLL, misclassification rate, variance and squared bias for $\widehat{\boldsymbol{\beta}}$, false positive rate and false negative rate for feature selection and outlier detection based on 30 simulation replicates.

n	p	Method	MNLL	MR	var($\widehat{\beta}$)	bias($\widehat{\beta}$)^2	FPR($\widehat{\beta}$)	FNR($\widehat{\beta}$)	FPR($\widehat{\phi}$)	FNR($\widehat{\phi}$)
100	20	Oracle	0.24(0.08)	0.11(0.04)	0.85(0.00)	0.29(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		enetLTS	0.50(0.08)	0.28(0.06)	0.03(0.00)	1.11(0.00)	0.18(0.09)	0.00(0.00)	0.005(0.008)	0.00(0.00)
		MIProb	0.27(0.04)	0.11(0.04)	0.01(0.00)	0.46(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		MIP	0.64(0.07)	0.31(0.04)	0.02(0.00)	1.52(0.00)	0.06(0.09)	0.75(0.00)	0.00(0.00)	1.00(0.00)
		Lasso	0.62(0.03)	0.29(0.04)	0.005(0.00)	1.52(0.00)	0.00(0.00)	0.75(0.00)	0.00(0.00)	1.00(0.00)
100	50	Oracle	0.22(0.04)	0.09(0.03)	0.10(0.00)	0.02(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		enetLTS	0.51(0.10)	0.27(0.13)	0.02(0.00)	0.45(0.00)	0.12(0.06)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		MIProb	0.26(0.04)	0.10(0.02)	0.01(0.00)	0.19(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		MIP	0.69(0.09)	0.37(0.06)	0.02(0.00)	0.63(0.00)	0.04(0.03)	0.75(0.00)	0.00(0.00)	1.00(0.00)
		Lasso	0.62(0.02)	0.29(0.03)	0.002(0.00)	0.63(0.00)	0.00(0.00)	0.75(0.00)	0.00(0.00)	1.00(0.00)
100	150	Oracle	0.21(0.05)	0.09(0.03)	0.07(0.00)	0.02(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		enetLTS	0.54(0.07)	0.29(0.03)	0.007(0.00)	0.16(0.00)	0.06(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		MIProb	0.34(0.12)	0.16(0.06)	0.03(0.00)	0.08(0.00)	0.00(0.00)	0.25(0.37)	0.00(0.00)	0.00(0.00)
		MIP	0.88(0.12)	0.42(0.03)	0.02(0.00)	0.21(0.00)	0.03(0.00)	0.75(0.00)	0.00(0.00)	1.00(0.00)
		Lasso	0.62(0.05)	0.30(0.06)	0.002(0.00)	0.21(0.00)	0.007(0.01)	0.75(0.00)	0.00(0.00)	1.00(0.00)

of the relevant features are not included. Recall that our simulations add mean shift contamination only to the relevant features; when some are missing, it is more challenging to detect contaminated units.

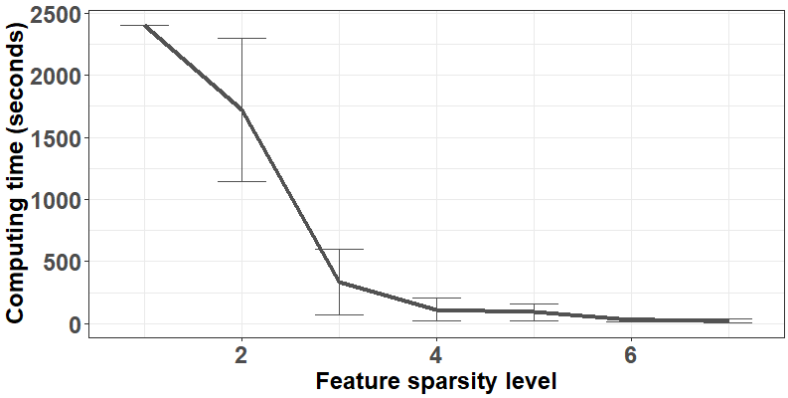


Figure 5.1: Average computing times across various feature sparsity levels k_p in simulated data following the data generation approach described above with $n = 50$, $p = 7$, $p_0 = 4$, and $k_n = 5$. Bars represent ± 1 standard deviations over 5 simulation replicates.

Regarding tuning, we utilized different approaches for each procedure as appropriate. The oracle operates on uncontaminated units and relevant features only, and requires no tuning. EnetLTS is tuned with cross-validation through the `enetLTS` package in `R` following the default settings with 5 folds (Kurnaz et al., 2018). For our proposal, MIProb, we used a robust version of BIC as described in Section 5.3.2, selecting the k_p corresponding to the minimum. Similarly, MIP is tuned based on the traditional BIC (without trimming incorporated). Finally, the non-robust lasso is tuned through 10-fold cross-validation in the `glmnet` package. We note that MIProb and MIP are implemented in `Julia 1.3.1` to interact with the `Mosek` solver through its `JuMP` package. MIProb and enetLTS utilize 24 cores per replication through their multi-thread options.

5.5 Investigating overwintering honey bee loss in Pennsylvania

Pollinators play a vital role supporting critical natural and agricultural ecosystem functions. Specifically, honey bees (*Apis mellifera*) are of great economic importance and play a primary role in pollination services (Calderone, 2012; Chopra et al.,

2015). The added value of honey bees pollination for the crops produced in the United States (in terms of higher yield and quality of the product) is annually estimated around 15–20 billion dollars (Morse and Calderone, 2000; Calderone, 2012), and according to the Pennsylvania Beekeepers Association, their yearly contribution has an estimated value of 60 million dollars in the state of Pennsylvania alone; see https://pastatebeekeepers.org/pdf/ValueofhoneybeesinPA3.pdf (accessed on 15 July 2021). Yet the decline of the honey bee populations is a widespread phenomenon around the globe (Becher et al., 2013; Pettis and Delaplane, 2010; Potts et al., 2010; Oldroyd and Nanork, 2009; Ellis et al., 2010). Major threats for honey bees include habitat fragmentation and loss, mites (van Dooremalen et al., 2012; Morawetz et al., 2019), parasites and diseases (Genersch et al., 2010), pesticides (Yasrebi-de Kom et al., 2019), climate change (Switanek et al., 2017), extreme weather conditions, the introduction of alien species (Stout, 2009), as well as the interactions between these factor (vanEngelsdorp and Meixner, 2010). Moreover, the overwintering period is often a major contributor to honey bee loss (Seeley and Visscher, 1985; Steinhauer et al., 2014; Döke et al., 2015; Bruckner et al., 2020). We thus focus on honey bee winter survival.

In the United States, beekeepers suffered an average 45.5% overwinter colony loss between 2020 and 2021 (Steinhauer et al., 2021). This figure was 41.2% in the state of Pennsylvania for the same overwintering period; see https://beeinformed.org/2021/06/21/united-states-honey-bee-colony-losses-2020-2021-preliminary-results/ (accessed on 15 July 2021). In both cases, this was an increase compared to the previous year, when the reported losses were 43.7% and 36.6% for the United States and Pennsylvania, respectively (Bee Informed Partnership, 2020). In recent years the trend of overwintering loss for Pennsylvania is comparable to the one at the national level, making it an interesting case study. Thus, in the following we analyze overwintering survey data for Pennsylvania covering the years 2016–2019.

5.5.1 Model formulation and data

Focusing on the state of Pennsylvania, honey bee winter survival was recently investigated in Calovi et al. (2021) based on winter loss survey data provided by the Pennsylvania State Beekeepers Association. The data cover three winter periods (2016–2017, 2017–2018, and 2018–2019), and the main goals of the analysis were to assess the importance of weather, topography, land use, and management factors on overwintering mortality, and to predict survival given current weather conditions and projected changes in climate. The authors utilized a random forest classifier to model overwintering survival. Importantly, they also controlled for the treatment of varroa mites (*Varroa destructor)* at both apiary and colony levels, since this represents a key factor in describing honey bee survival – all untreated colonies were excluded from the dataset. Their main findings suggest that growing degree days (see Table 5.2) and precipitations in the warmest quarter of the preceding year were the most important predictors, followed by precipitations in the wettest quarter, as well as maximum temperature in the warmest month. These results highlight the strong association between weather events and overwintering survival of honey bees.

The data set used in our analysis is extracted from the Supplementary Information published in Calovi et al. (2021) – see Table 5.2 for a description of the variables included into our model.

Since observations in the original data set represent colonies that may belong to the same apiary, we aggregated the data to obtain unique apiary information. This is particularly important in order to reduce dependence across observations, and leads to a sample of $n = 257$ apiaries from 1429 colonies (in the absence of publicly available geo-localized information, apiary identification was made possible through the features "bioc02" and "slope").

We created a binary response taking the proportion of survived colonies per apiary, and assigning the label 1 if such a mean is greater than 0.8, and the label -1 if it is smaller than 0.6. These thresholds are motivated by the "average" winter colony loss rate described above and they allow us to study the most "extreme" behavior (significantly higher or lower losses); they also provide a balanced labeling

for the response variable. The remaining observations are completely removed from the data set in use and thus decreasing the sample size to $n = 216$.

Table 5.2: Description of the features included into our logistic model formulation to describe honey bee overwintering survival. See Calovi et al. (2021) for details. The bioc# variables refer to bioclimatic variables of the WorldClim database; see https://www.worldclim.org/data/bioclim.html (accessed on 15 July 2021).

	Variable	Description
1	survival	Binary survival response at the apiary level
2	bee2	Winter total precipitation
3	bee4	Winter days with maximum temperature above 16°C and precipitation below 3 mm
4	gdd	Growing degree days (base 5°C) as the accumulation of average daily temperatures
5	dd rain	Days between rain events ¿ 0.25 mm
6	bioc02	Mean diurnal temperature range
7	bioc04	Temperature seasonality
8	bioc08	Mean temperature of the wettest Quarter
9	bioc09	Mean temperature of the driest quarter
10	bioc18	Precipitation of the warmest quarter
11	bioc19	Precipitation of the coldest quarter
12	slope	Terrain slope
13	sol rad	Potential incident solar radiation, 21 December
14	pcurv	Profile curvature
15	tcurv	Terrain curvature
16	TWI	Topographic wetness index
17	EW	East/West orientation of slope
18	ITL	Distance-weighted insect toxic load
19	col nov	Number of colonies in November
20	exp 1-2	Beekeeper years of experience between 1 and 2 (binary variable)
21	exp 2-5	Beekeeper years of experience between 2 and 5 (binary variable)
22	exp <1	Beekeeper years of experience less than 1 (binary variable)
23	exp >10	Beekeeper years of experience greater than 10 (binary variable)

We compared the same procedures considered in our simulation study (see Section 5.4) without introducing a ridge-like penalty for any of the methods. Relatedly, we did not use all features in the original study, which presented sizable collinearities. In particular, for each pair of features with an absolute pair-wise correlation above 0.7, we computed the mean absolute correlation of each feature against all the others and removed the one with the largest mean absolute correlation from our pool.

Each column of the design matrix $\boldsymbol{X}$ (excluding the intercept and categorical factors) was standardized to have zero median and median absolute deviation (MAD) equal to the average MAD across columns (standardization does not affect our proposal and each of the other approaches included in our comparison performs its

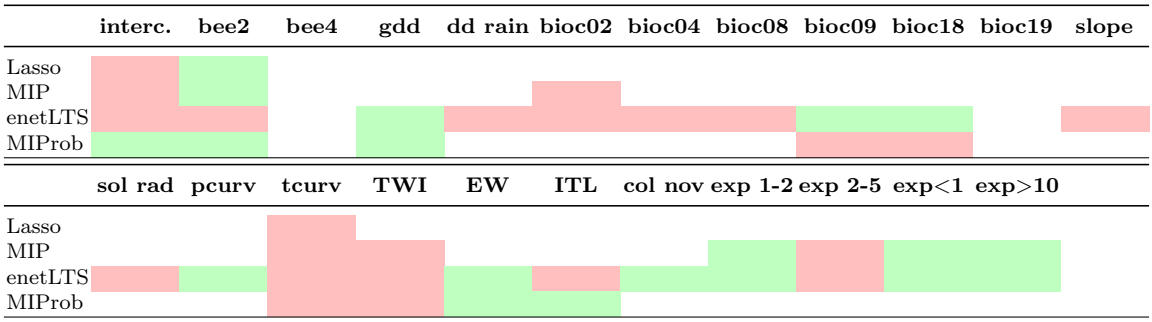
own standardization as needed). Importantly, for MIP and MIProb we introduced *group sparsity constraints* (Yuan and Lin, 2006) to tackle the categorical feature "beekeepers' experience"; the reference category is "between 5 and 10 years" and all coefficients for the dummy variables are included or excluded from the fit together.

5.5.2 Results

We randomly split the data into training and test sets, encompassing 100 and 116 points, respectively. For robust methods, we fix the trimming proportion at 10% after exploring a range of values suitable for the nature of the problem, and only tune the sparsity level. Figure 5.2 compares the *balanced accuracy*, defined as (sensitivity+specificity)/2, on the test set across different methods. Here sensitivity is defined as (# true positives)/(# true positives + # true negatives) and specificity is defined as (# true negatives)/(# true negatives + # false positives). While this is a function of the sparsity level imposed on MIP and MIProb, for enetLTS and lasso we show the mean values across eight repetitions due to the intrinsic randomness induced by cross-validation methods (horizontal dashed lines). Here MIP and MIProb are quite comparable and generally outperform competing methods, although we notice a drop in predictive performance for MIProb if the sparsity level $k_p \geq 9$ – which is likely a result of overfitting due to data trimming compared to MIP. Based on these findings, in the following we present the results based on $k_p = 8$ (including the intercept), where the balanced accuracy for both methods is very close to their maximum.

Table 5.3 displays the features selected by each method on the training set. We focus on the interpretation of the signs of the estimated coefficients, represented as green (positive) and red (negative) cells, respectively. The estimates provided by MIProb are in line with the findings of the original study (Calovi et al., 2021). Specifically, MIProb estimates a positive association between honey bee survival and "bee2" (winter total precipitation), "gdd" (growing degree days), "EW" (East/West orientation of slope) and "ITL" (distance-weighted insect toxic load; see Douglas et al. 2020). This suggests that the impact of precipitations and the accumulation of average daily temperatures (gdd), which influence the growth of crops, have

Table 5.3: Features selected by lasso, MIP, enetLTS and MIProb (robust MIP) on a training set encompassing 100 points. Green and red cells indicate estimated coefficients with positive and negative signs, respectively. White cells indicate non-selected features.



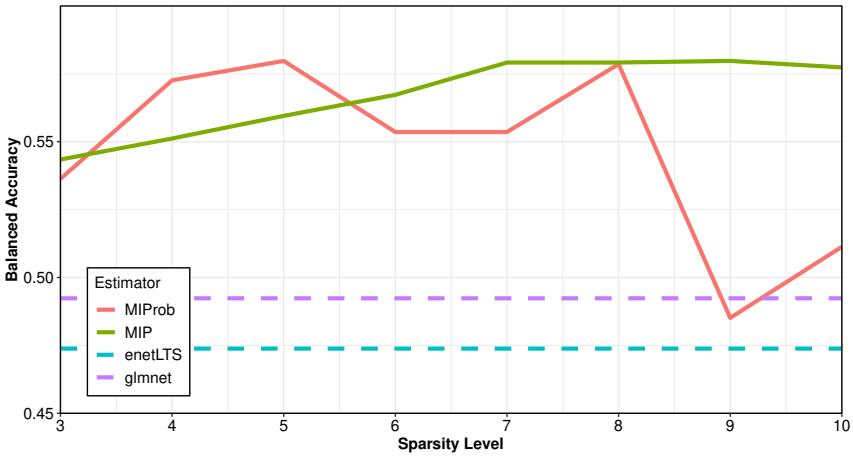


Figure 5.2: Balanced accuracy computed on a test set encompassing 126 points, as a function of the sparsity level k_n for MIP and MIProb (using a 10% trimming for the latter). The average balanced accuracy over 8 repetitions is shown also for lasso and enetLTS.

an overall positive effect on honey bee survival. In contrast, MIProb estimates a negative association between honey bee survival and "bioc09" (mean temperature of the driest quarter), "bioc18" (precipitation of the warmest quarter), "tcurv" (terrain curvature) and "TWI" (topographic wetness index). This highlights once more the major impact of weather predictors, as well as topographic factors and humidity levels. Notably, beekeepers' experience was not selected as a relevant feature by MIProb, which further supports the findings in Calovi et al. (2021).

Considering the other procedures, enetLTS appears to produce denser solutions (this was also observed in the simulations in Section 5.4), excluding only three features from the fit, and the non-robust lasso appears to produce sparser solutions, selecting only three features – which is indeed due to the presence of outliers. This is supported by the fact that a lasso fit after the exclusion of the outliers detected

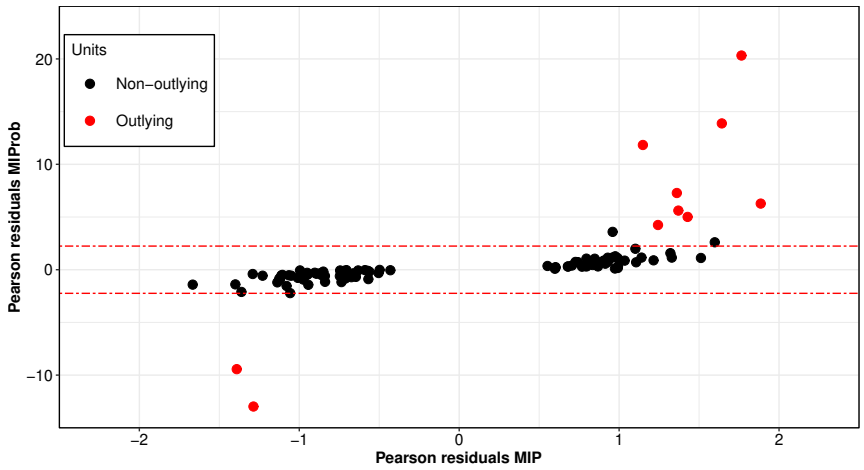


Figure 5.3: Pearson residuals for MIProb and MIP. Outlying cases detected by MIProb are highlighted in red. Horizontal red lines represent the 0.0125 and 0.9875 quantiles of the standard normal distribution.

by MIProb provides richer solutions, corresponding to clearer minima of the cross-validation error, where approximately 10 features are selected and several of these are shared with MIProb (data not shown). MIP uses the same sparsity level of MIProb but selects a different set of features, which is again due to the presence of outliers (e.g., it selects "bioc02" and the dummies related to beekeepers' experience).

Figure 5.3 compares Pearson residuals for MIProb and MIP estimators. The outlying cases detected by MIProb, which are highlighted in red, deviate substantially from the remaining observations and are undetected by the non-robust MIP algorithm. Moreover, focusing on the set of features selected by MIProb, Figure 5.4 compares the boxplots of outliers selected by MIProb against the remaining non-outlying cases. We notice that the two distributions are indeed quite different for variables such as "bee2", "gdd", "EW" and "ITL". This provides further evidence that the data set contains some outlying cases which significantly differ from the rest of the points.

5.6 Final remarks

We propose a discrete approach based on L_0-constraints to simultaneously perform feature selection and multiple outlier detection for logistic regression models. This is important since modern (binary) classification studies often encompass a large

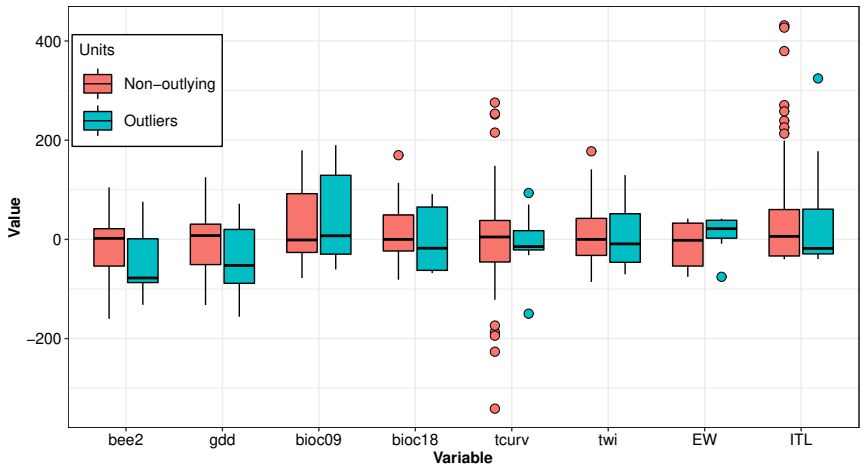


Figure 5.4: Box-plots comparing the values assumed by the features selected by MIProb contrasting outlying and non-outlying case. The values of each feature are scaled to have zero median and MAD equal to the average MAD across columns.

number of features, which tends to increase the probability of data contamination. Outliers need to be detected and treated appropriately, since they can hinder classical estimation methods. Specifically, we focus on the logistic slippage model, which leads to the exclusion (or trimming) of the most influential cases from the fit, and a "strong" sparsity assumption on the coefficients. To solve such a double combinatorial problem, we rely on state-of-the-art solvers for mixed-integer conic programming which, unlike existing heuristic methods for robust and sparse logistic regression, provide guarantees of optimality even if the algorithm is stopped before convergence. Our proposal, MIProb, provides robust and sparse estimates with an optional ridge-like penalization term.

MIProb, outperforms existing methods in our simulation study. It provides sparser solutions with lower false positive and negative rates for both feature selection and outlier detection while maintaining stronger predictive power under most settings. Moreover, MIProb performs very well in our honey bee overwintering survival application. Based on three years of publicly available data from Pennsylvania beekeepers, it outperforms existing heuristic methods in terms of predictive power, robustness and sparsity of the estimates, and it produces results consistent with previous studies (Calovi et al., 2021). In particular, we found that weather variables appear to be strong contributors. Winter total precipitation and growing degree days are positively associated with honey bee survival, while the mean temperature

of the driest quarter and the precipitation of the warmest quarter show a negative association. Moreover, our results indicate that the lower the exposure to pesticide (i.e., as their distance increases) the higher honey bee survival is. These findings are important in order to understand the main drivers of honey bee loss and highlight the importance of multi-source data to study and predict honey bee overwintering survival.

Our work can be extended in several directions. We are exploring additional and more complex simulation settings (e.g., higher dimensionality, collinear features, etc.). We did not experiment with the ridge-like penalty in the current paper, but this is an important tool and requires further investigation. However, computing time is currently the main bottleneck to more extensive exploration. Thus, in the future , we plan to consider additional modeling strategies that can reduce the computational burden. For instance, developing more suitable big-$\mathcal{M}$ bounds and using outer approximation techniques with dynamic constraint generation and first-order techniques as in Bertsimas et al. (2017). We are also exploring more efficient tuning strategies for the sparsity and trimming levels, as well as the ridge-like parameter, if present. Utilizing approaches such as warm-starts or integrated cross-validation (Kenney et al., 2021) can substantially reduce the computational burden for subsequent runs of the MIP algorithm, and allow better tuning. If the trimming level for MIProb is inflated, a re-weighting approach may also be included in order to increase the efficiency of the estimator as in Kurnaz et al. (2017), as well as approaches based on the forward search Atkinson and Riani (2001). However, larger trimming levels might increase the computational burden, and the procedure does not take into account the feature selection process. Thus, the forward search might be combined with diagnostic methods that simultaneously study the effect of outliers and features (Menjoge and Welsch, 2010). Moreover, the theoretical properties of our procedure require further investigation, and its extension to other generalized linear models such as Poisson or multinomial regressions is of great interest.

Source code for the implementation of our procedure and to replicate our simulation and application results is openly available at `https://github.com/LucaIns/SFSOD_logreg` (accessed on 29 July 2021).

"We are drowning in information and starving for knowledge"

Rutherford D. Roger (1985)

Chapter 6

Final Remarks

In this final chapter we summarize our work, discuss some of its limitations and possible extensions, as well as other recent developments in the areas of high-dimensional modeling and robust statistics that can drive future directions of research.

6.1 Discussion

We studied high-dimensional regression models affected by the presence of various forms of data contamination. To establish our framework, in Insolia et al. (2021b) we first focused on low-dimensional linear regression settings affected by the co-occurrence of multiple outliers arising from mean-shift (MSOM) and variance-inflation outlier models (VIOM), which lead to the exclusion and down-weighting of outlying cases, respectively (Beckman and Cook, 1983). Here the MSOM provides a safeguard against outliers arising both in the response and the explanatory variables. Combining hard trimming methods such as the forward search (Atkinson and Riani, 2000) with restricted maximum likelihood estimation (Harville, 1977), we proposed a novel approach to detect and treat both types of outliers. Unlike commonly employed soft trimming procedures which down-weight all points (Maronna et al., 2006), in our proposal MSOMs are excluded from the fit, VIOMs are down-weighted, and non-outlying cases receive full weights. We also developed regression diagnostics that help to guide the analysis, and demonstrated the effectiveness of our proposal through synthetic and real-world data. Notably, we combined hard

and soft trimming procedures in a principled way and, to our knowledge, ours is the first approach that effectively deals with multiple VIOM outliers, as well as the joint presence of MSOMs and VIOMs – which were traditionally considered as alternative contamination mechanisms (Cook et al., 1982; Thompson, 1985).

The MSOM, unlike the VIOM, has received some attention in the context of high-dimensional modeling during the last decade (albeit often implicitly, through the use of hard trimming estimators; Alfons et al. 2013; Kurnaz et al. 2017). However, existing methods rely on heuristic algorithms to solve the non-convex robust loss, which can only attain a local minimum, and enforce sparsity through convex relaxations of best subset selection. To fill this gap, in Insolia et al. (2021d) we proposed the use of discrete optimization techniques (Bertsimas et al., 2016) to simultaneously perform feature selection and outlier detection with optimality guarantees – meaning that the global optimum is indeed achievable. Specifically, we transformed MSOM detection into a feature selection problem and developed a general framework for *robust best subset selection* which encompasses a broad class of models and loss functions. Focusing on the L_2-loss subject to cardinality constraints (enforcing the selection of both units and features), we proposed a mixed integer quadratic formulation and derived stronger theoretical results under weaker assumptions compared to state-of-the-art procedures. These include what we called the *robustly strong oracle property*, which ensures that our proposal asymptotically behaves as if the sets of truly relevant features and MSOM outliers were known in advance. This also holds in the ultra-high dimensional setting, where the number of predictors increases exponentially with the number of uncontaminated samples. We showed that our proposal outperforms existing methods through numerical simulations and a real-world application where we analyzed the relationship between childhood obesity and the human oral microbiome. The model we obtained in the application was more interpretable and had higher predictive power than those reported in prior analyses of the same data. Moreover, the features we selected as relevant were consistent with prior literature on childhood obesity, and the cases we detected as outliers were very informative. We note that an approach similar to ours has been recently brought forth in Thompson (2022) and Jammal et al.

(2021), which employ effective heuristics based on, respectively, projected block-coordinate gradient descent and proximal alternating linearized minimization. This demonstrates the importance of the topic.

