Methodological advances in high- and infinite-dimensional statistics

Ryan Thompson
BCom (Hons), University of Sydney

A thesis submitted for the degree of
Doctor of Philosophy at Monash University in 2022

Monash Business School
Department of Econometrics and Business Statistics

Abstract

The 21st century has borne witness to significant technological advances, not least of which has been tremendous growth in computing power. At the same time, fields diverse as economics and psychology have seen an explosion in both the complexity of data and the questions asked of it. These changes have given rise to a new data analytic landscape and drawn into focus a host of intriguing statistical problems. Among the most relevant in this new landscape are those problems comprising a large or infinite number of unknowns, studied under the umbrella of high- and infinite-dimensional statistics. Motivated by numerous areas of application and enabled by remarkable advances in technology, we present a collection of works that seek to push the frontiers in this area.

First, we consider a fundamental tool for high-dimensional regression under sparsity—best subset selection. Notwithstanding its desirable statistical properties, best subset selection is susceptible to outliers and can break down from a single contaminated data point. To address this issue, we introduce robust subset selection, which generalises the concept of subset selection to predictors and observations, thereby achieving robustness and sparsity. Building on recent advances in combinatorial optimisation, we devise a flexible computational framework. The robustness of the new estimator in terms of its finite-sample breakdown point is established. Compared with existing proposals, robust subset selection yields lower false positive rates and improved prediction error.

Next, we look at high-dimensional regression and classification from the perspective of structured sparsity, a generalisation of sparsity that encompasses many statistical and machine learning problems. We propose structured sparse estimators that combine group subset selection with shrinkage. To accommodate sophisticated structures, the new estimators allow for overlap between groups. An optimisation framework for fitting the regularisation surface is developed, and finite-sample error bounds for estimation of the regression function are presented. As an application requiring structure, we study sparse semiparametric modelling, a procedure that allows the effect of each predictor to be zero, linear, or nonlinear. Our estimators improve on alternatives for this task.

Finally, we switch gears and revisit a classic statistical problem—testing statistical hypotheses of centre. Classical tests that assess statistical hypotheses of centre implicitly assume a specific centre, e.g., the mean or median. Yet, scientific hypotheses, from which statistical hypotheses are derived, often do not prescribe a specific centre. Rather than test a single centre, we propose to test a family of plausible centres, such as that induced by the Huber loss function. This new problem amounts to testing an infinite-dimensional parameter whose components are different centres. We devise a Bayesian nonparametric testing procedure for this task, enabled by a novel pathwise optimisation routine. The new test has several favourable properties not found in classical tests.

Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Ryan Thompson
16 September 2022

Acknowledgements

First and foremost, I wish to express my appreciation to my PhD advisors: Farshid Vahid and Catherine Forbes, whose intellectual and professional wisdom I benefited immensely from over the last three and a half years. I am particularly thankful for the freedom they afforded me in choosing the directions of my research and for their steadfast encouragement in all my academic endeavours. It truly has been a gratifying and pleasurable experience to produce this thesis under their collective mentorship.

Second, I thank the current and former members of my PhD panel: Gael Martin, Natalia Bailey, Anastasios Panagiotelis, and Didier Nibbering, who kindly committed their time overseeing my candidature to completion. I also gratefully acknowledge the contributions of my collaborators, Steven MacEachern and Mario Peruggia, with whom I coauthored a paper that formed a chapter of this thesis. I have learned much from them and am continually impressed by the depth and breadth of their statistical intuition.

Last but not least, I must mention my family, to whom I owe a debt of gratitude. Thank you to my partner Ariana Tammetta, who boldly left her life in Sydney to come and study in Melbourne with me. In the face of much uncertainty over the last few years, her unwavering companionship and support have been one of the few precious constants in life. Thank you also to my parents, grandparents, and brother. I could not have reached this stage in life if not for their unconditional love, kindness, and generosity.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The 21st century has borne witness to significant technological advances, not least of which has been tremendous growth in computing power. In the last ten years, for instance, the high-performance computing systems benchmarked by the TOP500 supercomputing project saw an aggregate forty-fold performance increase (Strohmaier et al. 2022). At the same time, fields diverse as economics and psychology have seen an explosion in both the complexity of data and the questions asked of it. It is no longer uncommon, for example, to encounter data sets where the number of variables exceeds the number of observations by an order of magnitude or more. These changes have given rise to a new data analytic landscape and drawn into focus a host of intriguing statistical problems. Among the most relevant in this new landscape are those problems comprising a large or infinite number of unknowns, studied under the umbrella of high- and infinite-dimensional statistics. Though this topic has a long and rich history in statistics, going back well into the 20th century, today's landscape has rendered it more important than ever.