In Insolia et al. (2021a), we further extended our work to high-dimensional linear regression models affected by the co-occurrence of multiple MSOM and VIOM outliers – the former concerning both response and design matrix. We employed a mixed-effects linear model where MSOM and VIOM outliers are represented as additional fixed and random components, respectively, and we developed a doubly robust class of nonconcave penalization methods which extend the approach in Fan and Li (2012). Specifically, based on restricted maximum likelihood principles, we showed that MSOM and VIOM outliers can be detected and treated independently, and proposed a 3-step procedure to perform model selection and detect outlying cases of each type. Our procedure satisfies several desirable properties, such as the *doubly robust strong oracle property* which, under mild assumptions, generalizes our robustly strong oracle property to the co-occurrence of multiple MSOM and VIOM outliers. This implies that one can estimate optimal units' weights, and thus achieve an optimal trade-off between high breakdown point and efficiency, as well as improve estimation of the error variance. We also proposed a data-driven procedure for VIOM detection, which provides a considerable gain with respect to existing soft trimming estimators that rely on a given nominal efficiency (Maronna et al., 2006; Chang et al., 2018), and thus are not adaptive. This improves the estimation of error variance, which is fundamental for several statistical learning goals (Fan et al., 2012a; Reid et al., 2016). Moreover, for the class of hard trimming methods, our proposal based on nonconcave penalties provides a bridge between robust estimators that rely on convex penalties (Alfons et al., 2013; Kurnaz et al., 2017) and our combinatorial approach based on L_0-constraints and discrete optimization techniques (Insolia et al., 2021d), which can be exploited by the latter to reduce its computing time. A comparison with more recent soft trimming methods based on nonconvex penalties would also be of interest (Amato et al., 2021; Kepplinger, 2021). The effectiveness of our proposal was demonstrated through Monte Carlo simulations and real-world applications related to the Boston housing market and

glioblastoma gene expression data.

Finally, in Insolia et al. (2021c) we focused on the logistic regression model and extended our approach in Insolia et al. (2021d) to high-dimensional binary classification studies affected by data contamination. We developed a sparse and robust estimation procedure that simultaneously selects and estimates regression parameters, and detects and excludes outliers from the fit. In particular, we transformed outlier detection into a feature selection problem through an over-parametrized *logistic slippage model* (Bedrick and Hill, 1990) – which is the counterpart of the MSOM for linear models – and used again discrete optimization techniques. To tackle the presence of a non-quadratic objective function, we developed a mixed integer conic programming formulation, which allows us to solve the underlying double combinatorial problem with optimality guarantees. We used simulations to show that our approach outperforms existing methods based on heuristic algorithms and convex penalties (Kurnaz et al., 2017). Notably, soft trimming estimators with nonconvex penalties have been recently developed by Bianco et al. (2021), and Bianco et al. (2022) studied the setting with a diverging number of parameters. This highlights the importance of this topic. We also used our proposal to study the main drivers of honey bee (*Apis mellifera*) colony loss analyzing data from the annual winter loss survey collected by the Pennsylvania State Beekeepers Association (Calovi et al., 2021). Our proposal provided a parsimonious and interpretable classification model, supporting previous findings in the literature, and highlighted the presence of outlying observations within the survey data.

6.2 Extensions

Our contributions to date can be expanded and strengthened in a number of ways. Below, we briefly sketch a few selected topics we are actively investigating at this time.

Computational efficiency for mixed integer programs The approaches for simultaneous feature selection and outlier detection we introduced in Insolia et al.

(2021d,c) are based on discrete optimization. While very effective, they can be computationally burdensome for larger problems. It is thus of great importance to develop more efficient estimation techniques.

Leveraging recent advances in the use of MIPs to solve best subset selection, we are particularly interested in the development of perspective cuts formulations to obtain tighter lower bounds (Frangioni and Gentile, 2009; Gómez, 2021), novel heuristics and model relaxations (Bertsimas et al., 2016; Willis and von Stosch, 2017), coordinate descent and local combinatorial optimization (Hazimeh and Mazumder, 2020), as well as in exploiting the dual formulation of classification tasks (Bertsimas et al., 2021b).

Relatedly, we are also developing a MIP-based implementation of the FAST-LTS algorithm (Rousseeuw and Van Driessen, 1999, 2006). This is an heuristic method that relies on resampling and so-called *concentration steps* (C-steps) which guarantee the attainment of a local optimum. It is computationally very lean and effectively provides high-breakdown point estimates. Notably, its core idea is widespread in a variety of statistical domains – including both supervised to unsupervised learning problems. Building upon existing studies (Agulló, 2001), we are combining FAST-LTS with the optimal approach based on a combinatorial search, where C-steps are embedded within a discrete optimization framework. This promises to significantly cut the computational burden of our MIP proposals, leading to tighter big-$\mathcal{M}$ bounds, warm-starts and heuristics.

Choice of the tuning parameters Most of our theoretical results rely on the choice of suitable tuning parameters (e.g., the sparsity and the contamination levels), which strongly affect the estimator performance. Although they are guaranteed to exist (at least asymptotically), finding appropriate tuning parameter values in a data-driven fashion is non-trivial.

This is a common problem in the literature on high-dimensional modeling and robust statistics; we provided general guidelines to choose tuning parameters, but there are issues that deserve further attention. Indeed, in our work to date, we often focused on tuning one key parameter fixing the others – which reduces computing

time but limits the search space. In the future, we plan to develop a computationally lean and simultaneous tuning strategy for all key parameters in our proposals, and study its theoretical properties.

Tuning procedures based on *robust information criteria* to detect outlying cases in low-dimensional linear models have been recently investigated by Riani et al. (2022). They provide additional insights for the methods we proposed in this thesis, and we plan to investigate their connection with recent advances for MIP-based techniques (Gómez and Prokopyev, 2021). In addition, we plan to develop *robust cross-validation* strategies that put more emphasis on out-of-sample performance. Cross-validation relies on less stringent assumptions compared to information criteria, but this comes at the cost of a larger computational burden – which is particularly taxing for MIP-based procedures. To mitigate this issue, several strategies have been recently incorporated into best subset selection algorithms (Kenney et al., 2021; Takano and Miyashiro, 2020; Kreber, 2019), and we plan to extend these strategies to settings involving data contamination. Robust cross validation based on soft trimming approaches was studied by Ronchetti et al. (1997). But for hard trimming methods, to our knowledge, existing approaches assume (often implicitly) that the fraction of contamination in training and testing sets is comparable. Some of our simulations highlighted the fact that this condition can be quite restrictive at times – e.g., in the presence of weak signals and small sample sizes. For this reason, in Insolia et al. (2021d) we used a much more conservative trimming proportion on testing data – but this, too, may be unsatisfactory in some settings. We are thus exploring the use of a "double tuning" strategy that controls for a different amount of contamination on training and testing sets.

When a problem exhibits strong collinearities among the predictors, a typical remedy is to include an additional ridge-like constraint in the feature selection formulation. But this adds a tuning parameter and can substantially increase the computational burden of MIP-based procedures. For our simultaneous feature selection and outlier identification set-ups, we plan to investigate the use of ridge-like constraints, as well as the use of *data whitening* techniques that de-correlate the original features while attempting to preserve the interpretability of the model (Ken-

ney, 2021). We note though that existing data whitening techniques are non-robust against the presence of data contamination since they assign full weight to each observation. Hence, we are developing a robust counterpart of whitening, which is similar in spirit to dimension reduction techniques such as robust principal component analysis (Hubert et al., 2005).

Mixed-effects linear models Our work in Insolia et al. (2021a) can be naturally extended to mixed-effects linear models, which have a greater flexibility in accommodating a large variety of problems and datasets, such as longitudinal studies, where outlying cases might also arise in the random design matrix.

High-breakdown point estimators based on hard trimming have been recently investigated in this context (Copt and Victoria-Feser, 2006; Chervoneva and Vishnyakov, 2011; Koller, 2013; Chervoneva and Vishnyakov, 2014; Agostinelli and Yohai, 2016). Relatedly, Gumedze et al. (2010) studied the presence of a single VIOM outlier which may also arise within a given random component. However, the treatment of outliers in high-dimensional mixed-effects linear models has not yet received much attention in the literature. Notably, Fan et al. (2012b, 2014b) proposed the use of a penalized likelihood to jointly estimate fixed and random effects, while employing Mahalanobis distances to estimate units' weights and accommodate outliers. However, estimated weights can be improved if one takes into account the underlying regression structure.

In full generality, we plan to combine and extend the approaches in Gumedze et al. (2010) and Fan and Li (2012) through the inclusion of random effects $\boldsymbol{Zb}$ in (4.2); see Chapter 4. In addition, a group of observations belonging to one (or more) random component(s), say the units indexed by $\mathcal{S}_\delta$, may be affected by a variance inflation parameter $\boldsymbol{\delta}$. Thus, multiple VIOM outliers may arise both at the unit level and at the group level, which are captured by non-zero entries in $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, respectively. Selecting relevant random effects (non-zero entries in $\boldsymbol{b}$) while detecting and accounting for VIOM outliers therein (non-zero entries in $\boldsymbol{\delta}$) is fundamental since the exclusion of relevant random components introduces bias in the estimated variance for fixed effects, and the inclusion of irrelevant random effects

might provide singular or unstable estimates (Laird and Ware, 1982). Here the key difference from our current proposal is that one needs additional proxy matrices for population covariances of random effects, which may lead to a generalized least square estimator as opposed to a weighted least squares fit. Furthermore, group constraints should be enforced for the penalization of random effects in $\boldsymbol{b}$ (Yuan and Lin, 2006); i.e. all q_k random coefficients belonging to the k-th random component should be fully included or excluded from the model, as proposed in Fan and Li (2012). Interestingly, although the use of a group sparsity penalty on $\boldsymbol{b}$ is similar in spirit to the VIOM parametrization through $\boldsymbol{\delta}$, these are not equivalent. Moreover, the inclusion of both $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ might be redundant in some applications (e.g., if only a few repeated measurements per unit are available), and iterative schemes can be also developed to tackle this problem.

Other extensions We plan to extend our approaches to a broader class of models. We explored logistic regression in Insolia et al. (2021c), where we expect to improve both algorithmic efficiency and theoretical characterization. Other members of the *generalized linear models* (GLM) family are of great interest as well, and we plan to develop a general theory for simultaneous feature selection and outlier detection in GLMs. Several strategies can be devised in tackling the presence of more complex loss functions, such as the use of piecewise-linear approximations that are amenable for modern MIP solvers (Sato et al., 2016, 2017; Bertsimas and Copenhaver, 2017; Saishu et al., 2021). Similar ideas can be extended also to the more flexible class of *generalized additive models*, which allow one to model continuous and discrete data in a semi-parametric fashion (Avella Medina, 2016).

Other modeling strategies for structured data are also of great interest to us – as they allow one to exploit richer information in an analysis. For instance, we are currently developing approaches for robust and sparse cluster-wise regression, where the data are characterized by intrinsic group structures (García-Escudero et al., 2010; Torti et al., 2019).

6.3 Directions for future research

We conclude this dissertation with an outline of some recent breakthroughs in high-dimensional modeling and robust statistics. They offer new insights that place the development of robust model selection into a broader perspective and raise a number of research questions that can guide future studies.

Propagation of outliers In the traditional Tukey-Huber contamination mechanism – which, among others, gives rise to the mean-shift and variance-inflation outlier models considered in this thesis – full observations are detected as outlying, meaning that the entire $(p+1)$-dimensional vector of a given i-th unit is estimated as such. This paradigm, also known as *case-wise contamination*, may suffer from several drawbacks in the presence of heterogeneous and high-dimensional data that are typical in modern applications. For instance: (i) as the problem dimension p increases, it is more likely that any given unit can in fact be contaminated in at least one of its measurements; (ii) even with a small number of features, removing or down-weighting an entire observation due to the contamination of a limited number of its entries may lead to a loss of information. As stated by Rousseeuw and Bossche (2018): "*recently researchers have come to realize that the outlying rows paradigm is no longer sufficient for modern high-dimensional datasets*".

The *independent contamination model*, also known as *cell-wise contamination*, can be effective in tackling these problems and was first introduced by Alqallaf et al. (2009). For a data matrix $\boldsymbol{Z} = (\boldsymbol{y}, \boldsymbol{X}) \in \mathbb{R}^{n\times(p+1)}$, it assumes that the Tukey-Huber contamination mechanism operates independently on each variable. Thus, for any j-th variable (where $j = 1, \ldots, p+1$), it postulates that $Z_{ij} \sim (1-\epsilon)F_j + \epsilon C_j$, indicating that each entry is drawn from the "true" distribution F_j with probability $(1-\epsilon)$, and from a contaminating distribution C_j with (a small) probability ϵ. Therefore, the expected proportion of units which are contaminated in at least one of their entries is equal to $1-(1-\epsilon)^{p+1}$. This quantity is very likely to exceed 50% as p increases, even for small ϵ values, and thus existing robust and equivariant estimators breakdown – a phenomenon called *propagation of outliers* in Alqallaf

et al. (2009).

The presence of cell-wise outliers motivated the rapid development of a new generation of robust estimation procedures across several domains, such as multivariate location and scatter (Van Aelst et al., 2012; Agostinelli et al., 2015; Leung et al., 2017), clustering (Farcomeni, 2014), data pre-processing (Rousseeuw and Bossche, 2018), linear regression (Leung et al., 2016; Öllerer et al., 2016; Filzmoser et al., 2020), feature selection (Wang, 2019; Su et al., 2021), and feature screening (Wang and Van Aelst, 2019).

Cell-wise outliers pose new and fascinating challenges to the field of robust statistics, since traditional results are not guaranteed to hold or can be more difficult to be derived (e.g., the breakdown point or the influence function of the estimator). It would be of great interest to extend our contributions to the co-occurrence of case-wise (such as MSOM and VIOM) and cell-wise contamination mechanisms, and several strategies can be envisaged to tackle this problem.

Robust sure screening properties Our main theoretical results, such as the doubly robust strong oracle property and the robustly strong oracle property, hold also if the number of explanatory variables increases exponentially with the sample size. Nevertheless, this setting poses serious challenges in terms of computing time and memory allocation. Indeed, even in the absence of outlying cases, sparse estimation techniques require that the data (or its sparse representation) can be loaded into computer memory altogether with intermediate results of the computations (Qian et al., 2020). Clearly, this problem is exacerbated for techniques based on discrete optimization, as well as in the presence of data contamination.

To overcome these issues, *feature screening* procedures are recently receiving a lot attention. They aim at considerably reducing the set of candidate predictors through computationally lean filters that can avoid the removal of possibly relevant predictors. The set of features passing this screening phase, can then be refined by a second step of sparse estimation based on "more traditional" feature selection methods.

Several screening rules for continuous penalization methods have been introduced

in the literature (Ghaoui et al., 2010; Wang et al., 2015), and they have recently been embedded within MIP-based techniques for best subset selection (Atamturk and Gómez, 2020; Deza and Atamturk, 2022). We focus on the *sure independence screening* (SIS) developed by Fan and Lv (2008) which uses marginal correlations to reduce the model dimension. Namely, assuming that $\log p = \mathcal{O}(n^{\xi})$ for some $\xi > 0$ and that each predictor has been standardized, SIS retains $q < n \ll p$ features based on $\widehat{\mathcal{S}}_{\alpha} = \{1 \leq j \leq p : |\tau_j| \geq |\tau_{\lfloor \alpha n \rfloor}|_{1:p}\}$, where $\boldsymbol{\tau} = \boldsymbol{X}^T \boldsymbol{y}$ denotes the marginal regression coefficients, and $|\tau_{\lfloor \alpha n \rfloor}|_{1:p}$ for $\alpha \in (0, 1)$ is the $\lfloor \alpha n \rfloor$-th term ($\lfloor \cdot \rfloor$ is the floor function) of the absolute coefficients $|\tau_j|$ sorted in in an ascending way. Indicating the true set of relevant features as $\mathcal{S}_{\beta}$, Fan and Lv (2008) provided the conditions for the *sure screening property* of SIS, i.e., $\Pr(\mathcal{S}_{\beta} \subset \widehat{\mathcal{S}}_{\alpha}) \to 1$ as $n \to \infty$, for a given α value. This guarantees that all the relevant features are retained by the screening procedure with a probability tending to one. Once that the sub-model $\widehat{\mathcal{S}}_{\alpha}$ of size comparable with n has been selected (i.e., $|\widehat{\mathcal{S}}_{\alpha}| = q = o(n)$), one can rely on penalization methods such as SCAD or adaptive lasso to perform sparse estimation on the q-dimensional coefficient vector. This approach has also been extended to broader settings, such as GLMs (Fan and Song, 2010) and nonparametric regression (Fan et al., 2011). On the other hand, the performance of SIS deteriorates as the predictors have strong dependence structures. To mitigate this problem, several extensions of SIS have been developed, such as *iterative SIS* (Fan and Lv, 2008), *factor profiled SIS* (FPSIS; Wang 2012), and *conditional SIS* (Barut et al., 2016).

Importantly, the existence of outliers has a significant impact on SIS-based methods, which prompted various researchers to investigate some robust counterparts. For instance, *robust rank correlation screening* uses Kendall's τ as opposed to sample correlation (Li et al., 2012), and *trimmed SIS-SCAD* replaces MLE with hard trimming estimators (Neykov et al., 2014) – see also the more recent approach proposed by Wang et al. (2018). However, similarly to SIS, these procedures suffer in the presence of strong multicollinearity. To cope with this issue, Wang and Van Aelst (2019) proposed a robust counterpart of FPSIS, where both the estimation of latent factors and the screening procedure rely on robust methods. This requires that correlations in the predictors can be fully modeled by only a few latent factors, and

assumes a case-wise contamination mechanism.

In the future, it would be of great interest to develop SIS-based procedures that can mitigate high levels of multicollinearity and tolerate the presence of case-wise and cell-wise outliers. Interestingly, the presence of cell-wise outliers can be tackled with some form of marginal screening (i.e., univariate outlier detection), where the estimated outlying cells are imputed through techniques developed for the analysis of missing data (Rousseeuw and Bossche, 2018). Similar strategies may be embedded into SIS itself, and it would be of interest to exploit and study the connections across these domains.

Complex data structures Structured, complex data such as longitudinal measurements are often of interest in high-dimensional studies. However, feature selection techniques cannot typically cope with strongly correlated and ultra-high dimensional features (e.g., p in the hundreds of thousands or millions).

Functional data analysis is very effective in exploiting the richness of information when one of the variables or units of interest can be naturally viewed as a smooth function (e.g., time series, stochastic processes, density functions, etc.). Functional data are generally evaluated over a discrete grid and then smoothed into curves, and this field is receiving an increasing attention across domains (Morris, 2015; Wang et al., 2016). Indeed, the statistical analysis of infinitely dimensional objects (in isolation or in combination with other vector or scalar objects) can provide invaluable insights and eager the extraction of complex, weak signals that are more difficult to detect using standard statistical tools. In full generality, functional regression considers the response variable and/or the predictors as a curve or trajectory – i.e., function-on-function, function-on-scalar, and scalar-on-function regression models (Ramsay and Silverman, 2005; Kokoszka and Reimherr, 2017).

In the presence of non-sparse signals, robust estimation methods were developed by Gervini (2008), Maronna and Yohai (2013) and Kalogridis and Van Aelst (2019) for the functional linear model, and by Denhere and Billor (2016) for the functional logistic model. Feature selection based on continuous penalties have also been extended to functional regression settings, especially for the class of linear

models; see for instance the review in Aneiros et al. (2019). Notably, robust model selection techniques for functional linear models have recently been developed for scalar-on-function (Pannu and Billor, 2020), function-on-scalar (Cai et al., 2022), and function-on-function regressions (Cai et al., 2021).

It would be of great interest to extend our proposals to functional regression models. Here one can distinguish two different forms of data contamination, namely, magnitude and shape outliers, which denote data points outlying within a curve or entire outlying curves, respectively (Dai et al., 2020). They closely resemble the MSOM and VIOM, and shape outliers are typically more difficult to detect similarly to MSOMs. In this setting, one can envision the use of MIP techniques to tackle the problem of robust model selection. Group constraints can be easily embedded within a discrete optimization framework to remove all points belonging to a given curve, and this also reduces the computational burden for MIP. Moreover, the flexibility of MIP allows one not only to select relevant features and detect outlying cases, but also choosing a suitable level of smoothness for functional regression coefficients – which is often fixed a priori based on a certain number of smooth functions that constitute a basis expansion of functional coefficient estimates.

Appendices

Appendix A

Supplementary Material to Chapter 3

A.1 Theoretical results

In this Appendix we provide proofs for the theoretical results discussed in Section 3.3.3.

Proof of Proposition 3.1. For the considered loss function $\rho(\cdot)$, with an additional ridge penalty and L_0 constraints on $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$, any feasible solution behaves similarly to the unpenalized OLS case – whose proof is usually based on the Sherman–Morrison formula (Atkinson, 1985; Chatterjee and Hadi, 1988). Indeed, the additional k_n degrees of freedom in (3.2) are used to zero-out the largest transformed residuals as in (3.3). We can write (3.2) as

$$\begin{aligned}
\min_{\substack{\|\boldsymbol{\beta}\|_0 \leq k_p \\ \|\boldsymbol{\beta}\|_2^2 \leq \lambda \\ \|\boldsymbol{\phi}\|_0 \leq k_n}} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - \phi_i) &= \min_{\substack{\|\boldsymbol{\beta}\|_0 \leq k_p \\ \|\boldsymbol{\beta}\|_2^2 \leq \lambda}} \min_{\|\boldsymbol{\phi}\|_0 \leq k_n} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - \phi_i) \\
&= \min_{\substack{\|\boldsymbol{\beta}\|_0 \leq k_p \\ \|\boldsymbol{\beta}\|_2^2 \leq \lambda}} \min_{\|\boldsymbol{\phi}\|_0 \leq k_n} \frac{1}{n} \sum_{i=1}^{n} \rho(e_i - \phi_i) \\
&= \min_{\substack{\|\boldsymbol{\beta}\|_0 \leq k_p \\ \|\boldsymbol{\beta}\|_2^2 \leq \lambda}} T(\boldsymbol{\beta}).
\end{aligned}$$

Then for fixed and feasible $\boldsymbol{\beta}$, we evaluate $T(\boldsymbol{\beta})$ before performing the outer mini-

mization. Since $\rho(\boldsymbol{e}) \geq 0$ and $\rho(\boldsymbol{e}) = 0$ at $\boldsymbol{e} = \mathbf{0}$, each $\rho(e_i - \phi_i)$ in $T(\boldsymbol{\beta})$ is minimized at $\phi_i = e_i$. However, only at most k_n of the $\rho(e_i - \phi_i)$ with $\rho(e_i) \neq 0$ can achieve this minimum as $n - k_n$ or more of ϕ_i must be set to 0. Hence, it follows that $T(\boldsymbol{\beta}) = n^{-1}[\sum_{i=1}^{n-k_n}\{\rho(e_i - 0)\}_{i:n} + \sum_{i=n-k_n+1}^{n}\{\rho(e_i - e_i)\}_{i:n}] = n^{-1}\sum_{i=1}^{n-k_n}\{\rho(e_i)\}_{i:n}$. ■

Proof of Proposition 3.2. Our approach is similar to Alfons et al. (2013) and Bertsimas and Mazumder (2014) and is not affected by the presence of an L_0 constraint on $\boldsymbol{\beta}$. For the first part of the proof, consider the presence of $n_0 \leq k_n = n - h$ outliers, so that $n - n_0 \geq h$ units are non-outlying. Let $\widetilde{\boldsymbol{Z}} = [\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$ be the contaminated sample and $M_y = \max_{i=1,\ldots,n} |y_i|$. If $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ (possibly excluding the intercept term), the corresponding loss $Q(\widehat{\boldsymbol{\beta}})$ in (3.2) satisfies $Q(\mathbf{0}) = \sum_{i=1}^{h}\{\rho(\widetilde{\boldsymbol{y}})\}_{i:n} \leq \sum_{i=1}^{h}\{\rho(\boldsymbol{y})\}_{i:n} \leq h\rho(M_y)$, where the first inequality follows from the fact that the contamination is arbitrary but not necessarily adversarial, and we also used the result in Theorem 3.1. Take any other estimate $\widehat{\boldsymbol{\beta}}$ such that $\|\widehat{\boldsymbol{\beta}}\|_2^2 \geq l_1$, where $l_1 = (h\rho(M_y) + 1)/\lambda^*$ is independent from the contamination structure. It follows that $Q(\widehat{\boldsymbol{\beta}}) \geq \lambda^*\|\boldsymbol{\beta}\|_2^2 \geq h\rho(M_y) + 1 > Q(\mathbf{0})$, leading to a contradiction since $Q(\mathbf{0}) \geq Q(\widehat{\boldsymbol{\beta}})$ (i.e., the objective is non-decreasing in the number of active features). Thus, $\|\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{Z}})\|_2^2 \leq l_1$ shows that the MIP estimator in (3.2) does not breakdown for $n_0 \leq k_n$ (i.e., $\epsilon^* \geq (k_n + 1)/n$).

For the second part of the proof, we need to show that $\epsilon^* \leq (k_n + 1)/n$. We assume that the estimator does not breakdown, so that $\|\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{Z}})\|_2^2 \leq u_1$. Let $n_0 > k_n$ and denote the corresponding contaminated sample as $\widetilde{\boldsymbol{Z}} = (\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}) = \boldsymbol{Z} + (\boldsymbol{\Delta}_y, \boldsymbol{\Delta}_X)$. It follows that

$$\begin{aligned} Q(\widehat{\boldsymbol{\beta}}) &= \sum_{i=1}^{n-n_0}\left\{\rho(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}})\right\}_{i=1:n} + \sum_{j=n-n_0+1}^{n-k_n}\left\{\rho(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}})\right\}_{j=1:n} + \lambda^*\|\widehat{\boldsymbol{\beta}}\|_2^2 \\ &\geq \left[\rho\left\{(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) + (\Delta_{y_i} - \boldsymbol{\Delta}_{X_i}^T\widehat{\boldsymbol{\beta}})\right\}\right]_{i=n-n_0+1} + \lambda^*\|\widehat{\boldsymbol{\beta}}\|_2^2 \end{aligned} \tag{A.1}$$

since at least one outlier is included in the fit; namely, the unit corresponding to the $(n - n_0 + 1)$-th position of the ordered transformed residuals. Thus, the possible unboundedness of (A.1), as both terms Δ_{y_i} and $\boldsymbol{\Delta}_{X_i}$ may take arbitrarily large

values, contradicts the assumption that $\|\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{Z}})\|_2^2 \leq u_1$. ■

Proof of Proposition 3.3. The following proof provides a necessary condition for any method to achieve SFSOD consistency. This has been proved for feature selection with an L_0 constraint (Shen et al., 2012, 2013), and also extended to the presence of group constraints (Xiang et al., 2015). Here we extend this further to account for the presence of, and identify, MSOM outliers. The main difference being that variable selection in (3.2) is performed on two disjoint sets of coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$. Similarly to Theorem 1 in Shen et al. (2013), we consider a least favorable scenario where Fano's inequality can be applied. Let $\mathcal{C}_\beta = \{\boldsymbol{\beta}_j\}_{j=0}^p$ be a collection of parameters with components equal to γ_β or 0 (e.g., one can think of $\boldsymbol{\beta}_0$ as being the true model). Similarly, define $\mathcal{C}_\phi = \{\boldsymbol{\phi}_i\}_{i=0}^n$ as a collection of parameters with components equal to γ_ϕ or 0. Assume also that $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}\|_2^2 \leq 4\gamma_\beta^2$ for any $j, j' \in 0, 1, \ldots, p$, and $\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{i'}\|_2^2 \leq 4\gamma_\phi^2$ for any $i, i' \in 0, 1, \ldots, n$. Let $\Delta_\beta^* = \min_{\boldsymbol{\beta}_0:|\boldsymbol{\beta}_0|\geq 1, \boldsymbol{\beta}_0 \in \mathcal{S}_\beta, |\mathcal{S}_\beta| \leq p_0} \Delta_\beta$ and $\Delta_\phi^* = \min_{\boldsymbol{\phi}_0:|\boldsymbol{\phi}_0|\geq 1, \boldsymbol{\phi}_0 \in \mathcal{S}_\phi, |\mathcal{S}_\phi| \leq n_0} \Delta_\phi$ such that $r_\beta = \max_{1\leq j\leq p} \|\boldsymbol{X}_j\|_2^2 (n\Delta_\beta^*)^{-1}$ and $r_\phi = (n\Delta_\phi^*)^{-1}$. For any $\boldsymbol{\beta}_j, \boldsymbol{\beta}_{j'} \in \mathcal{C}_\beta$ with densities $q(\boldsymbol{\beta}_j)$ and $q(\boldsymbol{\beta}_{j'})$, respectively, the corresponding Kullback–Leibler information is equal to $D[q(\boldsymbol{\beta}_j), q(\boldsymbol{\beta}_{j'})] = \|\boldsymbol{X}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'})\|_2^2 \leq 2\max_{1\leq k\leq p} \|\boldsymbol{X}_k\|_2^2 \gamma_\beta^2/(n\sigma^2) \leq 2r_\beta \Delta_\beta/\sigma^2$. Here the first bound is obtained using sub-additivity and the triangle inequality, and the second one is based on Lemma 1 in Shen et al. (2013). Similarly, $D[q(\boldsymbol{\phi}_i), q(\boldsymbol{\phi}_{i'})] = \|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{i'}\|_2^2 \leq 2\gamma_\phi^2/(n\sigma^2) \leq 2r_\phi \Delta_\phi/\sigma^2$. Thus, for any estimates $\boldsymbol{T}_\beta$ and $\boldsymbol{T}_\phi$, it follows from Fano's inequality that: $F_\beta = (p+1)^{-1} \sum_{j\in\mathcal{C}_\beta} P(\boldsymbol{T}_\beta = j) \leq (2nr_\beta\Delta_\beta + \sigma^2 \log 2)/(\sigma^2 \log(p))$ and $F_\phi = (n+1)^{-1} \sum_{i\in\mathcal{C}_\phi} P(\boldsymbol{T}_\phi = i) \leq (2nr_\phi\Delta_\phi + \sigma^2 \log 2)/(\sigma^2 \log(n))$. Using the fact that $P(\widehat{\mathcal{S}} \neq \mathcal{S}) = P\{(\widehat{\mathcal{S}}_\beta \neq \mathcal{S}_\beta) \cup (\widehat{\mathcal{S}}_\phi \neq \mathcal{S}_\phi)\} = 1 - P\{(\widehat{\mathcal{S}}_\beta = \mathcal{S}_\beta) \cap (\widehat{\mathcal{S}}_\phi = \mathcal{S}_\phi)\} \geq 1 - \min\{P(\widehat{\mathcal{S}}_\beta = \mathcal{S}_\beta), P(\widehat{\mathcal{S}}_\phi = \mathcal{S}_\phi)\}$ leads to the following lower bound:

$$\sup_{\{(\theta,A):\Delta_\theta \leq R^*\}} P(\widehat{\mathcal{S}} \neq \mathcal{S}) \geq 1 - \min(F_\beta, F_\phi), \tag{A.2}$$

where $R^* = \max\{(1-c_\beta^*)\sigma^2 \log(p)/(2nr_\beta), (1-c_\phi^*)\sigma^2 \log(n)/(2nr_\phi)\}$ and $c_\beta^*, c_\phi^* > 0$. For $\sup_{\theta_0 \in B_0(u,l)} (\widehat{\mathcal{S}} \neq \mathcal{S}) \to 0$ as in Theorem 3.3, it follows from (A.2) that the L_0-band $B(u,l)$ cannot interact with the L_0-ball $B(R^*, 0)$. Thus, a necessary condition for any estimator to achieve SFSOD consistency is that $l \geq$

$\sigma^2/n \max\{\log(p)/(4r_\beta), \log(n)/4r_\phi\}$; this provides a tighter bound compared to a naïve substitution of $p+n$ in place of p in Theorem 1 of Shen et al. (2013). ■

Proof of Proposition 3.4. The following result bounds the reconstruction error $P(\widehat{\boldsymbol{\theta}}_{L_0} \neq \widehat{\boldsymbol{\theta}}_0) \geq P(\widehat{\mathcal{S}}^{L_0} \neq \mathcal{S})$ and extends Theorem 2 in Shen et al. (2013) to the presence of MSOM outliers. Let $\overline{\mathcal{S}} \subset \{1, \ldots, (p+n)\}$ be any feasible estimate of the active set such that $\overline{\mathcal{S}}_\beta \neq \mathcal{S}_\beta$ and $\overline{\mathcal{S}}_\phi \neq \mathcal{S}_\phi$, with $|\overline{\mathcal{S}}_\beta| \leq p_0$ and $|\overline{\mathcal{S}}_\phi| \leq n_0$. Note that if $k_p = p_0$ and $k_n = n_0$, it follows that $|\widehat{\mathcal{S}}^{L_0}| = |\widehat{\mathcal{S}}^{L_0}_\beta| + |\widehat{\mathcal{S}}^{L_0}_\phi| \leq p_0 + n_0$. To simplify the notation, take $L(\boldsymbol{\theta}, \mathcal{S}_\beta, \mathcal{S}_\phi) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{S}_\beta}\boldsymbol{\beta}_{\mathcal{S}_\beta} - \boldsymbol{I}_{n_{\mathcal{S}_\phi}}\boldsymbol{\phi}_{\mathcal{S}_\phi}\|_2^2$. Partitioning $\overline{\mathcal{S}}$ as $(\overline{\mathcal{S}}_\beta \backslash \mathcal{S}_\beta) \cup (\overline{\mathcal{S}}_\beta \cap \mathcal{S}_\beta) \cup (\overline{\mathcal{S}}_\phi \backslash \mathcal{S}_\phi) \cup (\overline{\mathcal{S}}_\phi \cap \mathcal{S}_\phi)$, it follows that:

$$
\begin{aligned}
P\left(\widehat{\boldsymbol{\theta}}_{L_0} \neq \widehat{\boldsymbol{\theta}}_0\right) \leq & P\left(\widehat{\boldsymbol{\beta}}_{L_0} \neq \widehat{\boldsymbol{\beta}}_0\right) + P\left(\widehat{\boldsymbol{\phi}}_{L_0} \neq \widehat{\boldsymbol{\phi}}_0\right) \\
& \leq \sum_{\widehat{\mathcal{S}}^{L_0}_\beta \in \overline{\mathcal{S}}_\beta} P\left(L(\boldsymbol{\beta}, \widehat{\mathcal{S}}^{L_0}_\beta) - L(\boldsymbol{\beta}_0, \mathcal{S}_\beta) \leq 0\right) \\
& \quad + \sum_{\widehat{\mathcal{S}}^{L_0}_\phi \in \overline{\mathcal{S}}_\phi} P\left(L(\boldsymbol{\phi}, \widehat{\mathcal{S}}^{L_0}_\phi) - L(\boldsymbol{\phi}_0, \mathcal{S}_\phi) \leq 0\right) \\
& \leq \sum_{i_p=0}^{p_0-1} \sum_{j_p=0}^{p_0-i_p} \binom{p-p_0}{j_p} \binom{p_0}{i_p} P\left(L(\boldsymbol{\beta}, \widehat{\mathcal{S}}^{L_0}_\beta = \overline{\mathcal{S}}_\beta(i_p, j_p)) - L(\boldsymbol{\beta}_0, \mathcal{S}_\beta) \leq 0\right) \\
& \quad + \sum_{i_n=0}^{n_0-1} \sum_{j_n=0}^{n_0-i_n} \binom{n-n_0}{j_n} \binom{n_0}{i_n} P\left(L(\boldsymbol{\phi}, \widehat{\mathcal{S}}^{L_0}_\phi = \overline{\mathcal{S}}_\phi(i_n, j_n)) - L(\boldsymbol{\phi}_0, \mathcal{S}_\phi) \leq 0\right),
\end{aligned}
$$

where the first inequality follows from the union bound, the second inequality uses the probability of each feasible solution, and the third upper bound is based on the total number of possible solutions for a given size of correct (i_p and i_n) and incorrect (j_p and j_n) selections for SFSOD. Following the argument in Shen et al. (2013) separately on the two terms $P(\widehat{\boldsymbol{\beta}}_{L_0} \neq \widehat{\boldsymbol{\beta}}_0)$ and $P(\widehat{\boldsymbol{\phi}}_{L_0} \neq \widehat{\boldsymbol{\phi}}_0)$ leads to the

following result:

$$
\begin{aligned}
P\left(\widehat{\boldsymbol{\theta}}_{L_0} \neq \widehat{\boldsymbol{\theta}}_0\right) \leq & 2 \sum_{i_p=1}^{p_0} \sum_{j_p=0}^{i_p} (p-p_0)^{j_p} p_0^{i_p} \exp\left(-\frac{i_p}{18\sigma^2} n\Delta_\beta + \frac{2}{3} j_p\right) \\
& + 2 \sum_{i_n=1}^{n_0} \sum_{j_n=0}^{i_n} (n-n_0)^{j_n} n_0^{i_n} \exp\left(-\frac{i_n}{18\sigma^2} n\Delta_\phi + \frac{2}{3} j_n\right) \\
\leq & \frac{2e}{e-1}\left(R\left[\exp\left\{-\frac{n}{18\sigma^2}\left(\Delta_\beta - 36\frac{\log p}{n}\sigma^2\right)\right\}\right]\right. \\
& \left. + R\left[\exp\left\{-\frac{n}{18\sigma^2}\left(\Delta_\phi - 36\frac{\log n}{n}\sigma^2\right)\right\}\right]\right) \\
\leq & \frac{4e}{e-1} \max\left(R\left[\exp\left\{-\frac{n}{18\sigma^2}\left(\Delta_\beta - 36\frac{\log p}{n}\sigma^2\right)\right\}\right],\right. \\
& \left. R\left[\exp\left\{-\frac{n}{18\sigma^2}\left(\Delta_\phi - 36\frac{\log n}{n}\sigma^2\right)\right\}\right]\right),
\end{aligned}
$$

where $R(x) = x/(1-x)$. Using the fact that $P(\widehat{\boldsymbol{\theta}}_{L_0} \neq \widehat{\boldsymbol{\theta}}_0) \leq 1$ establishes the result in (3.5); this is a tighter bound compared to the naïve extension of Theorem 2 in Shen et al. (2013) using $p+n$ in place of p. ■

Proof of Proposition 3.5. The result in Theorem 3.5(1) immediately follows from Theorem 3.4 through a pointwise bound of (3.5) to $\boldsymbol{\theta}_0 \in B(u_\theta, l_\theta)$. For Theorem 3.5(2) our approach is similar to Liu and Yu (2013) and Zhu et al. (2020). Theorem 3.5(1) guarantees that $P(\widehat{\boldsymbol{\theta}}_{L_0} = \widehat{\boldsymbol{\theta}}_0) \to 1$ as $(n,p) \to \infty$. Therefore, with a probability tending to one, it follows that:

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}}_{L_0} &= (\boldsymbol{A}_{\widehat{\mathcal{S}}^{L_0}}^T \boldsymbol{A}_{\widehat{\mathcal{S}}^{L_0}})^{-1} \boldsymbol{A}_{\widehat{\mathcal{S}}^{L_0}} \boldsymbol{y} \\
&= (\boldsymbol{A}_{\widehat{\mathcal{S}}^{L_0}}^T \boldsymbol{A}_{\widehat{\mathcal{S}}^{L_0}})^{-1} \boldsymbol{A}_{\widehat{\mathcal{S}}^{L_0}} (\boldsymbol{A}_{\mathcal{S}} \boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}) \\
&= \boldsymbol{\theta}_0 + (\boldsymbol{A}_{\mathcal{S}}^T \boldsymbol{A}_{\mathcal{S}})^{-1} \boldsymbol{A}_{\mathcal{S}} \boldsymbol{\varepsilon}.
\end{aligned}
$$

Using the moment generating function with the fact that $\varepsilon_i \sim N(0, \sigma^2)$ for $i \notin \mathcal{S}_\phi$ leads to $(\boldsymbol{A}_{\mathcal{S}}^T \boldsymbol{A}_{\mathcal{S}})^{-1} \boldsymbol{A}_{\mathcal{S}} \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 (\boldsymbol{A}_{\mathcal{S}}^T \boldsymbol{A}_{\mathcal{S}})^{-1})$. Consequently, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{L_0} - \boldsymbol{\theta}_0) \to^d N(\boldsymbol{0}, \boldsymbol{\Sigma}_\theta)$. ■

Proof of Proposition 3.6. Both results immediately follow from Theorem 2 in Shen et al. (2013) considering our SFSOD formulation based on the two disjoint sets $\mathcal{S}_\beta$

and $\mathcal{S}_\phi$. ■

A.2 Simulation study details

In this Appendix we provide details and extensions of the simulation study in Section 3.4. Computations for this research were performed on the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer. This content is solely the responsibility of the authors and does not necessarily represent the views of the Institute for Computational and Data Sciences. We used basic memory option on the ACI-B cluster with an Intel Xeon 24 core processor at 2.2 GHz and 128 GB of RAM. The multi-thread option in `R` and `Gurobi` was limited to a maximum of 24 threads.

We compared the following estimators:

SparseLTS: combines an L_1-penalty with the LTS estimator (Alfons et al., 2013). Similarly to other methods, we do not perform a final re-weighting step. The algorithm starts with 1000 initial subsamples, where 20 subsamples with the lowest value of the objective function are used to compute additional concentration-steps until convergence. The sparsity level is tuned according to the BIC-type criterion proposed by the authors. Our implementation is based on the parallelized `sparseLTS` function of the `robustHD` package (Alfons, 2021) in `R` (we use the `R` version 3.5.2).

EnetLTS: combines an elastic net penalty with the LTS loss function (Kurnaz et al., 2017). Also here we use a lasso penalty and the algorithm starts with 1000 initial subsamples, where 20 subsamples with the lowest value of the objective function are used to compute additional concentration-steps until convergence. The sparsity level is tuned through a robust 10-folds cross-validation as advocated by the authors. It is implemented using the `enetLTS` `R` package (Kurnaz et al., 2018) without parallelization.

MIP: solves (3.2) based on the L_2-loss and excludes the ridge-like penalty. As customary in feature selection problems, we do not penalize the intercept and we

standardize features at the outset. Because the regressions we focus on comprise outliers, we use a robust standardization; both $\boldsymbol{y}$ and $\boldsymbol{X}$ are centered to have zero medians, and each $\boldsymbol{X}_j$ is also scaled to have unit median absolute deviation (MAD); results in the output are given in the original scale. Although binary features are often standardized (Tibshirani, 1997), since the constraints on $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are separate, we do not standardize the binary variables used in the outlier detection component of the problem. The interpretation of the entries in $\boldsymbol{\phi}$ as prediction residuals indicates that they are already on the same scale under the null model. The sparsity level k_p^* ranges from 1 (only the intercept term) to $2p_0$ and is tuned through a BIC-type criterion (see also Section 3.3.2). This is computed as: $\text{BIC}(k_p^*) = k_p^* \log(h) + h \log(L)$, where $h = n - n_0$ and $L = h^{-1}\|\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{\theta}}_{L_0}\|_2^2$. Instead of taking the minimum BIC, we aim at finding an elbow across the considered k_p^* values. As a simple approach, our final solution is the one with the largest absolute decrease along consecutive model sizes, i.e., $k_p = \min_{k_p^*}\{\text{BIC}(k_p^*) - \text{BIC}(k_p^* - 1)\}$. For each k_p^* value, the corresponding MIQP is warm-started using the result from the previous model of size $k_p^* - 1$. Big-$\mathcal{M}$ bounds are computed using the ensemble method described in Section 3.3.2, including all estimators used in our comparison (apart from the oracle). Here the multiplicative constant c is set to 10. Our implementation is based on the `Julia` programming language (version 0.6.0) in connection with the MIP commercial solver `Gurobi` (version 8.1.1) through the `JuMP` package. Our code can run in parallel and is provided in Appendix D. Each job runs with a scheduled time limit of 300 seconds.

We also provide an implementation for the tuning procedure described in Section 3.3.2. For cross-validation in (iii) we use the computationally efficient integrated scheme introduced in Kenney et al. (2021), robustifying the performance measure (the mean squared prediction error) with an upper trimmed sum. Choosing the trimming level is again not trivial, because cross-validation folds might contain different proportions of outliers. In order to be conservative, we fix the trimming proportion to $3k_n/n$ on the test fold and to $2k_n/n$ on the training folds. For information criteria in (iii) the situation is more straightforward, as one can compute robust values for them using only cases identified as non-outlying in any given MIP

run. Refining k_n downward in (iv) improves efficiency in estimating $\boldsymbol{\beta}$, which can be low if the starting k_n in (ii) is substantially larger than the true n_0, excluding non-outlying cases from the fit. Assuming that the selected model of size k_p in (iii) is close to the true p_0 active features, iteratively reducing k_n provides an effective strategy to pinpoint when outliers start to be included in the fit. Similarly to the Forward Search algorithm (Atkinson and Riani, 2000), this can be done monitoring an appropriate statistic (e.g., the minimum absolute deletion residuals) along iterations.

Table A.1 shows our results in terms of medians and MADs for the simulation setting discussed in Section 3.4. A comparison with Table 3.1 highlights the skewness of most metrics, with all methods performing better, especially our proposal (denoted MIP).

We also explored weak SNR scenarios. The following simulation setting is the same as in Section 3.4, with the only difference being that the signal-to-noise-ratio is reduced to SNR = 3. Table A.2 shows simulation results in term of medians and MADs. A comparison with Table 3.1 shows that similar conclusions hold, although all methods experience an overall decrease in performance. Our approach generally outperforms other methods and converges faster to the oracle solution. However, for scenarios with small sample sizes, the FNR in $\widehat{\boldsymbol{\beta}}$ for our method is worse. Moreover, while computing time for heuristic methods remains similar to the stronger SNR scenario, our proposal shows a marked increase.

We also explored simulation settings with multicollinearity structures. Table A.3 presents our results for a simulation scenario which mimics that of Section 3.4 (reporting medians and MADs), with the only difference being that $\boldsymbol{\Sigma}_X$ has an autoregressive correlation structure $\boldsymbol{\Sigma}_{X,ij} = 0.3^{|i-j|}$. Though this could be considered a "mild" level of correlation, we note that the addition of contamination increases the amount of multicollinearity present. Here our approach is often outperformed by other methods for small sample sizes, however as the latter increase we can again notice that our proposal converges faster to the oracle solution and results like those in Table 3.1 hold.

Table A.1: Median (MAD in parenthesis) of RMSPE, variance and squared bias for $\widehat{\boldsymbol{\beta}}$, FPR and FNR for $\widehat{\boldsymbol{\beta}}$ and outlier detection (as well as the corresponding F_1 scores), and computing time, based on 1000 simulation replications for the simulation setting in Section 3.4.

n	p	Method	RMSPE	var($\widehat{\beta}$)	bias($\widehat{\beta}$)^2	FPR($\widehat{\beta}$)	FNR($\widehat{\beta}$)	$F_1(\widehat{\beta})$	FPR($\widehat{\phi}$)	FNR($\widehat{\phi}$)	$F_1(\widehat{\phi})$	Time
50	50	Oracle	1.85(0.27)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	2.31(0.36)	0.05(0.01)	0.03(0.01)	0.11(0.07)	0.00(0.00)	0.95	0.00(0.00)	0.00(0.00)	1	12.10(0.67)
		SparseLTS	2.42(0.41)	0.06(0.01)	0.01(0.00)	0.53(0.07)	0.00(0.00)	0.79	0.00(0.00)	0.00(0.00)	1	3.46(0.78)
		MIP	1.91(0.40)	0.04(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	7.33(3.29)
100	50	Oracle	1.82(0.18)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	1.99(0.20)	0.01(0.00)	0.00(0.00)	0.22(0.16)	0.00(0.00)	0.90	0.00(0.00)	0.00(0.00)	1	9.69(0.31)
		SparseLTS	2.11(0.22)	0.03(0.00)	0.00(0.00)	0.67(0.07)	0.00(0.00)	0.75	0.00(0.00)	0.00(0.00)	1	4.05(0.74)
		MIP	1.83(0.19)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	29.34(11.80)
150	50	Oracle	1.81(0.15)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	1.92(0.16)	0.01(0.00)	0.00(0.00)	0.33(0.23)	0.00(0.00)	0.86	0.00(0.00)	0.00(0.00)	1	10.19(0.33)
		SparseLTS	1.99(0.17)	0.02(0.00)	0.00(0.00)	0.69(0.07)	0.00(0.00)	0.74	0.00(0.00)	0.00(0.00)	1	4.06(0.83)
		MIP	1.81(0.15)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	381.52(294.70)
50	200	Oracle	1.86(0.26)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	2.84(0.77)	0.03(0.00)	0.02(0.00)	0.18(0.05)	0.00(0.00)	0.92	0.00(0.00)	0.00(0.00)	1	36.06(3.20)
		SparseLTS	2.64(0.48)	0.02(0.00)	0.01(0.00)	0.16(0.02)	0.00(0.00)	0.92	0.00(0.00)	0.00(0.00)	1	3.70(0.94)
		MIP	1.96(0.46)	0.02(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	6.99(5.14)
100	200	Oracle	1.84(0.18)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	2.28(0.36)	0.02(0.00)	0.01(0.00)	0.25(0.10)	0.00(0.00)	0.89	0.00(0.00)	0.00(0.00)	1	45.21(3.13)
		SparseLTS	2.32(0.25)	0.01(0.00)	0.00(0.00)	0.31(0.02)	0.00(0.00)	0.86	0.00(0.00)	0.00(0.00)	1	10.10(2.42)
		MIP	1.85(0.19)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	64.92(26.67)
150	200	Oracle	1.81(0.14)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	2.03(0.22)	0.02(0.00)	0.00(0.00)	0.21(0.13)	0.00(0.00)	0.91	0.01(0.01)	0.00(0.00)	1	47.96(2.98)
		SparseLTS	2.24(0.19)	0.01(0.00)	0.00(0.00)	0.42(0.04)	0.00(0.00)	0.83	0.00(0.00)	0.00(0.00)	1	13.84(2.27)
		MIP	1.81(0.15)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	518.92(393.54)

Table A.2: Median (MAD in parenthesis) of RMSPE, variance and squared bias for $\widehat{\boldsymbol{\beta}}$, FPR and FNR for $\widehat{\boldsymbol{\beta}}$ and outlier detection (as well as the corresponding F_1 scores), and computing time, for 100 simulation replications similarly to Section 3.4 with SNR = 3.

n	p	Method	RMSPE	var($\widehat{\beta}$)	bias($\widehat{\beta}$)^2	FPR($\widehat{\beta}$)	FNR($\widehat{\beta}$)	$F_1(\widehat{\beta})$	FPR($\widehat{\phi}$)	FNR($\widehat{\phi}$)	$F_1(\widehat{\phi})$	Time
50	50	Oracle	2.49(0.34)	0.02(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	3.09(0.73)	0.14(0.00)	0.06(0.00)	0.18(0.13)	0.00(0.00)	0.92	0.00(0.00)	0.00(0.00)	1.00	14.76(0.81)
		SparseLTS	3.30(0.57)	0.15(0.00)	0.01(0.00)	0.58(0.07)	0.00(0.00)	0.78	0.00(0.00)	0.00(0.00)	1.00	2.56(0.72)
		MIP	2.93(0.99)	0.10(0.00)	0.02(0.00)	0.00(0.00)	0.20(0.30)	0.89	0.00(0.00)	0.00(0.00)	1.00	16.17(9.01)
100	50	Oracle	2.42(0.20)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	2.58(0.26)	0.02(0.00)	0.00(0.00)	0.30(0.16)	0.00(0.00)	0.87	0.00(0.00)	0.00(0.00)	1.00	12.30(0.21)
		SparseLTS	2.82(0.27)	0.05(0.00)	0.00(0.00)	0.71(0.10)	0.00(0.00)	0.74	0.00(0.00)	0.00(0.00)	1.00	2.40(0.26)
		MIP	2.48(0.30)	0.02(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	60.97(38.33)
150	50	Oracle	2.29(0.16)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	2.48(0.23)	0.02(0.00)	0.00(0.00)	0.50(0.25)	0.00(0.00)	0.80	0.00(0.00)	0.00(0.00)	1.00	12.82(0.18)
		SparseLTS	2.57(0.21)	0.03(0.00)	0.00(0.00)	0.76(0.07)	0.00(0.00)	0.73	0.00(0.00)	0.00(0.00)	1.00	4.05(1.26)
		MIP	2.31(0.21)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	751.73(302.70)
50	200	Oracle	2.44(0.31)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	3.93(1.28)	0.04(0.00)	0.03(0.00)	0.18(0.05)	0.00(0.00)	0.92	0.01(0.02)	0.00(0.00)	0.99	45.94(2.43)
		SparseLTS	3.67(1.06)	0.04(0.00)	0.03(0.00)	0.17(0.02)	0.00(0.00)	0.92	0.00(0.00)	0.00(0.00)	1.00	4.22(1.19)
		MIP	3.30(1.51)	0.05(0.00)	0.01(0.00)	0.00(0.00)	0.20(0.30)	0.89	0.00(0.00)	0.00(0.00)	1.00	18.76(17.94)
100	200	Oracle	2.40(0.27)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	3.18(0.65)	0.03(0.00)	0.01(0.00)	0.30(0.10)	0.00(0.00)	0.87	0.01(0.02)	0.00(0.00)	0.99	55.32(3.40)
		SparseLTS	3.09(0.30)	0.02(0.00)	0.00(0.00)	0.33(0.02)	0.00(0.00)	0.86	0.00(0.00)	0.00(0.00)	1.00	13.58(4.58)
		MIP	2.46(0.41)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	107.57(74.31)
150	200	Oracle	2.37(0.21)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
		EnetLTS	2.76(0.31)	0.02(0.00)	0.01(0.00)	0.28(0.14)	0.00(0.00)	0.88	0.00(0.00)	0.00(0.00)	1.00	56.74(3.69)
		SparseLTS	2.98(0.34)	0.02(0.00)	0.00(0.00)	0.45(0.02)	0.00(0.00)	0.82	0.00(0.00)	0.00(0.00)	1.00	12.60(0.89)
		MIP	2.39(0.22)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	840.34(519.30)

Smaller SNR regimes and stronger correlation structures were also explored; results are not reported as all methods performed quite poorly. In these settings, as advocated in Hastie et al. (2020), a ridge-like penalty may be beneficial.