Motivated by numerous areas of application and enabled by remarkable advances in technology, we present in this thesis a collection of works that seek to push the frontiers in high- and infinite-dimensional statistics. Our core contributions constitute new methodological tools for problems relating to sparsity, robustness, and nonparametrics. Throughout the thesis, we give special consideration to the issue of computation, a central concern when working in high- or infinite-dimensions. To help bridge theory and practice, open-source software implementations of all the new tools are made publicly available.

1.1 Robust subset selection

In Chapter 2, we consider a fundamental tool in high-dimensional regression under sparsity—best subset selection (or 'best subsets'). Thanks to recent developments in mathematical optimisation, the problem of choosing k from p predictors that defines best subsets is more computationally tractable than ever. Notwithstanding its desirable statistical properties, the best subsets estimator is susceptible to outliers and can break down from a single contaminated data point. To address this issue, we propose a robust adaption of best subsets. The adapted estimator generalises the notion of subset selection to predictors and observations, thereby achieving robustness in addition to sparsity. This procedure, which we call 'robust subset selection' (or 'robust subsets'), is defined by a combinatorial problem for which we apply modern discrete optimisation methods. We formally establish the robustness of the new estimator in terms of the finite-sample breakdown point of its objective value. Experiments on synthetic and real

data demonstrate the superiority of robust subsets over best subsets in the presence of contamination. Importantly, robust subsets offers lower false positive rates and improved prediction error compared with robust adaptions of continuous shrinkage estimators.

1.2 Group selection and shrinkage: Structured sparsity for semiparametric models

In Chapter 3, we study high-dimensional regression and classification from the perspective of structured sparsity. Sparse estimators that respect group structures have application to an assortment of statistical and machine learning problems, from multitask learning to sparse additive modelling to hierarchical selection. We introduce structured sparse estimators that combine group subset selection with shrinkage. To accommodate sophisticated structures, our estimators allow for arbitrary overlap between groups. We develop an optimisation framework for fitting the nonconvex regularisation surface and present finite-sample error bounds for estimation of the regression function. As an application requiring structure, we study sparse semiparametric modelling, a procedure that allows the effect of each predictor to be zero, linear, or nonlinear. For this task, the new estimators improve across several metrics on synthetic data compared to alternatives. Finally, we demonstrate their efficacy in modelling supermarket foot traffic and economic recessions using many predictors. These demonstrations suggest sparse semiparametric models, fit using the new estimators, are an excellent compromise between fully linear and fully nonparametric alternatives.

1.3 Familial inference

In Chapter 4, we turn to the issue of testing statistical hypotheses. In scientific research, statistical hypotheses are translations of scientific hypotheses into statements about one or more distributions, often concerning their centre. Tests that assess statistical hypotheses of centre implicitly assume a specific centre, e.g., the mean or median. Yet, scientific hypotheses do not always specify a particular centre. This ambiguity leaves the possibility for a gap between scientific theory and statistical practice that can lead to rejection of a true null. In the face of replicability crises in many scientific disciplines, 'significant results' of this kind are concerning. Rather than testing a single centre, we propose testing a family of plausible centres, such as that induced by the Huber loss function. This new problem amounts to testing an infinite-dimensional parameter whose components represent different centres. We devise a Bayesian nonparametric testing procedure for this task, enabled by a novel pathwise optimisation routine that shares relation to pathwise routines popularly used for sparse regression. The favourable properties of the new test are verified through numerical simulation in one- and two-sample settings. Two experiments from psychology serve as real-world case studies.