In order to further investigate the effects of mean shift parameters, we performed additional simulations and extended the results reported in Table 3.1 of the main text. For simplicity, we focus on the scenario with $n = 50$ and $p = 50$, where $n_0 = 5$ and $p_0 = 5$. In this setting, we use mean shifts $-\lambda_\varepsilon = \lambda_X \in \{1, 3, 5, 7, 9\}$ and replicate the analysis 100 independent times. Table A.4 shows that for all methods performance deteriorates as mean shifts decrease in magnitude, but that our proposal generally outperforms competitors across different metrics. In particular, outlier detection accuracy (based on FPR and FNR for $\widehat{\boldsymbol{\phi}}$, as well as their F_1 scores) drops for mean shifts of size 1, which leads to an overall decrease in performance. This is consistent with our Theorem 3.3; the minimal degree of separation between the true model and a least favorable model decreases with the size of the mean shifts, resulting in harder problems, and possibly in a failure of the necessary condition for SFSOD consistency. Note also that, since our MIP proposal (unlike other algorithms) aims for a global optimum, its expected computing will tend to increase as the size of the mean shifts decreases – making the problem harder (see Kenney et al. 2021). In contrast, computing time for heuristic methods remains quite stable.

A.3 Microbiome application details

This Appendix provides details of our microbiome application in Section 3.5. The analysis was performed through the same high-performance computing infrastructure as our simulations. We used an Intel Xeon 24 core processor at 2.2 GHz and 128 GB of RAM. The multi-thread option in `R` and `Gurobi` was limited to a maximum of 24 threads.

All competing robust methods were tuned as described in Appendix A.2, for the non-robust lasso we tuned via 10-fold cross-validation across a grid of at most 100 tuning parameters. It is implemented using the `cv.glmnet` function within the `glmnet` `R` package (Friedman et al., 2010). For our proposal, we selected the sparsity

Table A.3: Median (MAD in parenthesis) of RMSPE, variance and squared bias for $\widehat{\boldsymbol{\beta}}$, FPR and FNR for $\widehat{\boldsymbol{\beta}}$ and outlier detection (as well as the corresponding F_1 scores), and computing time, for 100 simulation replications similarly to Section 3.4 in presence of multicollinearity.

n	p	Method	RMSPE	var($\widehat{\beta}$)	bias($\widehat{\beta}$)^2	FPR($\widehat{\beta}$)	FNR($\widehat{\beta}$)	$F_1(\widehat{\beta})$	FPR($\widehat{\phi}$)	FNR($\widehat{\phi}$)	$F_1(\widehat{\phi})$	Time
50	50	Oracle	2.34(0.26)	0.02(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	2.75(0.36)	0.05(0.00)	0.02(0.00)	0.07(0.07)	0.00(0.00)	0.97	0.00(0.00)	0.00(0.00)	1	14.71(0.55)
		SparseLTS	2.92(0.49)	0.08(0.00)	0.00(0.00)	0.51(0.07)	0.00(0.00)	0.80	0.00(0.00)	0.00(0.00)	1	1.39(0.22)
		MIP	3.00(0.59)	0.09(0.00)	0.00(0.00)	0.11(0.03)	0.00(0.00)	0.95	0.00(0.00)	0.00(0.00)	1	9.05(2.67)
100	50	Oracle	2.30(0.21)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	2.43(0.23)	0.01(0.00)	0.01(0.00)	0.09(0.10)	0.00(0.00)	0.96	0.00(0.00)	0.00(0.00)	1	12.63(0.18)
		SparseLTS	2.57(0.25)	0.03(0.00)	0.00(0.00)	0.56(0.10)	0.00(0.00)	0.78	0.00(0.00)	0.00(0.00)	1	1.77(0.20)
		MIP	2.46(0.26)	0.02(0.00)	0.00(0.00)	0.04(0.03)	0.00(0.00)	0.98	0.00(0.00)	0.00(0.00)	1	26.37(8.51)
150	50	Oracle	2.21(0.17)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	2.36(0.17)	0.01(0.00)	0.00(0.00)	0.18(0.20)	0.00(0.00)	0.92	0.00(0.00)	0.00(0.00)	1	13.13(0.16)
		SparseLTS	2.38(0.19)	0.02(0.00)	0.00(0.00)	0.54(0.05)	0.00(0.00)	0.79	0.00(0.00)	0.00(0.00)	1	3.23(0.82)
		MIP	2.31(0.20)	0.01(0.00)	0.00(0.00)	0.02(0.03)	0.00(0.00)	0.99	0.00(0.00)	0.00(0.00)	1	340.60(175.80)
50	200	Oracle	2.33(0.33)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	3.19(0.96)	0.04(0.00)	0.01(0.00)	0.16(0.05)	0.00(0.00)	0.93	0.00(0.00)	0.00(0.00)	1	44.68(2.70)
		SparseLTS	3.01(0.55)	0.02(0.00)	0.00(0.00)	0.16(0.02)	0.00(0.00)	0.92	0.00(0.00)	0.00(0.00)	1	3.48(0.82)
		MIP	3.01(0.74)	0.03(0.00)	0.00(0.00)	0.02(0.01)	0.00(0.00)	0.99	0.00(0.00)	0.00(0.00)	1	8.30(5.25)
100	200	Oracle	2.27(0.23)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	2.63(0.32)	0.02(0.00)	0.00(0.00)	0.18(0.12)	0.00(0.00)	0.92	0.00(0.00)	0.00(0.00)	1	54.46(2.21)
		SparseLTS	2.76(0.32)	0.01(0.00)	0.00(0.00)	0.29(0.02)	0.00(0.00)	0.87	0.00(0.00)	0.00(0.00)	1	8.74(1.75)
		MIP	2.75(0.29)	0.01(0.00)	0.00(0.00)	0.03(0.00)	0.00(0.00)	0.99	0.00(0.00)	0.00(0.00)	1	66.16(20.62)
150	200	Oracle	2.24(0.20)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1	0.00(0.00)
		EnetLTS	2.46(0.24)	0.01(0.00)	0.00(0.00)	0.16(0.10)	0.00(0.00)	0.93	0.00(0.00)	0.00(0.00)	1	59.83(2.19)
		SparseLTS	2.67(0.23)	0.01(0.00)	0.00(0.00)	0.35(0.05)	0.00(0.00)	0.85	0.00(0.01)	0.00(0.00)	1	14.67(2.96)
		MIP	2.55(0.21)	0.01(0.00)	0.00(0.00)	0.03(0.00)	0.00(0.00)	0.99	0.00(0.00)	0.00(0.00)	1	375.21(247.13)

Table A.4: Size of the mean shifts for outliers, mean (SD in parenthesis) of RMSPE, variance and squared bias for $\widehat{\boldsymbol{\beta}}$, FPR and FNR for feature selection and outlier detection (as well as the corresponding F_1 scores), and computing time, based on 100 simulation replications for the simulation setting in Section 3.4 based on $n = p = 50$.

$-\lambda_\varepsilon$, λ_X	Method	RMSPE	var($\widehat{\beta}$)	bias($\widehat{\beta}$)^2	FPR($\widehat{\beta}$)	FNR($\widehat{\beta}$)	$F_1(\widehat{\beta})$	FPR($\widehat{\phi}$)	FNR($\widehat{\phi}$)	$F_1(\widehat{\phi})$	Time
1	Oracle	1.83(0.29)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
	EnetLTS	3.68(0.69)	0.07(0.00)	0.14(0.00)	0.24(0.23)	0.09(0.14)	0.85	0.05(0.03)	0.85(0.18)	0.24	11.15(0.53)
	SparseLTS	3.18(0.76)	0.10(0.00)	0.05(0.00)	0.48(0.07)	0.02(0.07)	0.79	0.06(0.03)	0.52(0.31)	0.63	3.25(0.63)
	MIP	2.70(1.14)	0.08(0.00)	0.02(0.00)	0.00(0.01)	0.23(0.28)	0.87	0.04(0.04)	0.34(0.40)	0.77	50.82(45.38)
3	Oracle	1.80(0.25)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
	EnetLTS	4.16(1.67)	0.16(0.00)	0.18(0.00)	0.32(0.23)	0.27(0.27)	0.71	0.05(0.04)	0.57(0.40)	0.58	11.74(0.55)
	SparseLTS	2.72(0.95)	0.10(0.00)	0.01(0.00)	0.54(0.07)	0.04(0.12)	0.77	0.01(0.03)	0.08(0.25)	0.95	3.05(0.36)
	MIP	2.25(1.00)	0.06(0.00)	0.01(0.00)	0.00(0.01)	0.11(0.23)	0.94	0.01(0.03)	0.08(0.26)	0.95	46.66(50.09)
5	Oracle	1.94(0.29)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
	EnetLTS	2.87(1.09)	0.08(0.00)	0.05(0.00)	0.20(0.21)	0.11(0.24)	0.85	0.02(0.03)	0.17(0.34)	0.90	11.96(0.73)
	SparseLTS	2.75(0.76)	0.09(0.00)	0.01(0.00)	0.56(0.07)	0.03(0.13)	0.77	0.01(0.03)	0.06(0.22)	0.97	3.24(0.46)
	MIP	2.22(0.73)	0.04(0.00)	0.00(0.00)	0.00(0.02)	0.07(0.16)	0.96	0.00(0.01)	0.02(0.13)	0.99	27.36(63.40)
7	Oracle	1.87(0.27)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
	EnetLTS	2.45(0.66)	0.04(0.00)	0.03(0.00)	0.15(0.18)	0.02(0.10)	0.92	0.01(0.02)	0.03(0.17)	0.98	12.02(0.61)
	SparseLTS	2.59(0.67)	0.07(0.00)	0.01(0.00)	0.56(0.07)	0.02(0.10)	0.77	0.01(0.02)	0.03(0.16)	0.98	3.81(0.59)
	MIP	2.17(0.79)	0.04(0.00)	0.00(0.00)	0.00(0.01)	0.07(0.16)	0.96	0.00(0.02)	0.03(0.16)	0.98	13.19(16.95)
9	Oracle	1.90(0.34)	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)	0.00(0.00)	1.00	0.00(0.00)
	EnetLTS	2.57(1.15)	0.05(0.00)	0.03(0.00)	0.17(0.20)	0.03(0.15)	0.90	0.01(0.02)	0.04(0.19)	0.98	12.26(0.78)
	SparseLTS	2.49(0.50)	0.06(0.00)	0.01(0.00)	0.55(0.07)	0.00(0.00)	0.79	0.00(0.01)	0.00(0.00)	1.00	3.52(0.48)
	MIP	2.17(0.78)	0.04(0.00)	0.00(0.00)	0.00(0.00)	0.08(0.17)	0.96	0.00(0.01)	0.01(0.08)	1.00	12.40(12.26)

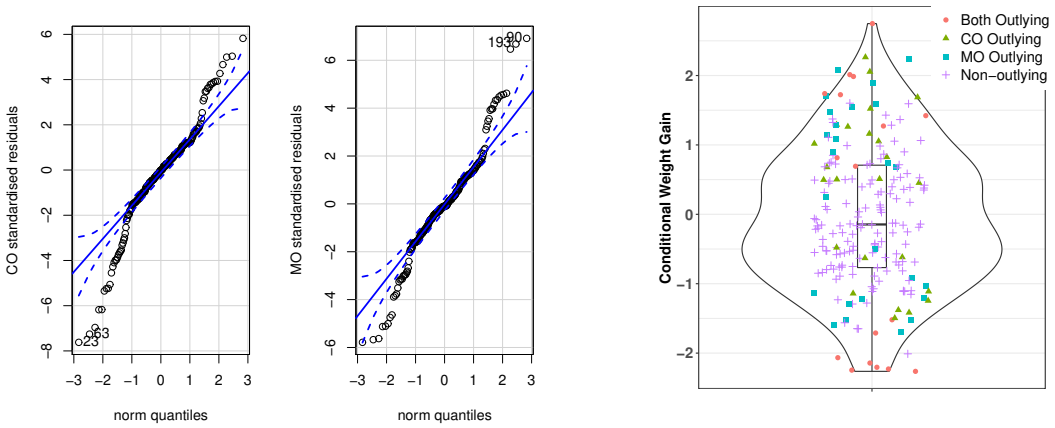


Figure A.1: Q-Q plots for the robust standardized residuals estimated by MIP in child and maternal oral datasets (left panel). Violin plot of the distribution of CWG values with outliers detected under the child and maternal oral regressions color and shape coded (right panel).

level based on the minimum BIC rather than the elbow as in simulations due to the shape of BIC curves. Here computing time for each job is increased to 6000 seconds and the multiplicative constant c for the ensemble method is reduced to 3.

The left panel of Figure A.1 shows the Q-Q plots for robust standardized residuals estimated by MIP for child and maternal oral data. This indicates that both type of residuals deviate from a normality assumption and motivates the use of robust estimation methods. The right panel of Figure A.2 shows that outlying cases for both models often have extreme CWG values, and some overlapping outliers have moderate CWG values.

Figure A.2 shows a plot of robust standardized residuals estimated by our MIP proposal on child and maternal oral data. Here a group of units, with extremely large or small residuals, is classified as outliers in both datasets.

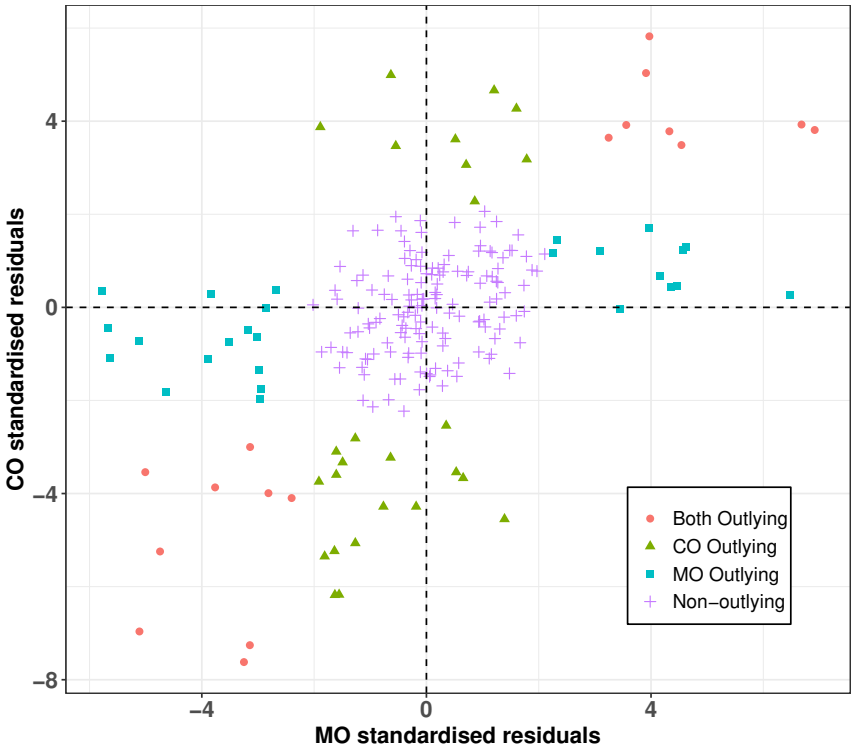


Figure A.2: Robust standardized residuals estimated by MIP under the child dataset against the ones from the maternal oral regression.

Figure A.3 shows the residuals obtained by our MIP proposal in child and maternal oral data against their respective robust measure of outlying-ness in the predictor space. The latter – which robustifies leverage values – is based on a robust location and scatter estimation using the minimum covariance determinant estimator (Rousseeuw and Van Driessen, 1999) in the `rrcov` `R` package (Todorov and Filzmoser, 2009). The trimming proportion is set to 20% as for MIP estimates. Here, especially for maternal oral data, some outliers are characterized by an extreme behavior also in the space of the relevant bacterial types being selected by MIP.

To determine how sensitive prediction results are to different training/test split ratios, we performed the same analysis with $n^{\text{tr}} \approx 0.9n$ and $n^{\text{te}} \approx 0.1n$. Results are summarized in Table A.5. As in the scenario with $n^{\text{tr}} \approx 0.8n$ and $n^{\text{te}} \approx 0.2n$, the two robust approaches produce very dense solutions without any substantial gains in predictive performance (and at times, higher prediction error) compared to the sparser solutions found through our proposal. The non-robust lasso selects an average of 1.5 features for the child oral regression (in addition to the intercept),

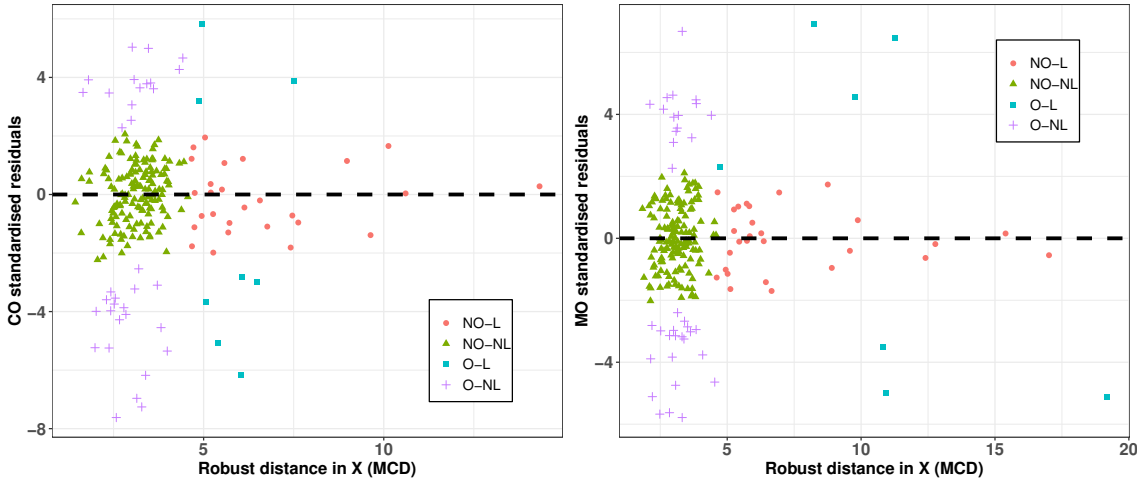


Figure A.3: Robust standardized residuals estimated by MIP under the child dataset (left panel) and maternal oral regression (right panel) against the corresponding robust measure of outlying-ness for the selected features in $\boldsymbol{X}$. In the legend "O" stands for "outliers" and "L" for "Leverage", the letter "N" stands for "non-".

Table A.5: Median (MAD in parenthesis) of TMSPE and the number of features selected on the training set (composed of 90% of the units) on eight train-test splits. Last column: number of features selected on the full data. Robust methods use 20% trimming.

Data	n^{tr}	n^{te}	p	Method	TMSPE	$\widehat{p}_0^{\text{tr}}$	$\widehat{p}_0^{\text{full}}$
Child oral	193	22	68	SparseLTS	0.20(0.01)	53.50(1.05)	52
				EnetLTS	0.12(0.02)	46.50(3.67)	52
				MIP	0.14(0.04)	14.50(0.26)	13
				Lasso	0.15(0.03)	2.50(0.79)	2
Maternal oral	193	22	63	SparseLTS	0.15(0.03)	50.50(0.52)	56
				EnetLTS	0.10(0.01)	50.00(3.41)	62
				MIP	0.10(0.02)	12.50(0.26)	13
				Lasso	0.18(0.03)	1.00(0.00)	1

and the intercept only model for the maternal oral regression.

It should also be noted that the computational burden of our MIP fluctuates based on the quality of warm-starts and big-$\mathcal{M}$ bounds (when the latter are used instead of SOS-1 constraints). Thus, as new, faster, heuristics are proposed, we expect computational efficiency to improve while still providing optimality guarantees. For instance, in our microbiome analysis, alternative robust procedures (i.e., enetLTS and sparseLTS) produced very dense solutions that either had weaker prediction power or showed very marginal improvement over a non-robust approach. As a consequence, the bounds were much wider, leading to higher computing time

(e.g., hitting the maximum time allotted). We noticed this appeared to be due to the relaxation producing weak lower bounds for the MIP; that is, there was very little change in the final solution after 600 seconds compared to, say, 6000 seconds, because the lower bound was slow to improve. This behavior has been well documented in prior literature (e.g., Bertsimas et al. 2016; Hastie et al. 2020; Hazimeh and Mazumder 2020; Kenney et al. 2021).

For further validation, we performed an additional analysis utilizing another phenotype previously explored in Craig et al. (2018). This is actually a richer, longitudinal phenotype expressing the dynamic of the growth index (defined as weight/height) from birth up to three years of age. While a full analysis extending our proposal to functional and/or longitudinal data is beyond the scope of this work, we studied the relationship between the fitted values of CWG produced by various approaches and the growth curves. We found that the CWG fitted values produced through the three robust methods more effectively differentiated higher and lower growth curves. This is illustrated in Figure A.4 where growth curves are color coded based on the corresponding fitted CWG from each method. The non-robust lasso provides very little differentiation for the child oral regression and none at all for the maternal oral regression. This is consistent with the fact that it selected only one bacterial group for the former and produced the intercept alone for the latter. For the robust procedures, the child oral regression was clearly more challenging, with more discrepancies between extreme fitted CWG values and growth curves (for instance, sparseLTS produced a very high fitted CWG for a child whose growth curve is mid-range).

To further test and compare across methods, we utilized the results from our training/test splits. For each iteration, we took the bacterial groups and non-outlying points selected from the three robust procedures and non-robust lasso and fit function-on-scalar regressions (Ramsay and Silverman, 2005) using the growth curves as response. As before, we calculated the prediction error using trimmed median squared prediction error (TMPSE) based on the bacterial groups and growth curves in the test set. Table A.6 summarizes the results of these fits. For the ma-

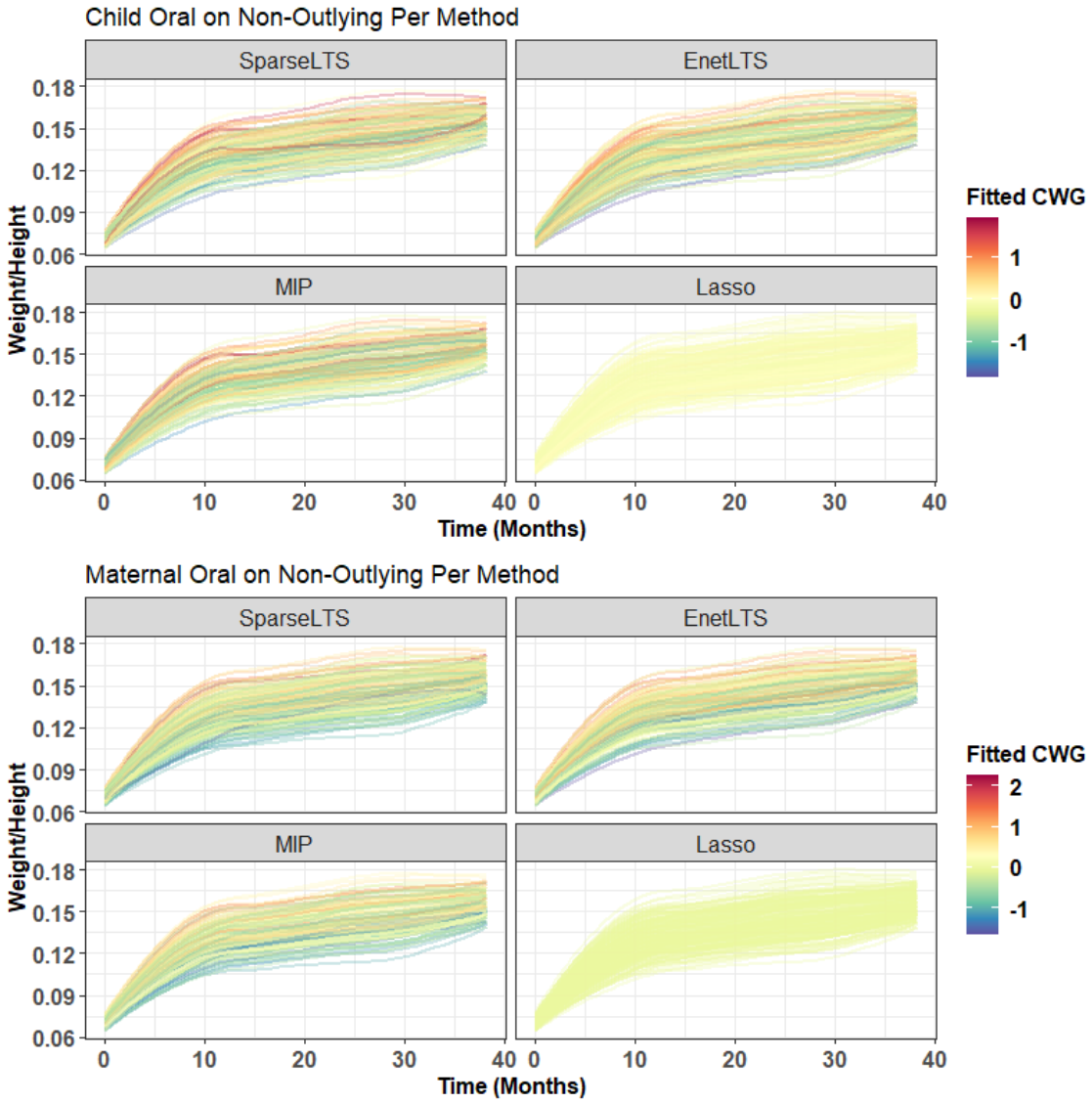


Figure A.4: Growth curves from birth to three years of age for children enrolled in the INSIGHT study. Curves are color-coded based on the fitted CWG values produced by each of the four procedures considered (sparseLTS, enetLTS, MIP and lasso) for the child oral (top) and maternal oral (bottom) regressions.

Table A.6: Median (MAD in parenthesis) of TMSPE for the function-on-scalar regressions and the number of features selected on the training set (from the CWG-based regressions) on eight train-test splits. Last column: number of features selected on the full data. Robust methods use 20% trimming.

Data	n^{tr}	n^{te}	p	Method	TMSPE	$\widehat{p}_0^{tr}$	$\widehat{p}_0^{full}$
Child oral	172	43	68	SparseLTS	0.038(0.006)	54.00(0.26)	52
				EnetLTS	0.041(0.007)	47.00(3.67)	52
				MIP	0.036(0.007)	13.00(0.52)	13
				Lasso	0.035(0.012)	2.00(0.26)	2
Maternal oral	172	43	63	SparseLTS	0.038(0.006)	52.00(1.31)	56
				EnetLTS	0.034(0.005)	52.00(4.46)	62
				MIP	0.031(0.004)	13.50(0.52)	13
				Lasso	0.036(0.012)	1.00(0.00)	1

ternal oral regressions, similar to prediction results obtained considering the scalar CWG response, we observe that our procedure maintains stronger prediction accuracy compared to other procedures. For the child oral regressions our procedure again attains stronger prediction accuracy than the other robust methods. However, we observe a phenomenon similar to the one already discussed in our main analysis, where the non-robust lasso produces a TMSPE similar (in fact, slightly lower) than our proposal. Again, we note that results for the non-robust lasso are markedly less stable; they fluctuate heavily across the different training/test splits producing the highest MAD.

A.4 Algorithmic implementation

The code to replicate both our simulation study and the microbiome application is available at https://github.com/LucaIns/SFSOD_MIP. It also contains features for additional comparisons, such as our proposed method for efficient tuning via cross-validation and the use of SOS-1 constraints. Further user information are provided in the `README` file therein. See also the Supporting Information at https://onlinelibrary.wiley.com/doi/10.1111/biom.13553.

Appendix B

Supplementary Material to Chapter 4

B.1 Theoretical results

Proof of Proposition 4.1. For any trimming level k_n, the objective function in (4.6) subject to integer constraints in (4.6a) can be equivalently formulated as

$$Q(\widehat{\boldsymbol{\beta}}) = \frac{1}{2}\sum_{i=1}^{n-k_n}[(y_i^* - \boldsymbol{\beta}^T\boldsymbol{x}_i^*)^2]_{i:n} + (n-k_n)\sum_{j=1}^{p}R_\lambda(|\beta_j|) \tag{B.1}$$

where $(t_1)_{1:n} \leq \ldots \leq (t_n)_{n:n}$ denote the order statistics of t_i, $\boldsymbol{y}^* = \sqrt{\boldsymbol{\mathcal{M}}_R}\boldsymbol{y}$ and $\boldsymbol{X}^* = \sqrt{\boldsymbol{\mathcal{M}}_R}\boldsymbol{X} = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*)^T$. This relies on the fact that a weighted regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ is equivalent to an unweighted regression of $\boldsymbol{y}^*$ on $\boldsymbol{X}^*$, and we also use Proposition 1 in Insolia et al. (2021d) to transform the mean-shift model based on $\boldsymbol{\phi}$ to a trimmed loss problem without explicit mean shift parameters. Then, denote the contaminated dataset as $\widetilde{\boldsymbol{Z}} = [\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}] = [(\boldsymbol{y} + \boldsymbol{\Delta}_y), (\boldsymbol{X} + \boldsymbol{\Delta}_X)]$. We first show that the BdP $\varepsilon^* \geq (n - k_n + 1)/n$, and then $\varepsilon^* \leq (n - k_n + 1)/n$.

For the first part of the proof assume that $\widetilde{\boldsymbol{Z}}$ contains $m_M \leq k_n$ outliers. Consider $\widehat{\boldsymbol{\beta}} = 0$, so that the associated loss

$$Q(\mathbf{0}) = \sum_{i=1}^{n-k_n}(\widetilde{y}_i^2)_{i:n} \leq \sum_{i=1}^{n-k_n}(y_i^2)_{i:n} \leq (n-k_n)M_y^2,$$

where the first inequality relies on the fact that contaminated data might contain inliers (i.e., mean shifts can be used to reduce the overall residuals sum of square), and $M_y = \max_{i=1,\dots,n} |y_i|$. Now consider any other estimate $\widehat{\boldsymbol{\beta}}$, and assume that $\|\widehat{\boldsymbol{\beta}}\|_2 \geq l$ – i.e., the estimator might break down – where $l = \{(n-k_n)M_y^2 + 1\}/c$ is independent from the contamination mechanism and $c > 0$. It follows that

$$Q(\widehat{\boldsymbol{\beta}}) \geq (n-k_n)\sum_{j=1}^{p} R_\lambda(|\beta_j|) \geq c(n-k_n)\|\boldsymbol{\beta}\|_2 \geq (n-k_n)M_y^2 + 1 > Q(\mathbf{0}),$$

where the first inequality immediately follows from (B.1), and the second inequality is based on the topological equivalence of norms and the definition of SCAD, since $\|\boldsymbol{\beta}\|_1 \geq \sum_{j=1}^{p} R_\lambda(|\beta_j|) \geq c\|\boldsymbol{\beta}\|_2$ for some constant $c > 0$ and any $\boldsymbol{\beta}$ vector. However, $Q(\widehat{\boldsymbol{\beta}}) > Q(\mathbf{0})$ leads to a contradiction as the objective function is non-decreasing in the number of non-zero $\widehat{\beta}_j$ components. Hence, $\|\widehat{\boldsymbol{\beta}}\|_2 < l$ implies that $\varepsilon^* \geq (n-k_n+1)/n$, which concludes the first part of the proof.

For the second part of the proof, consider $m_M > k_n$, and assume that $\|\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{Z}})\|_2 \leq u$ (i.e., the estimator does not breakdown). The objective in (B.1) can be decomposed as

$$\begin{aligned} Q(\widehat{\boldsymbol{\beta}}) &= \sum_{i=1}^{n-m_M} \left[(\widetilde{y}_i^* - \widehat{\boldsymbol{\beta}}^T\widetilde{\boldsymbol{x}}_i^*)^2\right]_{i:n} + \sum_{h=n-m_M+1}^{n-k_n} \left[(\widetilde{y}_h^* - \widehat{\boldsymbol{\beta}}^T\widetilde{\boldsymbol{x}}_h^*)^2\right]_{h:n} + (n-k_n)\sum_{j=1}^{p} R_\lambda(|\widehat{\beta}_j|) \\ &\geq \left[\{(y_i^* - \boldsymbol{\beta}^T\boldsymbol{x}_i^*) + (\Delta_{y_i} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{\Delta}_{x_i})\}^2\right]_{i=n-m_M+1} + (n-k_n)\sum_{j=1}^{p} R_\lambda(|\widehat{\beta}_j|) \end{aligned} \tag{B.2}$$

since at least one of the m_M outliers might be included in the fit – i.e., the $(n-n_0+1)$-th ordered squared residual if contamination is adversarial. Hence, since mean shifts Δ_{y_i} and $\boldsymbol{\Delta}_{x_i}$ can take arbitrary values, it is easy to see that (B.2) is unbounded similarly to OLS. This contradicts $\|\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{Z}})\|_2 \leq u$ and proves the result. ■

Proof of Proposition 4.2. It extends Theorem 1 in Fan and Li (2012) to the presence of MSOM contamination. Specifically, we can use the same argument, but their conditions must hold at least on $n - m_M$ (uncontaminated) points as opposed to n. Since k_n largest residuals (say, $k_n = m_M$) are always discarded from our loss in

(4.6), we thus need to ensure that these trimmed points encompass MSOM outliers. Condition 4.2(D) guarantees this, similarly to Theorem 3 in Insolia et al. (2021d), so that MSOM outliers have largest residuals for any model of size $k_p \leq p_0$. See Fan and Li (2012) for details. ■

Proof of Proposition 4.3. This result immediately follows from Theorem 2 in Fan and Li (2012) specifically focusing on VIOM outliers as random effects (i.e., our term $\boldsymbol{I}_n\boldsymbol{\gamma}$ instead of $\boldsymbol{Zb}$). However, in Fan and Li (2012) the dimension of the random effects $\boldsymbol{b}$ can increase exponentially with the sample size n, but in our formulation $\boldsymbol{\gamma}$ can only increase linearly with $n - k_n$. Thus, our conditions in list 4.2 might be relaxed to account only for VIOMs. Nevertheless, these more general conditions allow one to extend our results also to the presence of additional (pure) random effects, whose size can increase exponentially with $n - k_n$. ■

Proof of Theorem 4.1(1). The proofs for Theorem 4.1 follow some lines of the argument in Theorems 1 and 3 of Liu and Yu (2013), where an OLS or ridge fit is computed on top of the features selected by lasso.

Here with a slight abuse of notation, we denote $P(\mathcal{S}) = P(\widehat{\mathcal{S}} = \mathcal{S})$ and $P(\widetilde{\mathcal{S}}) = P(\widehat{\mathcal{S}} \neq \mathcal{S})$, where $\widehat{\mathcal{S}} = \{\widehat{\mathcal{S}}_\beta, \widehat{\mathcal{S}}_\phi, \widehat{\mathcal{S}}_\gamma\}$. Furthermore, we indicate as $\widehat{\boldsymbol{\beta}}|\widehat{\mathcal{S}}$ the estimated coefficients conditionally on the selected model, which is abbreviated as $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{S}}}$. It is also assumed that, conditionally on any selected model $\widehat{\mathcal{S}}$, units weights $\widehat{\boldsymbol{W}}$ are deterministic; i.e., they are treated as known constants (which is customary in robust statistics). Hence, conditionally on the true model $\mathcal{S}$, we have that $\widehat{\boldsymbol{W}} = \boldsymbol{W} = \boldsymbol{V}^{-1}$.

By the law of total expectations and using $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \leq 2(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2)$, it follows

that

$$\begin{aligned}
\|E\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 &= \|E\widehat{\boldsymbol{\beta}}_{\mathcal{S}} P(\mathcal{S}) + E\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}} P(\widetilde{\mathcal{S}}) - \boldsymbol{\beta}_0\|_2^2 \\
&\leq 2\|E\widehat{\boldsymbol{\beta}}_{\mathcal{S}} P(\mathcal{S}) - \boldsymbol{\beta}_0\|_2^2 + 2\|E\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}} P(\widetilde{\mathcal{S}})\|_2^2 \\
&= 2\|E\{(\boldsymbol{X}_{\mathcal{S}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\mathcal{S}})^+ \boldsymbol{X}_{\mathcal{S}}^T \widehat{\boldsymbol{W}} \boldsymbol{y}\} P(\mathcal{S}) - \boldsymbol{\beta}_0\|_2^2 + 2P(\widetilde{\mathcal{S}})\|E\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}}\|_2^2 \\
&= 2\|P(\mathcal{S})\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0\|_2^2 + 2P(\widetilde{\mathcal{S}})\|E\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}}\|_2^2 \\
&= 2\|\boldsymbol{\beta}_0(P(\mathcal{S}) - 1)\|_2^2 + 2P(\widetilde{\mathcal{S}})\|E\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}}\|_2^2 \\
&= 2P(\widetilde{\mathcal{S}})\{\|\boldsymbol{\beta}_0\|_2^2 + \|E\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}}\|_2^2\}. \qquad \text{(B.3)}
\end{aligned}$$

Further, using Jensen's inequality and the fact that $\|\boldsymbol{A}\boldsymbol{b}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{b}\|$ provides

$$\begin{aligned}
\|E\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}}\|_2^2 &\leq E\|(\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\widetilde{\mathcal{S}}})^+ \boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{y}\|_2^2 \\
&\leq E\|(\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\widetilde{\mathcal{S}}})^+ \boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}}^{1/2}\|_2^2 \|\widehat{\boldsymbol{W}}^{1/2} \boldsymbol{y}\|_2^2 \\
&= \Lambda_{\max}\{(\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\widetilde{\mathcal{S}}})^+\} E\|\widehat{\boldsymbol{W}}^{1/2} \boldsymbol{X}\boldsymbol{\beta}_0 + \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}\|_2^2 \\
&= \Lambda_{\max}\{(\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\widetilde{\mathcal{S}}})^+\} E(\|\widehat{\boldsymbol{W}}^{1/2} \boldsymbol{X}\boldsymbol{\beta}_0\|_2^2 + \varepsilon^T \widehat{\boldsymbol{W}} \varepsilon) \\
&= \Lambda_{\max}\{(\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\widetilde{\mathcal{S}}})^+\} (\|\widehat{\boldsymbol{W}}^{1/2} \boldsymbol{X}\boldsymbol{\beta}_0\|_2^2 + \mathrm{tr}(\widehat{\boldsymbol{W}})\sigma^2) \qquad \text{(B.4)} \\
&\leq \Lambda_{\max}\{(\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\widetilde{\mathcal{S}}})^+\} (\|\boldsymbol{X}\boldsymbol{\beta}_0\|_2^2 + n\sigma^2),
\end{aligned}$$

where $\Lambda_{\max}(\cdot)$ denotes the largest eigenvalue, and for a real matrix $\boldsymbol{A}$, the spectral norm $\|\boldsymbol{A}\|_2 = \sqrt{\Lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^T)} = \sqrt{\Lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})}$. In our case,

$$\begin{aligned}
\|(\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\widetilde{\mathcal{S}}})^+ \boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}}^{1/2}\|_2^2 &= \Lambda_{\max}\{(\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\widetilde{\mathcal{S}}})^+ \boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\widetilde{\mathcal{S}}} (\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\widetilde{\mathcal{S}}})^+\} \\
&= \Lambda_{\max}\{(\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\widetilde{\mathcal{S}}})^+\},
\end{aligned}$$

where the last equality follows from the property of a generalized inverse $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$. Combining (B.3) and (B.4) leads to the desired results. ■

Proof of Theorem 4.1(2). Introducing the WLS oracle estimator $\widehat{\boldsymbol{\beta}}_0$ and using the fact that

$$E\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\|_2 = E\|(\boldsymbol{X}_{\mathcal{S}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\mathcal{S}})^+ \boldsymbol{X}_{\mathcal{S}}^T \widehat{\boldsymbol{W}} \boldsymbol{\varepsilon}\|_2 = 0$$

provides

$$
\begin{aligned}
E\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2 &= E\|\widehat{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}_0-\widehat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0\|_2^2 \\
&= E\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_0\|_2^2+E\|\widehat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0\|_2^2 \\
&= E\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_0\|_2^2+\sigma^2\mathrm{tr}(\boldsymbol{\Sigma}_X^{-1})/\mathrm{tr}(\widehat{\boldsymbol{W}}) \qquad \text{(B.5)}
\end{aligned}
$$

the last equality follows from the MSE for the WLS oracle estimator and such term cannot be improved. Thus, we control the first term as follows

$$
\begin{aligned}
E\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_0\|_2^2 &= E\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}}-\widehat{\boldsymbol{\beta}}_0\|_2^2P(\mathcal{S})+E\|\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}}-\widehat{\boldsymbol{\beta}}_0\|_2^2P(\widetilde{\mathcal{S}}) \\
&= E\|\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}}-\widehat{\boldsymbol{\beta}}_0\|_2^2P(\widetilde{\mathcal{S}}), \qquad \text{(B.6)}
\end{aligned}
$$

where the first equality relies on the law of total expectations and the last one uses the fact that $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{S}}}=\widehat{\boldsymbol{\beta}}_0$ conditionally on $\{\widehat{\mathcal{S}}=\mathcal{S}\}$.

Further, note that

$$
\begin{aligned}
E\|\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}}-\widehat{\boldsymbol{\beta}}_0\|_2^2 &\leq 2\{E\|\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{S}}}\|_2^2+E\|\widehat{\boldsymbol{\beta}}_0\|_2^2\} \\
&= 2\{E\|(\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T\widehat{\boldsymbol{W}}\boldsymbol{X}_{\widetilde{\mathcal{S}}})^+\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T\widehat{\boldsymbol{W}}\boldsymbol{y}\|_2^2+E\|(\boldsymbol{X}_{\mathcal{S}}^T\widehat{\boldsymbol{W}}\boldsymbol{X}_{\mathcal{S}})^+\boldsymbol{X}_{\mathcal{S}}^T\widehat{\boldsymbol{W}}\boldsymbol{y}\|_2^2\} \\
&\leq 2E\|\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{y}\|_2^2\left[\Lambda_{\max}\{(\boldsymbol{X}_{\widetilde{\mathcal{S}}}^T\widehat{\boldsymbol{W}}\boldsymbol{X}_{\widetilde{\mathcal{S}}})^+\}+\Lambda_{\max}\{(\boldsymbol{X}_{\mathcal{S}}^T\widehat{\boldsymbol{W}}\boldsymbol{X}_{\mathcal{S}})^+\}\right], \\
& \qquad \text{(B.7)}
\end{aligned}
$$

where the first upper bound follows from $\|a+b\|^2\leq 2(\|a\|^2+\|b\|^2)$, and the second one uses $\|Ab\|\leq\|A\|\|b\|$. Finally, combining

$$
E\|\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{y}\|_2^2\leq E(\|\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}\boldsymbol{\beta}_0\|_2^2+\varepsilon^T\widehat{\boldsymbol{W}}\varepsilon)=\|\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}\boldsymbol{\beta}_0\|_2^2+\mathrm{tr}(\widehat{\boldsymbol{W}})\sigma^2\leq\|\boldsymbol{X}\boldsymbol{\beta}_0\|_2^2+n\sigma^2
$$

with (B.5), (B.6), and (B.7) concludes the proof. ■

Proof of Theorem 4.1(3). Under the conditions in lists 4.1-4.3, as $(n-m_M)\to\infty$, it follows that $P(\widehat{\mathcal{S}}=\mathcal{S})\to 1$ for some suitable constants. Thus, $\widehat{\boldsymbol{\beta}}$ has an asymptotic

normal distribution as it is a linear combination of normal distributions

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}_{\mathcal{S}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\mathcal{S}})^{-1} \boldsymbol{X}_{\mathcal{S}}^T \widehat{\boldsymbol{W}} \boldsymbol{y} \\ &= (\boldsymbol{X}_{\mathcal{S}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\mathcal{S}})^{-1} \boldsymbol{X}_{\mathcal{S}}^T \widehat{\boldsymbol{W}} (\boldsymbol{X}_{\mathcal{S}} \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}_0 + (\boldsymbol{X}_{\mathcal{S}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\mathcal{S}})^{-1} \boldsymbol{X}_{\mathcal{S}}^T \widehat{\boldsymbol{W}} \boldsymbol{\varepsilon} \sim N(\boldsymbol{\beta}_0, \sigma^2 (\boldsymbol{X}_{\mathcal{S}}^T \widehat{\boldsymbol{W}} \boldsymbol{X}_{\mathcal{S}})^{-1}),\end{aligned}$$

and $\widehat{\boldsymbol{W}} = \boldsymbol{V}^{-1}$ guarantees that it asymptotically reaches maximum efficiency. ■

B.2 Technical details

B.2.1 Choice of the proxy matrix $\mathcal{M}$

For mixed-effects linear models without data contamination as in Section 4.2.2, Fan and Li (2012) propose to replace $\sigma^{-2}\boldsymbol{\mathcal{B}}$ in (4.5) with a proxy matrix $\boldsymbol{\mathcal{M}}_b$. They show that under mild conditions it is safe to choose $\boldsymbol{\mathcal{M}}_b = \log(n)\boldsymbol{I}_n$, as the eigenvalues of $\boldsymbol{Z}^T\boldsymbol{P}_x\boldsymbol{Z}$ and $\boldsymbol{Z}\boldsymbol{Z}^T$ have magnitude increasing with n, so that they are likely to dominate the eigenvalues of $\boldsymbol{\mathcal{M}}_b$ for a large enough n. While this choice excludes cross-correlations in the random effects, it avoids the estimation of a large number of parameters as in the case of an unstructured covariance matrix.

In our formulation the terms $\boldsymbol{\mathcal{M}}_R$ and $\boldsymbol{\mathcal{M}}_\gamma$ in (4.6) and (4.8) are proxies for the unknown $\boldsymbol{P}_R$ and $\boldsymbol{\Gamma}$, respectively. Following Fan and Li (2012), in our implementation we use $\boldsymbol{\mathcal{M}}_\gamma = \log(n)\boldsymbol{I}_n$ on the first iteration of SCADws. If the 3-step procedure is re-iterated, such as in SCAD2s, we use estimated weights $\widehat{\boldsymbol{W}}$ from the previous iteration for their update; see Section B.3.1 for details.

B.2.2 Weights estimation

The formulation in (4.8) highlights that if $\widehat{\gamma}_i = 0$ also the corresponding variance inflation $\widehat{\omega}_i = 0$. However, it might be of interest to estimate ω_i when the corresponding $\widehat{\gamma}_i \neq 0$. A similar reasoning holds for step 3 of the heuristic method

described in Section 4.3.4. Note that

$$w_i = v_i^{-1} = (1 + \omega_i)^{-1} = (1 + \text{var}(\gamma_i)/\sigma^2)^{-1},$$

which can be estimated as follows:

1. Apply REMLE assuming that the units corresponding to non-zero components in $\widehat{\boldsymbol{\gamma}}$ arise from a VIOM. In principle, all weights should be jointly estimated, although this can be computationally heavy for large problems. A similar approach was used by Fan and Li (2012, p.2060) in one of their examples. Similarly to Insolia et al. (2021b), we also consider single-weights estimates as in FSRws, where each VIOM outlier is separately included in the model and estimated. This is the approach used in our simulations and applications.

2. The quantity γ_j^2/n can be used as an estimate of $\text{var}(\gamma_j)$ (Fan and Li, 2012, p. 2053 Eq. 20). Thus, one can consider $w_i = (1 + \widehat{\gamma}_i^2 c_1/\widehat{\sigma}^2)^{-1}$ where c_1 is a normalizing constant and the value $c_1 = 1/n$ was suggested by the authors.

3. One can treat the selected random effects γ_i as additional fixed effects and apply a ridge penalty (Hoerl and Kennard, 1970a). This can be considered optimal and is motivated by the fact that assuming a normal prior $N(\mathbf{0}, \sigma^2\boldsymbol{\Gamma})$ on $\boldsymbol{\gamma}$ leads to the ridge estimator as the maximum posterior probability estimator. Indeed, the estimates $\widehat{\boldsymbol{\gamma}}$ represent prediction residuals, so that their shrinkage performs a down-weighting scheme. Moreover, Grandvalet (1998) showed that adaptive ridge is equivalent to lasso estimation; this can be useful to simultaneously select and estimate optimal units' weights (e.g., combining Steps 2 and 3 of our main proposal and/or heuristic procedure).

B.2.3 Parameter tuning

In our proposal we tune the sparsity level for fixed-effects in Step 1, and the number of VIOM outliers as random-effects in Step 2, and we focus on the use of BIC-type criteria. Other approaches can be envisaged to achieve these goals, but this is left for future work.

In the first step of our proposal described in Section 4.3.1, we rely on a BIC-type criterion to tune the sparsity level of SCAD; note that MSOM outliers are removed according to the sparseLTS fit (see the main text for details). Namely, we select the λ which minimizes

$$\text{BIC}^F_{\lambda_k} = \bar{n} \log\{[(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_k})^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_k})/(\bar{n}\widehat{\sigma}^2_P)]^{1/2}\} + \log(\bar{n})k_p^{\lambda_k} \tag{B.8}$$

where $\bar{n} = n - k_n$, λ_k is the k-th element of the λ's sequence, $\widehat{\boldsymbol{\beta}}_{\lambda_k}$ is the minimizer of (4.6) for $\lambda = \lambda_k$, $k_p^{\lambda_k} = |\widehat{\mathcal{S}}_\beta(\lambda_k)|$ is the number of nonzero components in $\widehat{\boldsymbol{\beta}}_{\lambda_k}$, and the term $\widehat{\sigma}^2_P$ is a preliminary estimate of scale (e.g., the one obtained by our heuristic or SCADws).

Also in the second step of our proposal described in Section 4.3.2 we rely on a BIC-type criterion to adaptively detect VIOM outliers. Here we select the λ which minimizes

$$\text{BIC}^R_{\lambda_k} = 2\bar{n} \log\left[\frac{\widehat{\sigma}^2_S}{\sigma^2_h}(\boldsymbol{y} - \widehat{\boldsymbol{\gamma}}_{\lambda_k})^T \boldsymbol{P}_{\overline{x}}(\boldsymbol{y} - \widehat{\boldsymbol{\gamma}}_{\lambda_k})\right] + \log(\bar{n})k_W^{\lambda_k} \tag{B.9}$$

where $\widehat{\sigma}^2_S$ is an estimate of the error variance (e.g., based on an S-estimator with 50% breakdown point), $\widehat{\boldsymbol{\gamma}}_{\lambda_k}$ is the SCAD estimate of $\boldsymbol{\gamma}$ based on (4.8) for the k-th element of the λ's sequence, $k_W^{\lambda_k} = \bar{n} - \text{tr}(\widehat{\boldsymbol{W}}_{\lambda_k})$ with units' weights $\widehat{\boldsymbol{W}}_{\lambda_k}$ estimated (independently) according to $\widehat{\boldsymbol{\gamma}}_{\lambda_k}$ through REMLE, and σ^2_h is the variance of the truncated normal distribution containing a central portion $h/\bar{n}$ of the full distribution (for $h = \bar{n} - |\widehat{\mathcal{S}}_\gamma(\lambda_k)|$). Namely, if $h < \bar{n}$:

$$\sigma^2(h) = 1 - \frac{2\bar{n}}{h}\Phi^{-1}\left(\frac{\bar{n}+h}{2\bar{n}}\right)\phi\left\{\Phi^{-1}\left(\frac{\bar{n}+h}{2\bar{n}}\right)\right\},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density and cumulative density function for the standard normal distribution, respectively. This consistency correction factor is often used in robust regression (Rousseeuw and Leroy, 1987; Riani et al., 2014, 2022), and relies on the one-dimensional case of elliptical truncation of the multivariate normal distribution provided in Tallis (1963); see also Johnson et al. (1995, pp.156–162) for details. For $h = \bar{n}$, the correction term is $\sigma^2(h) = 1$ and thus

neglected.

In principle, as a generalization of (B.8), one can combine the BIC-like criteria in Insolia et al. (2021d) and Riani et al. (2022) to simultaneously tune the trimming proportion and the sparsity level in (4.6). Moreover, to take into account the co-occurrence of VIOM outliers, one can perform VIOM detection in (4.8) through an extension of our BIC-type criterion in (B.9) in the spirit of CAIC and extended CAIC discussed in Section 4.2.2. However, this is left for future work.

In Step 1 of our heuristic procedure described in Section 4.3.4, we rely on 10-fold cross-validation to select fixed effects through SCAD. Note that also here MSOMs are removed according to the sparseLTS fit when present. In Step 2 we rely on a BIC-type criterion similarly to (B.9) to detect VIOMs through SCAD. We select the λ which minimizes

$$\text{BIC}_{\lambda_k}^{\text{heur},R} = \frac{2}{\widehat{\sigma}_S^2}\left[\frac{(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\gamma}}_{\lambda_k})^T(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\gamma}}_{\lambda_k})}{\sigma_h^2}\right] + \log(\bar{n})k_V^{\lambda_k}, \tag{B.10}$$

where $\widehat{\boldsymbol{\beta}}$ is the regression vector estimated in the first step of our heuristic method, $\widehat{\boldsymbol{\gamma}}_{\lambda_k}$ are the penalized estimates from its second step, $k_V^{\lambda_k} = |\widehat{\mathcal{S}}_\gamma(\lambda_k)|$ is the number of nonzero components in $\widehat{\boldsymbol{\gamma}}_{\lambda_k}$, and the other terms are defined in (B.9).