1.4 Origin

The chapters of this thesis are drawn from three research papers written during the author's PhD candidature. Chapter 2 originates from Thompson (2022), a journal article in *Computational Statistics and Data Analysis* (arXiv: 2005.08217). Chapter 3 originates from Thompson and Vahid (2022), a preprint coauthored with Farshid Vahid of Monash

University (arXiv: 2105.12081). Chapter 4 originates from Thompson et al. (2022), a preprint coauthored with Catherine Forbes of Monash University and Steven MacEachern and Mario Peruggia of Ohio State University (arXiv: 2202.12540).

Chapter 2

Robust subset selection

2.1 Introduction

We study the canonical linear regression model $Y = X\beta + \varepsilon$ with response $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, predictors $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times p}$, regression coefficients $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$, and noise $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$. It is assumed that the response is centred and that the predictors are standardised. In the low-dimensional regime, where the number of predictors p is smaller than the number of observations n, it is straightforward to estimate β using the least squares estimator. However, in numerous contemporary statistical applications, p can be (much) greater than n, in which case the least squares estimator is no longer statistically meaningful. One way to navigate such situations is to assume that the underlying model is sparse, i.e., to assume only a small fraction of the available predictors are important for explaining the response. Even when $p < n$, estimators that induce sparsity are useful because they err on the side of simplicity and interpretability. The best subset selection (or 'best subsets') estimator is one of the earliest estimators that operates in the spirit of this idea, solving the constrained least squares problem:

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \\ \text{s. t.} \quad & \|\beta\|_0 \leq k, \end{aligned} \tag{2.1}$$

where k is an integer such that $0 \leq k \leq \min(n-1, p)$, and the ℓ_0-norm $\|\beta\|_0 := \sum_{j=1}^{p} 1(\beta_j \neq 0)$ is the number of nonzero elements in β. Best subset selection is a combinatorial problem due to the sparsity constraint on β. Unlike other well-known sparsity-inducing estimators such as lasso (Tibshirani 1996), the best subsets estimator (via its sparsity constraint) directly controls the number of predictors in the model.

The optimisation problem (2.1) suggests that to solve for the best subsets estimator, one must conduct a combinatorial search for the subset of (at most) k predictors that yields the best linear representation of the response. Although the resulting estimator has favourable statistical properties in terms of estimation, prediction, and selection (Bunea et al. 2007; Raskutti et al. 2011; Shen et al. 2013; Zhang et al. 2014), actually solving the (nonconvex) combinatorial problem is no small feat. In fact, finding the best subset(s) is an NP-hard problem (Natarajan 1995), and popular implementations such as the `R` package `leaps` do not scale well beyond $p \approx 30$. However, in recent work, Bertsimas et al. (2016) showed that the best subsets problem (2.1) can be formulated and solved (to global optimality) as a *mixed-integer program*, a class of mathematical optimisation

problems that has undergone remarkable advancements. In the last 10 years, for instance, the commercial mixed-integer solver `Gurobi` has experienced a nearly 60-fold hardware-independent speedup (Gurobi Optimization 2020). When used in conjunction with warm starts from a projected gradient descent method, Bertsimas et al. (2016) showed that their mixed-integer optimisation approach for best subsets can be applied to problems with dimensions as large as $p \approx 1,000$. This development represents the first time that the best subsets estimator has been tractable for contemporary high-dimensional data after at least 50 years of literature and has paved the way for exciting new research (Bertsimas and King 2016; Mazumder and Radchenko 2017; Kreber 2019; Bertsimas et al. 2020; Bertsimas and Van Parys 2020; Hastie et al. 2020; Hazimeh and Mazumder 2020; Takano and Miyashiro 2020; Kenney et al. 2021; Mazumder et al. 2023).