B.2.4 Parallel between our heuristic approach and M-estimation

The proposed heuristic method has a parallel with the following multi-stage, penalized M-estimation procedure.

Step 1 is equivalent to an adaptive hard-trimming, sparse estimator (i.e., it selects features and assigns binary weights) and guarantees an high-breakdown point. This step aims to exclude MSOMs and select only the relevant features (see for instance Alfons et al. 2013; Kurnaz et al. 2017; Insolia et al. 2021d). Step 2 corresponds to an adaptive "truncated" M-estimator, where only the most extreme cases are down-weighted. In full generality, this estimator takes the form $\widehat{\boldsymbol{\beta}} = \arg\min_\beta \sum_{i=1}^n \rho(u_i)$, where $u_i = (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})/\sigma$. Here the idea is that the $n - m_M - m_V$ uncontaminated

points receive full weights as in OLS, but only VIOMs are down-weighted according to the $\rho(\cdot)$ function in use, and MSOMs (if present) are excluded from the fit. For

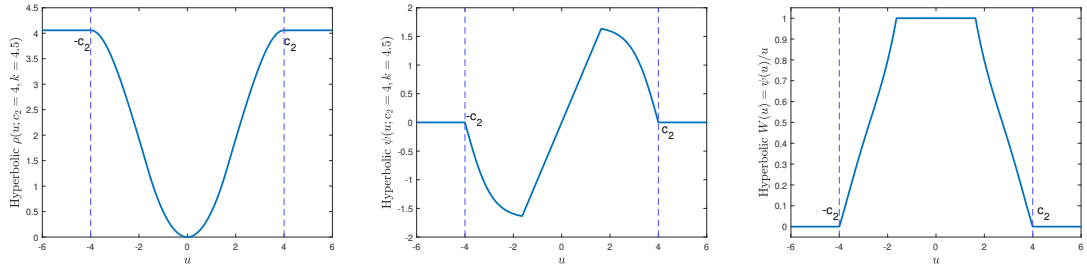


Figure B.1: Hyperbolic tangent ρ function (left panel), ψ function (central panel), and weight function (right panel) for $c_2 = 4$ and $k = 4.5$.

instance, this has a parallel with the *hyperbolic tangent* $\rho(\cdot)$ function, which can be considered as refinement of Hampel's piecewise linear redescending function and is related to the *change of variance curve* (Hampel et al., 1981). Tanh-estimators are more easily defined in terms of their derivatives, and the corresponding $\psi(\cdot)$ function is

$$\psi(u) = \begin{cases} u & \text{if } |u| \leq c_1 \\ \{A(k-1)\}^{1/2} \tanh\left[\frac{1}{2}\left\{(k-1)B^2/A\right\}^{1/2}(c_2 - |u|)\right] \text{sign}(u) & \text{if } c_1 \leq |u| \leq c_2 \\ 0 & \text{if } |u| > c_2 \end{cases}$$

for suitable constants k, A, B, c_1, and c_2, where $0 < c_1 < c_2$ satisfies

$$c_1 = \{A(k-1)\}^{1/2} \tanh\left[\frac{1}{2}\left\{(k-1)B^2/A\right\}^{1/2}(c_2 - c_1)\right].$$

These constants are traditionally computed iteratively, based on the Newton-Raphson algorithm and numerical integration. Figure B.1 shows the corresponding ρ, ψ, and weight functions for $c_2 = 4$ and $k = 4.5$.

Unlike tanh-estimators, our heuristic proposal does not pre-specify a trade-off between breakdown point and efficiency, but this is adaptively tuned as follows. The rejection point c_2 approximately corresponds to the smallest standardized residual for the MSOMs detected at Step 1. Similarly, the constant c_1 is set to the value of the largest standardized residual for points which are not affected by MSOM or VIOM. Specifically, for our heuristic proposal, c_1 and c_2 can be computed based on order

statistics from the scaled residuals obtained at Step 1. Ideally, assuming without loss of generality that all outliers have sizeable residuals, these corresponds to the $(n - m_V - m_M)$-th and $(n - m_M)$-th order statistics of the absolute standardized residuals, respectively.

B.3 Simulation study details

Experiments were carried out using `MATLAB 2021b` and `R` version `3.6.2`. The hardware in use has an Intel Core i7-7700HQ CPU @ 2.8 GHz $\times$ 4 processors and 16 GB RAM.

B.3.1 Algorithmic implementations

Scenario 1: low-dimensional setting

LTS, MM85, MM95 and FSRws are computed through the `FSDA MATLAB Toolbox` (FSDA, 2022), with default input parameters – note that MM-estimators rely on an S-estimator with breakdown point set to 0.5. Our heuristic procedure based on SCAD is computed through the `Penalized MATLAB toolbox` (McIlhagga, 2016) and VIOM detection is tuned according to (B.10) for 100 λ values. For SCADws, VIOM detection (Step 2 of our main proposal) is tuned through the BIC-type information criterion in (B.9) for 50 λ values. Units' weights for FSRws, Heur and SCADws are estimated (independently) through the `VIOM()` function which is available on `FSDA` (Insolia et al., 2021b).

Scenario 2: high-dimensional setting

Lasso (Tibshirani, 1996) is computed through the `Penalized MATLAB toolbox` and its sparsity level is tuned according to a 5-fold cross validation for 100 λ values.

SparseLTS, which combines an L_1-penalty with the LTS estimator (Alfons et al., 2013), is computed through the `sparseLTS()` function of the `robustHD` package (Alfons, 2021) in `R`. The latter is called from `MATLAB` itself through the `RunRcode()` function[1]. The algorithm starts with 500 random subsamples, where 10 subsam-

[1] https://it.mathworks.com/matlabcentral/fileexchange/50071-runrcode-rscriptfilename-rpath

ples with the lowest value of the objective function are used to compute additional concentration-steps (which are at most 100). The sparsity level is chosen according to the minimum root trimmed mean squared prediction error based on a repeated 10-fold cross validation for 100 λ values, where the number of replications is equal to 10. Note that we do not perform a final re-weighting step.

TaL, which is an adaptive lasso with Tukey's bisquare loss, is computed through the `MATLAB` code provided in Chang et al. (2018). Its tuning constant is set at $c = 4.685$ to achieve 85% nominal efficiency, and we use the sparseLTS solution to initialize the algorithm. Its sparsity level is tuned according to the BIC-type criterion proposed by the authors, and we consider 50 λ values.

SCADopt, SCADws and SCAD2s use sparseLTS as a preliminary estimator, which can be considered as the first step in solving SCAD through the LLA algorithm. Specifically, we apply our proposals on the set of cases detected as non-MSOMs by sparseLTS, and its set of active features. In terms of proxy matrices, SCADws uses $\boldsymbol{\mathcal{M}}_\gamma = \log(n - m_M)\boldsymbol{I}_{n-m_M}$ (see Section B.1); SCAD2s, for the set of cases $\widehat{\mathcal{S}}_\gamma$ estimated as VIOMs by SCADws, relies on $\boldsymbol{\mathcal{M}}_\gamma^{\widehat{\mathcal{S}}_\gamma} = \widehat{\boldsymbol{W}}_{\widehat{\mathcal{S}}_\gamma}^{-1} - \boldsymbol{I}_{|\widehat{\mathcal{S}}_\gamma|}$; similarly, SCADopt relies on population weights for the true VIOM outliers using $\boldsymbol{\mathcal{M}}_\gamma^{\mathcal{S}_\gamma} = \boldsymbol{W}_{\mathcal{S}_\gamma}^{-1} - \boldsymbol{I}_{m_V}$. In SCAD2s and SCADopt, the remaining diagonal entries of $\boldsymbol{\mathcal{M}}_\gamma$ are set to 0 in the first step. However, in the second step, they are set to 1 – this allows us to detect as VIOMs also cases that were not detected by SCADws and non-outlying cases with sizeable residuals, respectively. For SCADws, SCAD2s and SCADopt, we tune their sparsity level (Step 1 of our proposal) and detect VIOMs (Step 2 of our proposal) through the information criteria described in (B.8) and (B.9). Here SCADws uses the estimate of s^2 provided by our heuristic for $\widehat{\sigma}_P^2$ in (B.8), and the estimated error variance provided by an S-estimator with 50% BdP for $\widehat{\sigma}_S^2$ in (B.9). SCADopt and SCAD2s rely on the population variance and the estimate of s^2 provided by SCADws, respectively, and we consider $\widehat{\sigma}_P^2 = \widehat{\sigma}_S^2$. At each step, similarly to TaL, our proposals are evaluated over a sequence of 50 λ values.

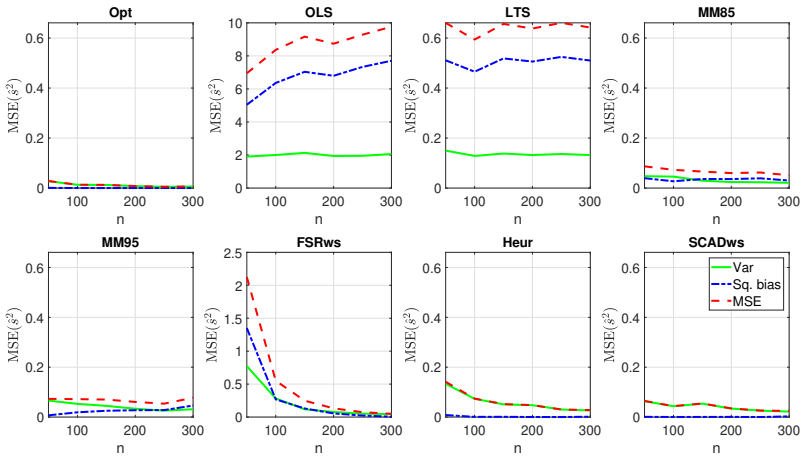


Figure B.2: Scenario 1. Trimmed MSE($\widehat{s}^2$) comparisons (with 20% upper trimming) across procedures and sample sizes.

B.3.2 Additional settings

Scenario 1: low-dimensional VIOMs

Figure B.2 extends Figure 4.2 and shows the trimmed MSE for $\widehat{s}^2$. Here, for each method, we trimmed 20% of the largest squared entries in $\sigma^2 - \widehat{s}^2$ and then computed the MSE decomposition on the remaining instances. SCADws greatly outperform other methods both in terms of biases and variances sample sizes, and Heur performs comparably but has higher variability. A comparison with Figure 4.2 indicates that only a few instances (at most 20 out of 100) were affecting the MSE of SCADws and Heur, but other methods report very similar results. In particular, MM85 and MM95 have persistent biases across sample sizes, and their variability is comparable to SCADws. This highlights the drawback of using M-estimators with a pre-specified nominal efficiency.

We also explored various SNR regimes. The following simulation settings are the same as scenario 1 in Section 4.4.1, with the only difference being that SNR $= 1, 2, 5$. Table B.1-B.3 show simulation results in term of MSE decompositions for $\widehat{\boldsymbol{\beta}}$ and $\widehat{s}^2$, and mean (with standard deviations in parenthesis) FPR and FNR for VIOM detection. A comparison with the results in Section 4.4.1 shows that similar conclusions hold, although all methods report an overall decrease in performance for weaker SNR regimes. Also here, SCADws converges faster to the oracle solution and outperforms other feasible estimation methods.

Table B.1: Scenario 1 with SNR = 1. MSE for $\widehat{\boldsymbol{\beta}}$ and $\widehat{s}^2$ (decomposed into squared bias and variance), and mean (SD in parenthesis) FPR and FNR for feature selection and outlier detection, based on 100 simulation replications.

n	p	Method	bias$(\widehat{\beta})^2$	var$(\widehat{\beta})$	bias$(\widehat{s}^2)^2$	var$(\widehat{s}^2)$	FPR$(\widehat{\tau})$	FNR$(\widehat{\tau})$
50	2	Opt	0.00092	0.0967	0.0057	0.5358	0.00(0.00)	0.00(0.00)
		OLS	0.00692	0.2514	71.8625	18.7466	0.00(0.00)	1.00(0.00)
		LTS	0.00259	0.2458	5.6526	0.3960	0.14(0.04)	0.37(0.13)
		MM85	0.00106	0.1485	0.7348	0.7163	1.00(0.00)	0.00(0.00)
		MM95	0.00140	0.1427	0.3959	1.5487	1.00(0.00)	0.00(0.00)
		FSRws	0.00400	0.2072	25.4549	14.4381	0.00(0.00)	0.91(0.14)
		Heur	0.00276	0.1556	0.2688	2.6999	0.07(0.11)	0.50(0.21)
		SCADws	0.00144	0.1545	0.0135	1.8640	0.03(0.04)	0.53(0.21)
100	2	Opt	0.00034	0.0582	0.0006	0.3243	0.00(0.00)	0.00(0.00)
		OLS	0.00180	0.1210	77.2597	9.5786	0.00(0.00)	1.00(0.00)
		LTS	0.00158	0.1467	4.8607	0.1985	0.13(0.03)	0.38(0.08)
		MM85	0.00086	0.0888	0.4516	0.6457	1.00(0.00)	0.00(0.00)
		MM95	0.00092	0.0863	0.6477	1.2266	1.00(0.00)	0.00(0.00)
		FSRws	0.00103	0.0954	7.5363	8.1701	0.00(0.00)	0.79(0.15)
		Heur	0.00128	0.0909	0.0000	2.2979	0.04(0.05)	0.56(0.16)
		SCADws	0.00107	0.0918	0.0308	1.6272	0.02(0.03)	0.57(0.14)
150	2	Opt	0.00069	0.0305	0.0011	0.2358	0.00(0.00)	0.00(0.00)
		OLS	0.00010	0.0742	81.7951	8.5131	0.00(0.00)	1.00(0.00)
		LTS	0.00010	0.0761	5.1616	0.1196	0.13(0.02)	0.38(0.07)
		MM85	0.00037	0.0454	0.5344	0.3351	1.00(0.00)	0.00(0.00)
		MM95	0.00054	0.0464	0.6117	0.6909	1.00(0.00)	0.00(0.00)
		FSRws	0.00038	0.0501	2.7022	2.1681	0.00(0.01)	0.71(0.13)
		Heur	0.00042	0.0475	0.0079	1.0925	0.03(0.03)	0.53(0.12)
		SCADws	0.00036	0.0475	0.0026	0.9926	0.02(0.03)	0.55(0.12)
200	2	Opt	0.00033	0.0261	0.0040	0.1745	0.00(0.00)	0.00(0.00)
		OLS	0.00053	0.0613	76.2648	5.7191	0.00(0.00)	1.00(0.00)
		LTS	0.00011	0.0510	5.0271	0.0907	0.13(0.02)	0.38(0.06)
		MM85	0.00014	0.0339	0.5330	0.2268	1.00(0.00)	0.00(0.00)
		MM95	0.00019	0.0380	0.5367	0.4445	1.00(0.00)	0.00(0.00)
		FSRws	0.00029	0.0407	1.3658	1.5142	0.00(0.01)	0.68(0.12)
		Heur	0.00025	0.0344	0.0063	0.9036	0.02(0.03)	0.56(0.12)
		SCADws	0.00015	0.0366	0.0691	0.8392	0.02(0.02)	0.58(0.11)
250	2	Opt	0.00014	0.0202	0.0037	0.1187	0.00(0.00)	0.00(0.00)
		OLS	0.00074	0.0384	78.5485	3.1807	0.00(0.00)	1.00(0.00)
		LTS	0.00115	0.0473	5.1892	0.0844	0.13(0.02)	0.37(0.05)
		MM85	0.00018	0.0248	0.5391	0.1883	1.00(0.00)	0.00(0.00)
		MM95	0.00013	0.0249	0.5225	0.3533	1.00(0.00)	0.00(0.00)
		FSRws	0.00013	0.0273	0.5807	1.2080	0.01(0.01)	0.64(0.11)
		Heur	0.00024	0.0261	0.0009	0.7908	0.02(0.02)	0.55(0.10)
		SCADws	0.00017	0.0250	0.0420	0.7442	0.02(0.02)	0.57(0.10)
300	2	Opt	0.00005	0.0169	0.0000	0.1176	0.00(0.00)	0.00(0.00)
		OLS	0.00052	0.0466	82.0060	3.4206	0.00(0.00)	1.00(0.00)
		LTS	0.00011	0.0400	4.9945	0.0703	0.13(0.02)	0.38(0.05)
		MM85	0.00021	0.0246	0.4337	0.1752	1.00(0.00)	0.00(0.00)
		MM95	0.00029	0.0261	0.7822	0.3896	1.00(0.00)	0.00(0.00)
		FSRws	0.00030	0.0260	0.3107	0.8086	0.01(0.01)	0.61(0.10)
		Heur	0.00034	0.0254	0.0651	0.5022	0.02(0.01)	0.57(0.09)
		SCADws	0.00031	0.0239	0.0922	0.4944	0.01(0.01)	0.57(0.09)

Table B.2: Scenario 1 with SNR = 2. MSE for $\widehat{\boldsymbol{\beta}}$ and $\widehat{s}^2$ (decomposed into squared bias and variance), and mean (SD in parenthesis) FPR and FNR for feature selection and outlier detection, based on 100 simulation replications.

n	p	Method	bias$(\widehat{\beta})^2$	var$(\widehat{\beta})$	bias$(\widehat{s}^2)^2$	var$(\widehat{s}^2)$	FPR$(\widehat{\tau})$	FNR$(\widehat{\tau})$
50	2	Opt	0.00046	0.0483	0.0014	0.1340	0.00(0.00)	0.00(0.00)
		OLS	0.00346	0.1257	17.9656	4.6867	0.00(0.00)	1.00(0.00)
		LTS	0.00129	0.1229	1.4131	0.0990	0.14(0.04)	0.37(0.13)
		MM85	0.00053	0.0743	0.1837	0.1791	1.00(0.00)	0.00(0.00)
		MM95	0.00070	0.0713	0.0990	0.3872	1.00(0.00)	0.00(0.00)
		FSRws	0.00200	0.1036	6.3637	3.6095	0.00(0.00)	0.91(0.14)
		Heur	0.00116	0.0739	0.0432	0.6911	0.07(0.11)	0.50(0.21)
		SCADws	0.00071	0.0769	0.0023	0.4642	0.03(0.04)	0.54(0.21)
100	2	Opt	0.00017	0.0291	0.0001	0.0811	0.00(0.00)	0.00(0.00)
		OLS	0.00090	0.0605	19.3149	2.3947	0.00(0.00)	1.00(0.00)
		LTS	0.00079	0.0733	1.2152	0.0496	0.13(0.03)	0.38(0.08)
		MM85	0.00043	0.0444	0.1129	0.1614	1.00(0.00)	0.00(0.00)
		MM95	0.00046	0.0431	0.1619	0.3067	1.00(0.00)	0.00(0.00)
		FSRws	0.00052	0.0477	1.8841	2.0425	0.00(0.00)	0.79(0.15)
		Heur	0.00073	0.0448	0.0000	0.5765	0.04(0.05)	0.55(0.16)
		SCADws	0.00056	0.0460	0.0086	0.4044	0.02(0.03)	0.57(0.14)
150	2	Opt	0.00034	0.0153	0.0003	0.0589	0.00(0.00)	0.00(0.00)
		OLS	0.00005	0.0371	20.4488	2.1283	0.00(0.00)	1.00(0.00)
		LTS	0.00005	0.0381	1.2904	0.0299	0.13(0.02)	0.38(0.07)
		MM85	0.00019	0.0227	0.1336	0.0838	1.00(0.00)	0.00(0.00)
		MM95	0.00027	0.0232	0.1529	0.1727	1.00(0.00)	0.00(0.00)
		FSRws	0.00019	0.0250	0.6756	0.5420	0.00(0.01)	0.71(0.13)
		Heur	0.00016	0.0243	0.0017	0.2759	0.03(0.03)	0.53(0.12)
		SCADws	0.00015	0.0238	0.0011	0.2509	0.02(0.03)	0.55(0.12)
200	2	Opt	0.00016	0.0130	0.0010	0.0436	0.00(0.00)	0.00(0.00)
		OLS	0.00026	0.0306	19.0662	1.4298	0.00(0.00)	1.00(0.00)
		LTS	0.00005	0.0255	1.2568	0.0227	0.13(0.02)	0.38(0.06)
		MM85	0.00007	0.0170	0.1333	0.0567	1.00(0.00)	0.00(0.00)
		MM95	0.00009	0.0190	0.1342	0.1111	1.00(0.00)	0.00(0.00)
		FSRws	0.00014	0.0203	0.3415	0.3786	0.00(0.01)	0.68(0.12)
		Heur	0.00013	0.0176	0.0027	0.2158	0.02(0.03)	0.57(0.11)
		SCADws	0.00009	0.0182	0.0173	0.2128	0.01(0.02)	0.58(0.11)
250	2	Opt	0.00007	0.0101	0.0009	0.0297	0.00(0.00)	0.00(0.00)
		OLS	0.00037	0.0192	19.6371	0.7952	0.00(0.00)	1.00(0.00)
		LTS	0.00057	0.0236	1.2973	0.0211	0.13(0.02)	0.37(0.05)
		MM85	0.00009	0.0124	0.1348	0.0471	1.00(0.00)	0.00(0.00)
		MM95	0.00006	0.0124	0.1306	0.0883	1.00(0.00)	0.00(0.00)
		FSRws	0.00006	0.0136	0.1452	0.3020	0.01(0.01)	0.64(0.11)
		Heur	0.00013	0.0126	0.0010	0.1983	0.02(0.02)	0.56(0.10)
		SCADws	0.00009	0.0126	0.0116	0.1881	0.02(0.02)	0.57(0.10)
300	2	Opt	0.00003	0.0084	0.0000	0.0294	0.00(0.00)	0.00(0.00)
		OLS	0.00026	0.0233	20.5015	0.8552	0.00(0.00)	1.00(0.00)
		LTS	0.00005	0.0200	1.2486	0.0176	0.13(0.02)	0.38(0.05)
		MM85	0.00011	0.0123	0.1084	0.0438	1.00(0.00)	0.00(0.00)
		MM95	0.00015	0.0130	0.1956	0.0974	1.00(0.00)	0.00(0.00)
		FSRws	0.00015	0.0130	0.0777	0.2022	0.01(0.01)	0.61(0.10)
		Heur	0.00014	0.0125	0.0147	0.1304	0.02(0.02)	0.57(0.09)
		SCADws	0.00015	0.0122	0.0246	0.1211	0.01(0.01)	0.57(0.09)

Table B.3: Scenario 1 with SNR = 5. MSE for $\widehat{\boldsymbol{\beta}}$ and $\widehat{s}^2$ (decomposed into squared bias and variance), and mean (SD in parenthesis) FPR and FNR for feature selection and outlier detection, based on 100 simulation replications.

n	p	Method	bias$(\widehat{\beta})^2$	var$(\widehat{\beta})$	bias$(\widehat{s}^2)^2$	var$(\widehat{s}^2)$	FPR$(\widehat{\tau})$	FNR$(\widehat{\tau})$
50	2	Opt	0.00018	0.0193	0.0002	0.0214	0.00(0.00)	0.00(0.00)
		OLS	0.00138	0.0503	2.8745	0.7499	0.00(0.00)	1.00(0.00)
		LTS	0.00052	0.0492	0.2261	0.0158	0.14(0.04)	0.37(0.13)
		MM85	0.00021	0.0297	0.0294	0.0287	1.00(0.00)	0.00(0.00)
		MM95	0.00028	0.0285	0.0158	0.0619	1.00(0.00)	0.00(0.00)
		FSRws	0.00080	0.0414	1.0182	0.5775	0.00(0.00)	0.91(0.14)
		Heur	0.00048	0.0302	0.0067	0.1092	0.06(0.10)	0.50(0.21)
		SCADws	0.00027	0.0301	0.0002	0.0747	0.03(0.04)	0.54(0.21)
100	2	Opt	0.00007	0.0116	0.0000	0.0130	0.00(0.00)	0.00(0.00)
		OLS	0.00036	0.0242	3.0904	0.3831	0.00(0.00)	1.00(0.00)
		LTS	0.00032	0.0293	0.1944	0.0079	0.13(0.03)	0.38(0.08)
		MM85	0.00017	0.0178	0.0181	0.0258	1.00(0.00)	0.00(0.00)
		MM95	0.00018	0.0173	0.0259	0.0491	1.00(0.00)	0.00(0.00)
		FSRws	0.00021	0.0191	0.3015	0.3268	0.00(0.00)	0.79(0.15)
		Heur	0.00031	0.0181	0.0000	0.0901	0.04(0.05)	0.56(0.16)
		SCADws	0.00027	0.0180	0.0017	0.0688	0.02(0.03)	0.58(0.14)
150	2	Opt	0.00014	0.0061	0.0000	0.0094	0.00(0.00)	0.00(0.00)
		OLS	0.00002	0.0148	3.2718	0.3405	0.00(0.00)	1.00(0.00)
		LTS	0.00002	0.0152	0.2065	0.0048	0.13(0.02)	0.38(0.07)
		MM85	0.00007	0.0091	0.0214	0.0134	1.00(0.00)	0.00(0.00)
		MM95	0.00011	0.0093	0.0245	0.0276	1.00(0.00)	0.00(0.00)
		FSRws	0.00008	0.0100	0.1081	0.0867	0.00(0.01)	0.71(0.13)
		Heur	0.00008	0.0099	0.0003	0.0420	0.03(0.03)	0.53(0.12)
		SCADws	0.00006	0.0095	0.0002	0.0407	0.02(0.03)	0.55(0.12)
200	2	Opt	0.00007	0.0052	0.0002	0.0070	0.00(0.00)	0.00(0.00)
		OLS	0.00011	0.0123	3.0506	0.2288	0.00(0.00)	1.00(0.00)
		LTS	0.00002	0.0102	0.2011	0.0036	0.13(0.02)	0.38(0.06)
		MM85	0.00003	0.0068	0.0213	0.0091	1.00(0.00)	0.00(0.00)
		MM95	0.00004	0.0076	0.0215	0.0178	1.00(0.00)	0.00(0.00)
		FSRws	0.00006	0.0081	0.0546	0.0606	0.00(0.01)	0.68(0.12)
		Heur	0.00004	0.0070	0.0004	0.0376	0.02(0.03)	0.57(0.12)
		SCADws	0.00003	0.0071	0.0026	0.0347	0.02(0.02)	0.58(0.11)
250	2	Opt	0.00003	0.0040	0.0001	0.0047	0.00(0.00)	0.00(0.00)
		OLS	0.00015	0.0077	3.1419	0.1272	0.00(0.00)	1.00(0.00)
		LTS	0.00023	0.0095	0.2076	0.0034	0.13(0.02)	0.37(0.05)
		MM85	0.00004	0.0050	0.0216	0.0075	1.00(0.00)	0.00(0.00)
		MM95	0.00003	0.0050	0.0209	0.0141	1.00(0.00)	0.00(0.00)
		FSRws	0.00003	0.0055	0.0232	0.0483	0.01(0.01)	0.64(0.11)
		Heur	0.00004	0.0050	0.0001	0.0314	0.02(0.02)	0.55(0.10)
		SCADws	0.00004	0.0049	0.0018	0.0293	0.02(0.02)	0.57(0.09)
300	2	Opt	0.00001	0.0034	0.0000	0.0047	0.00(0.00)	0.00(0.00)
		OLS	0.00010	0.0093	3.2802	0.1368	0.00(0.00)	1.00(0.00)
		LTS	0.00002	0.0080	0.1998	0.0028	0.13(0.02)	0.38(0.05)
		MM85	0.00004	0.0049	0.0173	0.0070	1.00(0.00)	0.00(0.00)
		MM95	0.00006	0.0052	0.0313	0.0156	1.00(0.00)	0.00(0.00)
		FSRws	0.00006	0.0052	0.0124	0.0323	0.01(0.01)	0.61(0.10)
		Heur	0.00006	0.0048	0.0026	0.0214	0.02(0.02)	0.57(0.09)
		SCADws	0.00005	0.0049	0.0040	0.0198	0.01(0.01)	0.57(0.09)

Scenario 2: high-dimensional VIOMs and MSOMs

We also explored weaker SNR regimes for the simulation scenario 2 in Section 4.4.2. Setting SNR = 3, Table B.4 shows simulation results in term of MSE decompositions for $\widehat{\boldsymbol{\beta}}$ and $\widehat{s}^2$, and mean (with standard deviations in parenthesis) FPR and FNR for feature selection and outlier detection. These results are consistent with the ones presented in Table 4.1. However, all methods experience an overall decrease in performance, which is more marked for the FNR in $\widehat{\boldsymbol{\beta}}$ for smaller sample size. Overall, SCAD2s outperforms other feasible estimation methods also in this setting

We also explored simulation scenarios with denser models. The following simulation setting is the same as scenario 2 in Section 4.4.2, with the only difference being that the true sparsity level is increased to $p_0 = 7$ (including the intercept term). Table B.5 shows simulation results in term of MSE decompositions for $\widehat{\boldsymbol{\beta}}$ and $\widehat{s}^2$, and mean (with standard deviations in parenthesis) FPR and FNR for feature selection and outlier detection. The results are again consistent with the ones presented in Table 4.1, although all methods experience a decrease in performance for smaller sample sizes, with an overall increase in the FNR for $\widehat{\boldsymbol{\beta}}$ and the MSE for s^2. Also in this setting, SCAD2s converges faster to the oracle solution compared to other feasible estimation methods.