Despite the impressive developments in computational tools for best subset selection, certain fundamental limitations in the estimator itself remain. Particularly relevant to real-world applications is the robustness of best subsets to contamination in the data or lack thereof. Similar to the nonsparse least squares estimator, best subsets is highly susceptible to contamination in both the response and the predictors. Specifically, in the *casewise contamination framework*, where a portion of the rows of Y and X are outliers, a single contaminated data point can have an arbitrarily severe effect on best subsets. The absence of robustness to casewise contamination is an important practical limitation of best subsets and raises doubts about the appropriateness of the estimator for many applications of sparse regression involving outliers, e.g., earnings forecasting (Wang et al. 2007), analytical chemistry (Smucler and Yohai 2017), and biomarker discovery (Cohen Freue et al. 2019). Although robust adaptions of other sparse estimators such as lasso have been studied fairly intensively (Rosset and Zhu 2007; Wang et al. 2007; Lambert-Lacroix and Zwald 2011; Alfons et al. 2013; Nguyen and Tran 2013; Wang et al. 2013; Lozano et al. 2016; Smucler and Yohai 2017; Chang et al. 2018; Yang et al. 2018; Amato et al. 2021), a lack of similar research is available on the topic of best subsets due to computational considerations. The objective of this chapter is to address this gap.

2.1.1 Robust subset selection

In view of the preceding discussion, we study a robust adaption of best subset selection: robust subset selection (or 'robust subsets'). Motivated by ideas related to robust statistics and advances in mathematical optimisation, robust subsets generalises the problem of selection to include both predictors *and* observations, leading to the combinatorial problem:

$$\begin{aligned} \min_{\beta, I} \quad & \frac{1}{2} \sum_{i \in I} (y_i - x_i^T \beta)^2 \\ \text{s.t.} \quad & \|\beta\|_0 \leq k \\ & I \subseteq [n] \\ & |I| \geq h, \end{aligned} \tag{2.2}$$

where h is an integer such that $k \leq h \leq n$, and the notation $[n]$ denotes the set $\{1, \ldots, n\}$. In effect, robust subsets performs a best subsets fit on the h observations that produce the smallest square error while the most anomalous $n - h$ observations are 'trimmed'. The idea of trimming anomalous observations is inspired by the method of *least trimmed squares* (LTS), an estimator that is highly resistant to contamination in both Y and X and is well-established in the robust statistics literature (Rousseeuw 1984). Because

minimising the sum of squares in (2.2) without the sparsity constraint on β leads to the LTS estimator, robust subsets can be interpreted as subset selection under LTS loss.

Although solving (2.2) exactly is theoretically intractable (it is NP-hard), this chapter demonstrates that modern methods from mathematical optimisation can be applied to tackle practical-sized problem instances with n and p in the hundreds, including the high-dimensional case when $p \gg n$. The resulting estimator is shown to have favourable statistical properties, both theoretically in terms of its finite-sample breakdown point and empirically in terms of its performance on synthetic and real data. Unlike robust adaptions of sparse estimators that rely on continuous shrinkage, the robust subsets estimator (via its nonconvex sparsity constraint on β) exhibits excellent support recovery and produces fitted models with few nonzeros.

2.1.2 Contributions and organisation

The contributions of this work are summarised as follows. We first show that the problem of robust subset selection is amenable to formulation as a mixed-integer program, allowing us to leverage advancements in mixed-integer solvers to compute exact solutions. To complement this approach, we develop tailored heuristics to quickly obtain good feasible solutions to the robust subsets problem. These heuristics include a projected block-coordinate gradient descent method, for which we derive convergence properties, and a neighbourhood search method, which exploits neighbourhood information across a grid of the parameters k and h to generate an entire set of fitted models. Our heuristics can also rapidly generate warm start solutions that the solver can refine to produce high-quality fitted models. The breakdown point for the objective value of the robust subsets estimator is subsequently derived. Finally, numerical experiments are conducted on synthetic data under a comprehensive set of contamination settings. A real data application is also illustrated. Our implementation `robustsubsets` is made available as an open-source `R` package.

This chapter is structured as follows. Section 2.2 provides a brief review of related work. Section 2.3 introduces computational methods for robust subset selection. Section 2.4 discusses robustness properties vis-à-vis the breakdown point. Section 2.5 presents results from numerical experiments. Section 2.6 describes a real data application. Section 2.7 closes the chapter with concluding remarks. Proofs are provided in the appendices.

2.2 Background

In light of the extensive literature on the topics of sparse and robust regression (and their intersection), we provide a brief review of select work related to the present chapter.