Finally, we explored simulation settings with multicollinearity in the predictors. Table B.6 presents simulation results for a setting as scenario 2 in Section 4.4.2, but the covariance matrix $\boldsymbol{\Sigma}_{p-1}$ has an autoregressive correlation structure $\boldsymbol{\Sigma}_{p-1,ij} = \eta^{|i-j|}$ with $\eta = 0.4$. Note that the level of multicollinearity increases after MSOM contamination. Overall, also here the results are consistent with the ones presented in Table 4.1. SCADws and SCAD2s suffer for small sample sizes (higher FNR in $\widehat{\boldsymbol{\beta}}$ compared to previous settings) but, as n increases, we can again notice that SCAD2s outperforms other feasible estimation methods and converges faster to the oracle solution.

Table B.4: Scenario 2 with SNR = 3. MSE for $\widehat{\boldsymbol{\beta}}$ and $\widehat{s}^2$ (decomposed into squared bias and variance), and mean (SD in parenthesis) FPR and FNR for feature selection and outlier detection, based on 100 simulation replications.

n	p	Method	bias$(\widehat{\beta})^2$	var$(\widehat{\beta})$	bias$(\widehat{s}^2)^2$	var$(\widehat{s}^2)$	FPR$(\widehat{\beta})$	FNR$(\widehat{\beta})$	FPR$(\widehat{\tau})$	FNR$(\widehat{\tau})$
100	150	Opt	0.00001	0.00156	0.01	0.29	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.08323	0.00575	553.76	30.95	0.03(0.05)	0.60(0.16)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.01019	0.07271	91.08	45.08	0.45(0.04)	0.09(0.21)	0.00(0.01)	0.84(0.02)
		TaL	0.00973	0.03453	5.28	12.19	0.05(0.08)	0.14(0.28)	1.00(0.00)	0.00(0.00)
		Heur	0.00389	0.05639	4.63	13.63	0.08(0.05)	0.12(0.26)	0.01(0.02)	0.75(0.10)
		SCADopt	0.00218	0.01366	0.07	2.98	0.00(0.00)	0.13(0.27)	0.07(0.15)	0.45(0.14)
		SCADws	0.00675	0.03739	3.27	9.30	0.01(0.02)	0.22(0.28)	0.04(0.07)	0.57(0.14)
		SCAD2s	0.00414	0.01803	0.29	5.33	0.00(0.00)	0.18(0.30)	0.10(0.17)	0.50(0.16)
150	150	Opt	0.00001	0.00096	0.03	0.25	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.08436	0.00357	556.31	15.87	0.03(0.04)	0.57(0.17)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.00224	0.04584	108.00	14.74	0.53(0.06)	0.01(0.07)	0.00(0.00)	0.84(0.01)
		TaL	0.00164	0.00605	6.02	2.55	0.01(0.02)	0.02(0.10)	1.00(0.00)	0.00(0.00)
		Heur	0.00032	0.02035	1.89	3.84	0.06(0.04)	0.02(0.09)	0.01(0.03)	0.66(0.12)
		SCADopt	0.00004	0.00319	0.11	0.96	0.00(0.00)	0.02(0.11)	0.01(0.06)	0.48(0.10)
		SCADws	0.00043	0.01106	0.32	3.70	0.01(0.02)	0.04(0.12)	0.02(0.03)	0.51(0.11)
		SCAD2s	0.00011	0.00426	0.04	1.25	0.00(0.00)	0.02(0.13)	0.03(0.08)	0.49(0.10)
200	150	Opt	0.00000	0.00065	0.06	0.13	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.08580	0.00277	596.61	17.34	0.03(0.04)	0.47(0.18)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.00147	0.03920	91.42	10.30	0.58(0.05)	0.01(0.07)	0.00(0.00)	0.84(0.01)
		TaL	0.00115	0.00597	4.46	1.72	0.01(0.03)	0.02(0.09)	1.00(0.00)	0.00(0.00)
		Heur	0.00022	0.01617	1.49	4.44	0.06(0.05)	0.01(0.06)	0.01(0.01)	0.62(0.11)
		SCADopt	0.00002	0.00205	0.20	1.68	0.00(0.00)	0.01(0.07)	0.00(0.01)	0.48(0.09)
		SCADws	0.00016	0.00720	0.04	2.17	0.00(0.01)	0.02(0.11)	0.02(0.03)	0.49(0.11)
		SCAD2s	0.00009	0.00393	0.03	3.57	0.00(0.00)	0.02(0.12)	0.03(0.06)	0.48(0.10)

Table B.5: Scenario 2 with $p_0 = 7$. MSE for $\widehat{\boldsymbol{\beta}}$ and $\widehat{s}^2$ (decomposed into squared bias and variance), and mean (SD in parenthesis) FPR and FNR for feature selection and outlier detection, based on 100 simulation replications.

n	p	Method	bias$(\widehat{\beta})^2$	var$(\widehat{\beta})$	bias$(\widehat{s}^2)^2$	var$(\widehat{s}^2)$	FPR$(\widehat{\beta})$	FNR$(\widehat{\beta})$	FPR$(\widehat{\tau})$	FNR$(\widehat{\tau})$
100	150	Opt	0.00003	0.00344	0.09	0.42	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.16356	0.00803	1459.98	101.12	0.03(0.05)	0.79(0.09)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.01549	0.09663	179.57	62.40	0.44(0.04)	0.08(0.22)	0.00(0.01)	0.84(0.02)
		TaL	0.01809	0.04415	19.30	24.19	0.03(0.06)	0.14(0.29)	1.00(0.00)	0.00(0.00)
		Heur	0.00492	0.08776	4.95	19.41	0.10(0.04)	0.10(0.26)	0.02(0.03)	0.75(0.11)
		SCADopt	0.00396	0.02762	0.39	7.17	0.00(0.00)	0.14(0.31)	0.06(0.16)	0.47(0.12)
		SCADws	0.01647	0.06730	31.19	49.21	0.01(0.02)	0.31(0.29)	0.05(0.06)	0.63(0.14)
		SCAD2s	0.01526	0.04684	9.21	33.77	0.00(0.01)	0.28(0.34)	0.11(0.18)	0.56(0.16)
150	150	Opt	0.00001	0.00225	0.06	0.34	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.16330	0.00431	1395.76	42.30	0.02(0.03)	0.76(0.12)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.00393	0.05647	186.48	32.35	0.53(0.06)	0.00(0.04)	0.00(0.00)	0.84(0.01)
		TaL	0.00373	0.00816	13.12	5.35	0.01(0.01)	0.01(0.10)	1.00(0.00)	0.00(0.00)
		Heur	0.00035	0.03514	2.95	4.55	0.10(0.04)	0.01(0.07)	0.01(0.01)	0.69(0.09)
		SCADopt	0.00006	0.00481	0.08	2.03	0.00(0.00)	0.01(0.09)	0.01(0.05)	0.47(0.11)
		SCADws	0.00154	0.02003	2.37	15.35	0.01(0.02)	0.07(0.13)	0.03(0.03)	0.52(0.13)
		SCAD2s	0.00024	0.00821	0.37	8.18	0.00(0.00)	0.02(0.11)	0.03(0.05)	0.52(0.12)
200	150	Opt	0.00002	0.00146	0.07	0.23	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.16476	0.00431	1496.92	30.59	0.03(0.05)	0.72(0.12)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.00205	0.04376	149.62	17.96	0.58(0.05)	0.00(0.00)	0.00(0.00)	0.83(0.00)
		TaL	0.00170	0.00414	8.56	3.25	0.01(0.01)	0.00(0.00)	1.00(0.00)	0.00(0.00)
		Heur	0.00023	0.02021	2.24	3.31	0.08(0.04)	0.00(0.00)	0.01(0.02)	0.64(0.11)
		SCADopt	0.00004	0.00199	0.12	0.64	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.47(0.08)
		SCADws	0.00006	0.00437	0.08	2.83	0.00(0.00)	0.01(0.04)	0.02(0.02)	0.48(0.08)
		SCAD2s	0.00004	0.00257	0.02	1.32	0.00(0.00)	0.00(0.00)	0.02(0.02)	0.49(0.08)

Table B.6: Scenario 2 in presence of multicollinearity. MSE for $\widehat{\boldsymbol{\beta}}$ and $\widehat{s}^2$ (decomposed into squared bias and variance), and mean (SD in parenthesis) FPR and FNR for feature selection and outlier detection, based on 100 simulation replications.

n	p	Method	bias$(\widehat{\beta})^2$	var$(\widehat{\beta})$	bias$(\widehat{s}^2)^2$	var$(\widehat{s}^2)$	FPR$(\widehat{\beta})$	FNR$(\widehat{\beta})$	FPR$(\widehat{\tau})$	FNR$(\widehat{\tau})$
100	150	Opt	0.00000	0.00196	0.05	0.31	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.08193	0.00487	947.37	40.57	0.02(0.04)	0.72(0.09)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.00175	0.03819	96.19	33.95	0.34(0.04)	0.01(0.06)	0.00(0.00)	0.83(0.01)
		TaL	0.00132	0.01112	5.33	8.07	0.01(0.02)	0.03(0.11)	1.00(0.00)	0.00(0.00)
		Heur	0.00036	0.02982	1.32	4.26	0.06(0.04)	0.03(0.11)	0.01(0.03)	0.67(0.14)
		SCADopt	0.00004	0.00683	0.51	5.06	0.00(0.00)	0.03(0.11)	0.00(0.01)	0.47(0.13)
		SCADws	0.00108	0.02531	3.49	11.05	0.00(0.01)	0.16(0.17)	0.03(0.04)	0.55(0.14)
		SCAD2s	0.00022	0.01127	1.50	7.22	0.00(0.00)	0.06(0.15)	0.03(0.05)	0.55(0.14)
150	150	Opt	0.00000	0.00128	0.04	0.22	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.08190	0.00387	1007.29	29.58	0.03(0.03)	0.71(0.10)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.00102	0.02115	90.18	14.03	0.36(0.05)	0.00(0.00)	0.00(0.00)	0.84(0.00)
		TaL	0.00034	0.00387	4.11	1.72	0.01(0.02)	0.00(0.00)	1.00(0.00)	0.00(0.00)
		Heur	0.00019	0.01976	1.40	2.69	0.07(0.05)	0.00(0.02)	0.01(0.03)	0.65(0.12)
		SCADopt	0.00001	0.00152	0.03	0.60	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.45(0.09)
		SCADws	0.00005	0.00857	0.19	3.61	0.00(0.01)	0.04(0.11)	0.02(0.02)	0.49(0.10)
		SCAD2s	0.00001	0.00334	0.06	2.65	0.00(0.00)	0.01(0.06)	0.02(0.02)	0.48(0.10)
200	150	Opt	0.00002	0.00086	0.03	0.15	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Lasso	0.08323	0.00521	1021.70	22.19	0.04(0.05)	0.67(0.14)	0.00(0.00)	1.00(0.00)
		SparseLTS	0.00064	0.01530	91.55	7.03	0.37(0.04)	0.00(0.00)	0.00(0.00)	0.83(0.00)
		TaL	0.00020	0.00240	4.20	1.01	0.01(0.01)	0.00(0.00)	1.00(0.00)	0.00(0.00)
		Heur	0.00013	0.01152	1.32	2.12	0.06(0.04)	0.00(0.00)	0.01(0.02)	0.61(0.10)
		SCADopt	0.00003	0.00113	0.14	0.54	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.47(0.07)
		SCADws	0.00002	0.00249	0.00	1.06	0.00(0.00)	0.01(0.04)	0.02(0.02)	0.47(0.08)
		SCAD2s	0.00003	0.00142	0.00	0.92	0.00(0.00)	0.00(0.00)	0.02(0.02)	0.47(0.08)

Table B.7: Estimated coefficients across different methods for the Boston housing data.

Variable	OLS	LTS	MM85	MM95	FSRws	Lasso	sparseLTS	TaL	Heur	SCADws	SCAD2s
intercept	36.459	13.481	13.862	7.275	8.152	35.455	-2.666	-2.600	7.893	4.942	-0.571
crim	-0.108	-0.089	-0.145	-0.129	-0.156	-0.106	-0.053	-0.117	-0.109		
zn	0.046	0.025	0.029	0.027	0.021	0.046			0.028		
indus	0.021	0.036	-0.004	-0.015	0.000				-0.008		
chas	2.687	1.394	1.390	1.221	1.320	2.743	0.889	1.645	1.300		
nox	-17.767	-10.203	-6.042	-5.912	-4.917	-16.232			-5.698		
rm	3.810	6.268	4.663	6.092	5.196	3.735	6.332	6.430	6.221	6.354	6.400
age	0.001	-0.032	-0.039	-0.042	-0.044		-0.008		-0.036		
dis	-1.476	-1.093	-0.878	-0.942	-0.811	-1.453	-0.028	-0.216	-0.933	-0.315	-0.334
rad	0.306	0.168	0.144	0.150	0.142	0.281			0.155		
tax	-0.012	-0.012	-0.009	-0.011	-0.008	-0.011	-0.005		-0.012	-0.010	-0.008
ptratio	-0.953	-0.876	-0.602	-0.685	-0.579	-0.907	-0.668	-0.807	-0.711	-0.687	-0.714
black	0.009	0.011	0.013	0.013	0.014	0.009	0.009	0.014	0.011		0.012
lstat	-0.525	-0.278	-0.267	-0.219	-0.228	-0.532	-0.285	-0.388	-0.266	-0.406	-0.374

B.4 Application study details

The methods used in our application studies mimic the ones discussed for our simulations. On Boston housing data, for SCADws and SCAD2s, we replace the second term in (B.8) with $2\log(\bar{n})k_p^{\lambda_k}$ to enforce sparser solutions. TaL uses an S-estimator with 50% BdP as a preliminary estimate of scale, which is computed on the active set estimated by sparseLTS. However, for glioblastoma gene expression data, TaL uses a preliminary OLS estimator computed on the set of active features and non-outlying cases estimated by sparseLTS.

B.4.1 Boston housing data: extended comparison

Table B.7 shows the estimated regression coefficients by different methods on Boston housing data. SCADws, SCAD2s and TaL provide sparser solutions and more interpretable models. Namely, SCAD2s selects only 6 predictors: *rm*, *dis*, *tax*, *ptratio*, *black* and *lstat* (plus the intercept term). SCADws selects the same set of features, except from *black*. TaL provides a slightly denser solution and selects 7 predictors: the ones selected by SCAD2s, aside from *tax*, plus *crim* and *chas*. Other methods provide much denser solutions.

B.5 Software availability

Source code for computing our proposals, and to replicate our simulation and application studies, is available at https://github.com/LucaIns/doubly_robust_sparse.

Bibliography

Agostinelli, C., Leung, A., Yohai, V. J., and Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, 24(3):441–461.

Agostinelli, C. and Yohai, V. J. (2016). Composite robust estimators for linear mixed models. *Journal of the American Statistical Association*, 111(516):1764–1774.

Agulló, J. (2001). New algorithms for computing the least trimmed squares regression estimator. *Computational Statistics & Data Analysis*, 36(4):425–439.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Alfons, A. (2021). robustHD: An R package for robust regression with high-dimensional data. *Journal of Open Source Software*, 6(67):3786.

Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7(1):226–248.

Algamal, Z. Y. and Lee, M. H. (2015). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in Biology and Medicine*, 67:136–145.

Alqallaf, F. A., Van Aelst, S., Yohai, V. J., and Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, 37(1):311–331.

Amato, U., Antoniadis, A., De Feis, I., and Gijbels, I. (2021). Penalised robust estimators for sparse and high-dimensional linear models. *Statistical Methods & Applications*, 30(1):1–48.

Aneiros, G., Cao, R., Fraiman, R., Genest, C., and Vieu, P. (2019). Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis*, 170:3–9.

Atamturk, A. and Gómez, A. (2020). Safe screening rules for L0-regression from perspective relaxations. In *Proceedings of Machine Learning Research*, volume 119, pages 421–430. PMLR.

Atkinson, A. C. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Clarendon Press, Oxford.

Atkinson, A. C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer–Verlag, New York.

Atkinson, A. C. and Riani, M. (2001). Regression diagnostics for binomial data from the forward search. *Journal of the Royal Statistical Society: Series D*, 50(1):63–78.

Atkinson, A. C. and Riani, M. (2006). Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics*, 15(2):460–476.

Atkinson, A. C., Riani, M., and Torti, F. (2016). Robust methods for heteroskedastic regression. *Computational Statistics & Data Analysis*, 104:209–222.

Avella Medina, M. A. (2016). *Robust Penalized M-estimators for Generalized Linear and Additive Models*. PhD thesis, University of Geneva.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. Wiley, New York, 3rd edition.

Barratt, S., Angeris, G., and Boyd, S. (2020). Minimizing a sum of clipped convex functions. *Optimization Letters*, 14(8):2443–2459.

Barut, E., Fan, J., and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277.

Becher, M. A., Osborne, J. L., Thorbek, P., Kennedy, P. J., and Grimm, V. (2013). Towards a systems approach for understanding honeybee decline: A stocktaking and synthesis of existing models. *Journal of Applied Ecology*, 50(4):868–880.

Beckman, R. J. and Cook, R. D. (1983). Outlier .... s. *Technometrics*, 25(2):119–149.

Bedrick, E. J. and Hill, J. R. (1990). Outlier tests for logistic regression: A conditional approach. *Biometrika*, 77(4):815–827.

Bee Informed Partnership (2020). 2019-2020 honey bee colony losses in the United States: Preliminary results.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2004). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

Bernholt, T. (2006). Robust estimators are hard to compute. Technical Report 52/2005, University of Dortmund, Dortmund, Germany.

Bertsimas, D. and Copenhaver, M. S. (2017). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 270(3):931–942.

Bertsimas, D., Cory-Wright, R., and Pauphilet, J. (2021a). A unified approach to mixed-integer optimization problems with logical constraints. *SIAM Journal on Optimization*, 31(3):2340–2367.

Bertsimas, D., King, A., et al. (2017). Logistic regression: From art to science. *Statistical Science*, 32(3):367–384.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852.

Bertsimas, D. and Mazumder, R. (2014). Least quantile regression via modern optimization. *The Annals of Statistics*, 42(6):2494–2525.

Bertsimas, D., Pauphilet, J., and Van Parys, B. (2021b). Sparse classification: A scalable discrete optimization perspective. *Machine Learning*, 110(11):3177–3209.

Bertsimas, D. and Van Parys, B. (2020). Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *The Annals of Statistics*, 48(1):300–323.

Beyer, M., Junk, J., Eickermann, M., Clermont, A., Kraus, F., Georges, C., Reichart, A., and Hoffmann, L. (2018). Winter honey bee colony losses, Varroa destructor control strategies, and the role of weather conditions: Results from a survey among beekeepers. *Research in Veterinary Science*, 118:52–60.

Bianco, A. M., Boente, G., and Chebi, G. (2021). Penalized robust estimators in sparse logistic regression. *TEST*, Forthcoming:1–32.

Bianco, A. M., Boente, G., and Chebi, G. (2022). Asymptotic behaviour of penalized robust estimators in logistic regression when dimension increases. *arXiv preprint arXiv:2201.12449*.

Bianco, A. M. and Yohai, V. J. (1996). Robust estimation in the logistic regression model. In Rieder, H., editor, *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday*, pages 17–34. Springer New York.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077.

Borio, D. (2017). Robust signal processing for GNSS. In *2017 European Navigation Conference (ENC)*, pages 150–158. IEEE.

Bottmer, L., Croux, C., and Wilms, I. (2022). Sparse regression for large data sets with outliers. *European Journal of Operational Research*, 297(2):782–794.

Box, G. E. (1953). Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335.

Box, G. E. and Tiao, G. C. (1968). A Bayesian approach to some outlier problems. *Biometrika*, 55(1):119–129.

Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B*, 73(3):325–349.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Bruckner, S., Nathalie, S., Jonathan, E., Anne Marie, F., Kelly, K., Eric, M., Annette, M., Meghan, M., Elina, N., Juliana, R., Karen, R., Daniel, R., Ramesh, S., Jennifer, T., Dennis, v., S. Dan, A., Michaela, W., and Geoffrey, W. (2020). 2019-2020 honey bee colony losses in the United States: Preliminary results. *Technical report*, 579:1581–1587.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin.

Buscemi, S. and Plaia, A. (2020). Model selection in linear mixed-effect models. *AStA Advances in Statistical Analysis*, 104(4):529–575.

Cabrera-Rubio, R., Collado, M. C., Laitinen, K., Salminen, S., Isolauri, E., and Mira, A. (2012). The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery. *The American Journal of Clinical Nutrition*, 96(3):544–551.

Cai, X., Xue, L., and Cao, J. (2021). Robust penalized M-estimation for function-on-function linear regression. *Stat*, 10(1):e390.

Cai, X., Xue, L., Cao, J., and Initiative, A. D. N. (2022). Robust estimation and variable selection for function-on-scalar regression. *Canadian Journal of Statistics*, 50(1):162–179.

Calderone, N. W. (2012). Insect pollinated crops, insect pollinators and US agriculture: Trend analysis of aggregate data for the period 1992–2009. *PLoS One*, 7(5):1–27.

Calovi, M., Grozinger, C. M., Miller, D. A., and Goslee, S. C. (2021). Summer weather conditions influence winter survival of honey bees (*Apis mellifera*) in the northeastern United States. *Scientific Reports*, 11(1553).

Carroll, R. J. and Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society: Series B*, 55(3):693–706.

Cerioli, A., Atkinson, A. C., and Riani, M. (2011). Some Perspectives on Multivariate Outlier Detection. In Ingrassia, S., Rocci, R., and Vichi, M., editors, *New Perspectives in Statistical Modeling and Data Analysis*, pages 231–238. Springer.

Cerioli, A., Atkinson, A. C., and Riani, M. (2016). How to marry robustness and applied statistics. In *Topics on Methodological and Applied Statistical Inference*, pages 51–64. Springer.

Cerioli, A., Farcomeni, A., and Riani, M. (2014). Strong consistency and robustness of the Forward Search estimator of multivariate location and scatter. *Journal of Multivariate Analysis*, 126:167–183.

Cerioli, A., Riani, M., Atkinson, A. C., and Corbellini, A. (2018). The power of monitoring: How to make the most of a contaminated multivariate sample. *Statistical Methods & Applications*, 27:1–29.

Chang, L., Roberts, S., and Welsh, A. (2018). Robust Lasso Regression Using Tukey's Biweight Criterion. *Technometrics*, 60(1):36–47.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*. John Wiley & Sons, New York.

Chervoneva, I. and Vishnyakov, M. (2011). Constrained S-estimators for linear mixed effects models with covariance components. *Statistics in Medicine*, 30(14):1735–1750.

Chervoneva, I. and Vishnyakov, M. (2014). Generalized S-estimators for linear mixed effects models. *Statistica Sinica*, 24(3):1257–1276.

Chopra, S. S., Bakshi, B. R., and Khanna, V. (2015). Economic dependence of U.S. industrial sectors on animal-mediated pollination service. *Environmental Science & Technology*, 49(24):14441—14451.

Christmann, A. (1994). Least median of weighted squares in logistic regression with large strata. *Biometrika*, 81(2):413–417.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society: Series B*, 48(2):133–155.

Cook, R. D., Holschuh, N., and Weisberg, S. (1982). A note on an alternative outlier model. *Journal of the Royal Statistical Society: Series B*, 44(3):370–376.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

Copas, J. B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society: Series B*, 50(2):225–253.

Copt, S. and Victoria-Feser, M.-P. (2006). High-breakdown inference for mixed linear models. *Journal of the American Statistical Association*, 101(473):292–300.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*. Chapman & Hall, London, second edition.

Craig, S. J., Blankenberg, D., Parodi, A. C. L., Paul, I. M., Birch, L. L., Savage, J. S., et al. (2018). Child weight gain trajectories linked to oral microbiota composition. *Scientific Reports*, 8(14030).

Cramer, J. S. (2002). The origins of logistic regression. Technical Report 2002-119/4, Tinbergen Institute, Amsterdam, The Netherlands.

Croux, C., Flandre, C., and Haesbroeck, G. (2002). The breakdown behavior of the maximum likelihood estimator in the logistic regression model. *Statistics & Probability Letters*, 60(4):377–386.

Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics & Data Analysis*, 44(1-2):273–295.

Dai, W., Mrkvička, T., Sun, Y., and Genton, M. G. (2020). Functional outlier detection and taxonomy by sequential transformations. *Computational Statistics & Data Analysis*, 149:106960.

dbGaP (2017). INSIGHT cohort microbiome study. http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001498.v1.p1.

De Finetti, B. (1961). The Bayesian approach to the rejection of outliers. In *Proceedings of the Fourth Berkeley Symposium on Probability and Statistics*, volume 1, pages 199–210. University of California Press Berkeley.

Denhere, M. and Billor, N. (2016). Robust principal component functional logistic regression. *Communications in Statistics - Simulation and Computation*, 45(1):264–281.

Deza, A. and Atamturk, A. (2022). Safe screening for logistic regression with $\ell_0 - \ell_2$ regularization. *arXiv preprint arXiv:2202.00467*.

Döke, M. A., Frazier, M., and Grozinger, C. M. (2015). Overwintering honey bees: Biology and management. *Current Opinion in Insect Science*, 10:185–193.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In Bickel, P., Doksum, K. A., and Hodges, J. L., editors, *A Festschrift for Erich L. Lehmann*, pages 157–184. Wadsworth.

Douglas, M. R., Sponsler, D. B., Lonsdorf, E. V., and Grozinger, C. M. (2020). County-level analysis reveals a rapidly shifting landscape of insecticide hazard to honey bees *Apis mellifera* on US farmland. *Scientific Reports*, 10(797).

Duffy, D. E. and Santner, T. J. (1989). On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Communications in Statistics - Theory and Methods*, 18(3):959–980.

Ellis, J. D., Evans, J. D., and Pettis, J. (2010). Colony losses, managed colony population decline, and Colony Collapse Disorder in the United States. *Journal of Apicultural Research*, 49(1):134–136.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.

Fan, J., Guo, S., and Hao, N. (2012a). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B*, 74(1):37–65.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5):849–911.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.

Fan, J., Xue, L., and Zou, H. (2014a). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42(3):819–849.

Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. *The Annals of Statistics*, 40(4):2043–2068.

Fan, Y., Qin, G., and Zhu, Z. (2012b). Variable selection in robust regression models for longitudinal data. *Journal of Multivariate Analysis*, 109:156–167.

Fan, Y., Qin, G., and Zhu, Z. Y. (2014b). Robust variable selection in linear mixed models. *Communications in Statistics - Theory and Methods*, 43(21):4566–4581.

Farcomeni, A. (2014). Robust constrained clustering in presence of entry-wise outliers. *Technometrics*, 56(1):102–111.

Ferigato, C., Insolia, L., Perrotta, D., and Sordini, E. (2018). First report on the collaboration E.2/I.3 on the application of robust statistical techniques to the computation of position from satellite data in the context of project 346 GALILEO, WP 848 GAL-innova. Technical Report JRC110442, Joint Research Centre, European Commission.

Filzmoser, P., Höppner, S., Ortner, I., Serneels, S., and Verdonck, T. (2020). Cellwise robust M regression. *Computational Statistics & Data Analysis*, 147:106944.

Frangioni, A. and Gentile, C. (2006). Perspective cuts for a class of convex 0–1 mixed integer programs. *Mathematical Programming*, 106(2):225–236.

Frangioni, A. and Gentile, C. (2009). A computational comparison of reformulations of the perspective relaxation: SOCP vs. cutting planes. *Operations Research Letters*, 37(3):206–210.

Freue, G. V. C., Kepplinger, D., Salibián-Barrera, M., and Smucler, E. (2019). Robust elastic net estimators for variable selection and identification of proteomic biomarkers. *The Annals of Applied Statistics*, 13(4):2065–2090.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

FSDA (2022). Flexible Statistics & Data Analysis toolbox for MATLAB, with extensions to R, SAS and other platforms. https://github.com/UniprJRC/FSDA.

Gaglione, S., Innac, A., Carbone, S. P., Troisi, S., and Angrisano, A. (2017). Robust estimation methods applied to GPS in harsh environments. In *2017 European Navigation Conference (ENC)*, pages 14–25. IEEE.

García-Escudero, L. A., Gordaliza, A., Mayo-Íscar, A., and San Martín, R. (2010). Robust clusterwise linear regression through trimming. *Computational Statistics & Data Analysis*, 54(12):3057–3069.

Gatu, C., Yanev, P. I., and Kontoghiorghes, E. J. (2007). A graph approach to generate all possible regression submodels. *Computational Statistics & Data Analysis*, 52(2):799–815.

Genersch, E., Von Der Ohe, W., Kaatz, H., Schroeder, A., Otten, C., Büchler, R., Berg, S., Ritter, W., Mühlen, W., Gisder, S., Meixner, M., Liebig, G., and Rosenkranz, P. (2010). The German bee monitoring project: A long term study to understand periodically high winter losses of honey bee colonies. *Apidologie*, 41(3):332–352.

Gervini, D. (2005). Robust adaptive estimators for binary regression models. *Journal of Statistical Planning and Inference*, 131(2):297–311.

Gervini, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600.

Ghaoui, L. E., Viallon, V., and Rabbani, T. (2010). Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*.

Gijbels, I. and Vrinssen, I. (2019). Robust estimation and variable selection in heteroscedastic linear regression. *Statistics*, 53(3):489–532.

Gómez, A. (2021). Outlier detection in time series via mixed-integer conic quadratic optimization. *SIAM Journal on Optimization*, 31(3):1897–1925.

Gómez, A. and Prokopyev, O. A. (2021). A mixed-integer fractional optimization approach to best subset selection. *INFORMS Journal on Computing*, 33(2):551–565.

Grandvalet, Y. (1998). Least absolute shrinkage is equivalent to quadratic penalization. In *International Conference on Artificial Neural Networks*, pages 201–206. Springer.

Gumedze, F. N. (2019). Use of likelihood ratio tests to detect outliers under the variance shift outlier model. *Journal of Applied Statistics*, 46(4):598–620.

Gumedze, F. N., Welham, S. J., Gogel, B. J., and Thompson, R. (2010). A variance shift model for detection of outliers in the linear mixed model. *Computational Statistics & Data Analysis*, 54(9):2128–2144.

Hadi, A. S. and Luceño, A. (1997). Maximum trimmed likelihood estimators: A unified approach, examples, and algorithms. *Computational Statistics & Data Analysis*, 25(3):251–272.

Haffajee, A. D. and Socransky, S. S. (2009). Relation of body mass index, periodontitis and Tannerella forsythia. *Journal of Clinical Periodontology*, 36(2):89–99.

Hampel, F. R. (1968). *Contributions to the Theory of Robust Estimation*. PhD thesis, University of California, Berkeley.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York.

Hampel, F. R., Rousseeuw, P. J., and Ronchetti, E. (1981). The change-of-variance curve and optimal redescending M-estimators. *Journal of the American Statistical Association*, 76(375):643–648.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York, USA, second edition.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2020). Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC.

Hazimeh, H. and Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5):1517–1537.

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17–22.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., et al. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies *ASPM* as a molecular target. *Proceedings of the National Academy of Sciences*, 103(46):17402–17407.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. (1996). *Robust Statistical Procedures*. SIAM, Philadelphia, second edition.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. John Wiley & Sons, New Jersey.

Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, 47(1):64–79.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503.

Imon, A. R. and Hadi, A. S. (2008). Identification of multiple outliers in logistic regression. *Communications in Statistics - Theory and Methods*, 37(11):1697–1709.

Insolia, L., Chiaromonte, F., Li, R., and Riani, M. (2021a). Doubly robust feature selection with mean and variance outlier detection and oracle properties. *arXiv preprint, arXiv:2106.11941*.

Insolia, L., Chiaromonte, F., and Riani, M. (2021b). A robust estimation approach for mean-shift and variance-inflation outliers. In Bura, E. and Li, B., editors, *Festschrift in Honor of R. Dennis Cook: Fifty Years of Contribution to Statistical Science*, pages 17–41. Springer.

Insolia, L., Kenney, A., Calovi, M., and Chiaromonte, F. (2021c). Robust variable selection with optimality guarantees for high-dimensional logistic regression. *Stats*, 4(3):665–681.

Insolia, L., Kenney, A., Chiaromonte, F., and Felici, G. (2021d). Simultaneous feature selection and outlier detection with optimality guarantees. *Biometrics*, Forthcoming:1–12.

Insolia, L., Molinari, R., Rogers, S. R., Williams, G. R., Chiaromonte, F., and Calovi, M. (2022). Honey bee colony loss linked to parasites, pesticides and extreme weather across the United States. *Under revision*.

Insolia, L. and Perrotta, D. (2023). Tk-merge: Computationally efficient robust clustering under general assumptions. In García-Escudero, L. A., Gordaliza, A., Mayo, A., Lubiano Gomez, M. A., Gil, M. A., Grzegorzewski, P., and Hryniewicz, O., editors, *Building Bridges between Soft and Statistical Methodologies for Data Science*, volume 1433 of *Advances in Intelligent Systems and Computing*, pages 216–223. Springer International Publishing.

Jammal, M., Canu, S., and Abdallah, M. (2021). Joint outlier detection and variable selection using discrete optimization. *SORT-Statistics and Operations Research Transactions*, 45(1):47–66.

Johansen, S., Nielsen, B., et al. (2016). Analysis of the Forward Search using some new results for martingales and empirical processes. *Bernoulli*, 22(2):1131–1183.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, volume 2. John Wiley & Sons.

Kalogridis, I. and Van Aelst, S. (2019). Robust functional regression based on principal components. *Journal of Multivariate Analysis*, 173:393–415.

Kenney, A. (2021). *Mixed Integer Programming, Whitening, and Fucntional Data Analysis: Improving Feature Selection in "OMICS" Research*. PhD thesis, The Pennsylvania State University.

Kenney, A., Chiaromonte, F., and Felici, G. (2021). MIP-BOOST: Efficient and effective L_0 feature selection for linear regression. *Journal of Computational and Graphical Statistics*, 30(3):566–577.

Kepplinger, D. (2021). Robust variable selection and estimation via adaptive elastic net S-estimators for linear regression. *arXiv preprint arXiv:2107.03325*.

Koh, K., Kim, S.-J., and Boyd, S. (2007). An interior-point method for large-scale ℓ_1-regularized logistic regression. *Journal of Machine Learning Research*, 8(Jul):1519–1555.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, USA.

Koller, M. (2013). *Robust Estimation of Linear Mixed Models*. PhD thesis, ETH Zurich.

Kong, D., Bondell, H. D., and Wu, Y. (2018). Fully efficient robust estimation, outlier detection and variable selection via penalized regression. *Statistica Sinica*, 28(2):1031–1052.

Kreber, D. (2019). A mixed-integer optimization approach to an exhaustive cross-validated model selection for regression. *Preprint, http://www. optimization-online. org/DB_HTML/2019/05/7188. html*.

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, 84(406):460–466.

Kurnaz, F. S., Hoffmann, I., and Filzmoser, P. (2017). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems*, 172:211–222.

Kurnaz, F. S., Hoffmann, I., and Filzmoser, P. (2018). enetLTS: Robust and sparse methods for high dimensional linear and logistic regression. https://CRAN.R-project.org/package=enetLTS.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, 79(385):61–71.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C*, 41(1):191–201.

Lee, Y., MacEachern, S. N., Jung, Y., et al. (2012). Regularization of case-specific parameters for robustness and efficiency. *Statistical Science*, 27(3):350–372.

Leung, A., Yohai, V., and Zamar, R. (2017). Multivariate location and scatter matrix estimation under cellwise and casewise contamination. *Computational Statistics & Data Analysis*, 111:59–76.

Leung, A., Zhang, H., and Zamar, R. (2016). Robust regression estimation and inference in the presence of cellwise and casewise contamination. *Computational Statistics & Data Analysis*, 99:1–11.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877.

Liang, H., Wu, H., and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, 95(3):773–778.

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326.

Liu, H., Wang, J., He, T., Becker, S., Zhang, G., Li, D., and Ma, X. (2018). Butyrate: A double-edged sword for health? *Advances in Nutrition*, 9(1):21–29.

Liu, H., Yao, T., and Li, R. (2016). Global solutions to folded concave penalized nonconvex learning. *The Annals of Statistics*, 44(2):629–659.

Liu, H. and Yu, B. (2013). Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics*, 7:3124–3169.

Liu, J., Cosman, P. C., and Rao, B. D. (2017). Robust Linear Regression via ℓ_0 Regularization. *IEEE Transactions on Signal Processing*, 66(3):698–713.

Liu, T.-Y. and Jiang, H. (2019). Minimizing sum of truncated convex functions and its applications. *Journal of Computational and Graphical Statistics*, 28(1):1–10.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45(2):866–896.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, 37(6A):3498–3528.

Ma, S., Fildes, R., and Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra-and inter-category promotional information. *European Journal of Operational Research*, 249(1):245–257.

Maronna, R. A. (2011). Robust ridge regression for high-dimensional data. *Technometrics*, 53(1):44–53.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons.

Maronna, R. A. and Yohai, V. J. (2013). Robust functional linear regression based on splines. *Computational Statistics & Data Analysis*, 65:46–55.

McCann, L. et al. (2006). *Robust Model Selection and Outlier Detection in Linear Regressions*. PhD thesis, Massachusetts Institute of Technology.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, second edition.

McIlhagga, W. (2016). Penalized: A MATLAB toolbox for fitting generalized linear models with penalties. *Journal of Statistical Software*, 72:1–21.

Medina, D., Li, H., Vilà-Valls, J., and Closas, P. (2019). Robust statistics for GNSS positioning under harsh conditions: A useful tool? *Sensors*, 19(24):5402.

Menjoge, R. S. and Welsch, R. E. (2010). A diagnostic method for simultaneous feature selection and outlier identification in linear regression. *Computational Statistics & Data Analysis*, 54(12):3181–3193.

Miller, A. J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society: Series A*, pages 389–425.

Miller, A. J. (2002). *Subset Selection in Regression*. Chapman and Hall/CRC, Boca Raton, Florida, second edition.

Morawetz, L., Köglberger, H., Griesbacher, A., Derakhshifar, I., Crailsheim, K., Brodschneider, R., and Moosbeckhofer, R. (2019). Health status of honey bee colonies (Apis mellifera) and disease-related risk factors for colony losses in Austria. *PLoS One*, 14(7):1–28.

Morgenthaler, S. (2007). A survey of robust statistics. *Statistical Methods and Applications*, 15(3):271–293.

Morgenthaler, S., Welsch, R. E., and Zenide, A. (2004). Algorithms for robust model selection in linear regression. In *Theory and Applications of Recent Robust Methods*, pages 195–206. Springer.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359.

Morse, R. A. and Calderone, N. W. (2000). The value of honey bees as pollinators of US crops in 2000. *Bee Culture*, 128(3):1–15.

Müller, C. H. and Neykov, N. (2003). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference*, 116(2):503–519.

Müller, S., Scealy, J. L., and Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, 28(2):135–167.

Müller, S. and Welsh, A. H. (2005). Outlier robust model selection in linear regression. *Journal of the American Statistical Association*, 100(472):1297–1310.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234.

Needell, D. and Tropp, J. A. (2009). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321.

Neykov, N., Filzmoser, P., and Neytchev, P. (2014). Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. *Statistical Papers*, 55(1):187–207.

Oldroyd, B. P. and Nanork, P. (2009). Conservation of Asian honey bees. *Apidologie*, 40(3):296–312.

Öllerer, V., Alfons, A., and Croux, C. (2016). The shooting S-estimator for robust regression. *Computational Statistics*, 31(3):829–844.

Pannu, J. and Billor, N. (2020). Robust sparse functional regression model. *Communications in Statistics - Simulation and Computation*, Forthcoming:1–21.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

Paul, I. M., Williams, J. S., Anzman-Frasca, S., Beiler, J. S., Makova, K. D., Marini, M. E., et al. (2014). The intervention nurses start infants growing on healthy trajectories (INSIGHT) study. *BMC Pediatrics*, 14(1):1–15.

Peng, H. and Lu, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis*, 109:109–129.

Perrotta, D., Cerasa, A., Torti, F., and Riani, M. (2020). The robust estimation of monthly prices of goods traded by the European Union. Technical report, Technical report EUR 30188 EN, JRC120407, Publications Office of the European Union, Luxembourg.

Perrotta, D. and Torti, F. (2010). Detecting price outliers in European trade data with the forward search. In *Data Analysis and Classification*, pages 415–423. Springer.

Pettis, J. S. and Delaplane, K. S. (2010). Coordinated responses to honey bee decline in the USA. *Apidologie*, 41(3):256–263.

Potts, S. G., Roberts, S. P., Dean, R., Marris, G., Brown, M. A., Jones, R., Neumann, P., and Settele, J. (2010). Declines of managed honey bees and beekeepers in Europe. *Journal of Apicultural Research*, 49(1):15–22.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705–724.

Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., Rivas, M. A., and Hastie, T. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genetics*, 16(10):1–30.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, second edition.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, 26(1):35–67.

Riani, M. and Atkinson, A. C. (2007). Fast calibrations of the forward search for testing multiple outliers in regression. *Advances in Data Analysis and Classification*, 1(2):123–141.

Riani, M., Atkinson, A. C., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society: Series B*, 71(2):447–466.

Riani, M., Atkinson, A. C., Corbellini, A., and Fabrizio, L. (2022). Information criteria for outlier detection avoiding arbitrary significance levels. *Econometrics and Statistics*, Forthcoming.

Riani, M., Cerioli, A., Atkinson, A. C., Perrotta, D., et al. (2014). Monitoring robust regression. *Electronic Journal of Statistics*, 8(1):646–677.

Riani, M., Perrotta, D., and Torti, F. (2012). FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, 116:17–32.

Ronchetti, E., Field, C., and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, 92(439):1017–1023.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880.

Rousseeuw, P. J. and Bossche, W. V. D. (2018). Detecting deviating data cells. *Technometrics*, 60(2):135–145.

Rousseeuw, P. J. and Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis*, 43(3):315–332.

Rousseeuw, P. J. and Hubert, M. (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2):e1236.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.

Rousseeuw, P. J., Perrotta, D., Riani, M., and Hubert, M. (2019). Robust monitoring of time series with application to fraud detection. *Econometrics and Statistics*, 9:108–121.

Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Rousseeuw, P. J. and Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, 12(1):29–45.

Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, 85(411):633–639.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In Franke, J., Härdle, W., and Martin, D., editors, *Robust and Nonlinear Time Series Analysis*, volume 26, pages 256–272. Springer, New York.

Saishu, H., Kudo, K., and Takano, Y. (2021). Sparse Poisson regression via mixed-integer optimization. *PLoS One*, 16(4):e0249916.

Santner, T. J. and Duffy, D. E. (1986). A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73(3):755–758.

Sato, T., Takano, Y., and Miyashiro, R. (2017). Piecewise-linear approximation for feature subset selection in a sequential logit model. *Journal of the Operations Research Society of Japan*, 60(1):1–14.

Sato, T., Takano, Y., Miyashiro, R., and Yoshise, A. (2016). Feature subset selection for logistic regression via mixed integer optimization. *Computational Optimization and Applications*, 64(3):865–880.

Savage, J. S., Birch, L. L., Marini, M., Anzman-Frasca, S., and Paul, I. M. (2016). Effect of the INSIGHT responsive parenting intervention on rapid infant weight gain and overweight status at age 1 year: A randomized clinical trial. *BMC Pediatrics*, 170(8):742–749.

Schelldorfer, J., Bühlmann, P., and van De Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214.

Schrijver, A. (1986). *Theory of Linear and Integer Programming*. John Wiley & Sons, New York.

Seal, H. L. (1967). Studies in the History of Probability and Statistics. XV The historical development of the Gauss linear model. *Biometrika*, 54(1-2):1–24.

Seeley, T. D. and Visscher, P. K. (1985). Survival of honeybees in cold climates: The critical timing of colony growth and reproduction. *Ecological Entomology*, 10(1):81–88.

Shafieezadeh Abadeh, S., Mohajerin Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 28:1576–1584.

She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639.

She, Y., Wang, Z., and Shen, J. (2021). Gaining outlier resistance with progressive quantiles: Fast algorithms and theoretical studies. *Journal of the American Statistical Association*, Forthcoming:1–14.

Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232.

Shen, X., Pan, W., Zhu, Y., and Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5):807–832.

Smucler, E. and Yohai, V. J. (2017). Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, 111:116–130.

SRA (2017). INSIGHT oral and gut microbiome. www.ncbi.nlm.nih.gov/bioproject/PRJNA420339.

Steinhauer, N., Aurell, D., Bruckner, S., Wilson, M., Rennich, K., vanEngelsdorp, D., and Williams, G. (2021). United States honey bee colony losses 2020-2021: Preliminary results. *Technical report*.

Steinhauer, N., Rennich, K., Wilson, M., Caron, D., Lengerich, E., Pettis, J., Rose, R., Skinner, J., Tarpy, D., Wilkes, J., and VanEngelsdorp, D. (2014). A national survey of managed honey bee 2012-2013 annual colony losses in the USA: Results from the Bee Informed Partnership. *Journal of Apicultural Research*, 53:1–18.

Stout, JC, M. C. (2009). Ecological impacts of invasive alien species on bees. *Apidologie*, 40:388–409.

Su, P., Tarr, G., and Muller, S. (2021). Robust variable selection under cellwise contamination. *arXiv preprint arXiv:2110.12406*.

Switanek, M., Crailsheim, K., Truhetz, H., and Brodschneider, R. (2017). Modelling seasonal effects of temperature and precipitation on honey bee winter mortality in a temperate climate. *Science of The Total Environment*, 579:1581–1587.

Takano, Y. and Miyashiro, R. (2020). Best subset selection via cross-validation criterion. *Top*, 28:475–488.

Tallis, G. M. (1963). Elliptical and radial truncation in normal populations. *The Annals of Mathematical Statistics*, 34(3):940–944.

Taveras, E. M., Rifas-Shiman, S. L., Belfort, M. B., Kleinman, K. P., Oken, E., and Gillman, M. W. (2009). Weight status in the first 6 months of life and obesity at 3 years of age. *Pediatrics*, 123(4):1177–1183.

Taylan, P., Yerlikaya-Özkurt, F., Bilgiç Uçak, B., and Weber, G.-W. (2021). A new outlier detection method based on convex optimization: Application to diagnosis of Parkinson's disease. *Journal of Applied Statistics*, 48(13-15):2421–2440.

Taylan, P., Yerlikaya-Özkurt, F., and Weber, G.-W. (2014). An approach to the mean shift outlier model by Tikhonov regularization and conic programming. *Intelligent Data Analysis*, 18(1):79–94.

Thompson, R. (1985). A note on restricted maximum likelihood estimation with an alternative outlier model. *Journal of the Royal Statistical Society: Series B*, 47(1):53–55.

Thompson, R. (2022). Robust subset selection. *Computational Statistics & Data Analysis*, 169:107415.

Tibshirani, J. and Manning, C. D. (2013). Robust logistic regression using shift parameters. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2: Short Papers, Baltimore, Maryland. Association for Computational Linguistics.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.

Todorov, V. and Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47.

Torti, F., Perrotta, D., Riani, M., and Cerioli, A. (2019). Assessing trimming methodologies for clustering linear regression data. *Advances in Data Analysis and Classification*, 13(1):227–257.

Tukey, J. (1960). A survey of sampling from contaminated distributions. In Olkin, I., editor, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 448–485. Standford University Press, Stanford, CA.

Van Aelst, S., Vandervieren, E., and Willems, G. (2012). A Stahel–Donoho estimator based on huberized outlyingness. *Computational Statistics & Data Analysis*, 56(3):531–542.

van Dooremalen, C., Gerritsen, L., Cornelissen, B., van der Steen, J. J. M., van Langevelde, F., and Blacquière, T. (2012). Winter survival of individual honey bees and honey bee colonies depends on level of Varroa destructor infestation. *PLoS One*, 7(4):1–8.

vanEngelsdorp, D. and Meixner, M. D. (2010). A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. *Journal of Invertebrate Pathology*, 103:S80–S95.

Vital, M., Howe, A. C., and Tiedje, J. M. (2014). Revealing the bacterial butyrate synthesis pathways by analyzing (meta) genomic data. *mBio*, 5(2):1–11.

Wang, D. and Loh, P.-L. (2020). Robust estimation in high-dimensional sparse heteroscedastic linear models. Technical report, University of Wisconsin-Madison.

Wang, H. (2012). Factor profiled sure independence screening. *Biometrika*, 99(1):15–28.

Wang, J., Wonka, P., and Ye, J. (2015). Lasso screening rules via dual polytope projection. *Journal of Machine Learning Research*, 16:1063–1101.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295.

Wang, S., Nan, B., Rosset, S., and Zhu, J. (2011). Random lasso. *The Annals of Applied Statistics*, 5(1):468.

Wang, T., Li, Q., Chen, B., and Li, Z. (2018). Multiple outliers detection in sparse high-dimensional regression. *Journal of Statistical Computation and Simulation*, 88(1):89–107.

Wang, Y. (2019). *Robust and Sparse Statistical Models for High-Dimensional Data*. PhD thesis, KU Leuven.

Wang, Y. and Van Aelst, S. (2019). Robust variable screening for regression using factor profiling. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(2):70–87.

Willis, M. J. and von Stosch, M. (2017). L0-constrained regression using mixed integer linear programming. *Chemometrics and Intelligent Laboratory Systems*, 165:29–37.

Xiang, S., Shen, X., and Ye, J. (2015). Efficient nonconvex sparse group feature selection via continuous and discrete optimization. *Artificial Intelligence*, 224:28–50.

Yasrebi-de Kom, I. A. R., Biesmeijer, J. C., and Aguirre-Gutiérrez, J. (2019). Risk of potential pesticide use to honeybee and bumblebee survival and distribution: A country-wide analysis for The Netherlands. *Diversity and Distributions*, 25(11):1709–1720.

Yerlikaya-Özkurt, F. and Taylan, P. (2020). New computational methods for classification problems in the existence of outliers based on conic quadratic optimization. *Communications in Statistics - Simulation and Computation*, 49(3):753–770.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2):642–656.

Yohai, V. J. and Zamar, R. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, 83(402):406–413.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

Zeigler, C. C., Persson, G. R., Wondimu, B., Marcus, C., Sobko, T., and Modéer, T. (2012). Microbiota in the oral subgingival biofilm is associated with obesity in adolescence. *Obesity*, 20(1):157–164.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

Zhu, Y., Shen, X., and Pan, W. (2020). On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association*, 115(529):217–230.

Zioutas, G. and Avramidis, A. (2005). Deleting outliers in robust regression with mixed integer programming. *Acta Mathematicae Applicatae Sinica*, 21(2):323–334.

Zioutas, G., Pitsoulis, L., and Avramidis, A. (2009). Quadratic mixed integer programming and support vectors for deleting outliers in robust regression. *Annals of Operations Research*, 166(1):339–353.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533.