2.2.1 Best subset selection and lasso

Since at least the 1960s, best subset selection has been recognised as an important problem in statistics (Garside 1965; Beale et al. 1967; Hocking and Leslie 1967). Furnival and Wilson (1974), in a seminal paper, introduced an exact algorithm for best subsets that relies on a branch-and-bound method and is still used today in `leaps`.[1] In more recent decades, computationally friendlier estimators have arisen, most notably lasso.

[1] See also Gatu and Kontoghiorghes (2006) and Hofmann et al. (2007) for later, related work.

Unlike best subsets, lasso is defined by a relatively simple convex problem:

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 \\ \text{s.t.} \quad & \|\beta\|_1 \leq t, \end{aligned} \tag{2.3}$$

where $t > 0$ controls the level of sparsity (albeit, indirectly). Any modern convex solver can optimise (2.3) or its popular Lagrangian form:

$$\min_{\beta} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_1, \tag{2.4}$$

which is equivalent to (2.3) for some $\lambda > 0$. In addition to convex solvers, efficient algorithms also exist for computing lasso that exploit its highly structured nature, including least angle regression (LARS, Efron et al. 2004) and pathwise coordinate descent (Friedman et al. 2007). Other noteworthy estimators include the Dantzig selector (Candes and Tao 2007), as well as those based on nonconvex penalties such as the smoothly clipped absolute deviation penalty (Fan and Li 2001) and the minimax concave penalty (Zhang 2010). For relevance, we limit our discussion to best subsets and lasso due to the prevalence of the latter in the robust statistics literature.

The lasso problems (2.3) and (2.4) are a convex relaxation of the best subsets problem (2.1); they replace the ℓ_0 constraint with a convex surrogate ℓ_1 constraint (or penalty). Therefore, lasso is often interpreted as a heuristic for best subsets. However, unlike the best subsets parameter k, the lasso parameters t or λ do not directly control the model sparsity. Lasso also induces shrinkage on the regression coefficients, which can help or hinder depending on the level of noise (Hastie et al. 2020; Mazumder et al. 2023). On the other hand, best subsets allows predictors to enter the model with a full least squares fit, removing the effect of other correlated predictors in the process.

From a theory standpoint, lasso requires that somewhat restrictive conditions hold to achieve good statistical properties (see, e.g., van de Geer and Bühlmann 2009). Zhao and Yu (2006) showed that lasso is only capable of selecting the true model consistently under the so-called irrepresentable condition, which places rather strong restrictions on the covariance of the predictors. Zhang et al. (2014) derived bounds on prediction loss from (a thresholded) lasso under a restricted eigenvalue condition on the predictor matrix. Even when this condition is satisfied, they showed that a substantial gap can still occur compared with the prediction loss from best subsets.

The empirical performance of best subsets compared with that of lasso is somewhat less well understood than the theory. Bertsimas et al. (2016) and Hastie et al. (2020) performed empirical comparisons of the estimators. Perhaps unsurprisingly, neither estimator was found to dominate uniformly, but their experiments validated the stylistic fact that best subsets tends to produce fitted models that are significantly sparser than those from lasso, especially when the number of predictors is large. The simulations in this chapter yield a similar finding in the contaminated setting in which the sparser models produced by robust subsets contain substantially fewer false positives than models generated by robust adaptions of lasso.

2.2.2 Sparse and robust regression

Like sparse regression, robust regression is a classical topic in statistics. A recent and detailed treatment of the subject is available in Maronna et al. (2019). With the

proliferation of high-dimensional datasets, estimators that are simultaneously robust and sparse have become a topic of intense interest in recent years. In particular, a fairly extensive body of literature is available in robust statistics on the topic of lasso. One of the earliest papers in this area is Wang et al. (2007), which introduced lasso with least absolute deviation (LAD) loss:

$$\min_{\beta} \quad \sum_{i=1}^{n} |y_i - x_i^T \beta| + \lambda \|\beta\|_1. \tag{2.5}$$

Huber loss was used by Rosset and Zhu (2007) and Lambert-Lacroix and Zwald (2011), and their estimators are related to the so-called extended lasso (Nguyen and Tran 2013). Indeed, lasso has been studied under numerous other loss functions including exponential square loss (Wang et al. 2013), minimum distance loss (Lozano et al. 2016), and Tukey's bisquare loss (Smucler and Yohai 2017; Chang et al. 2018). Alfons et al. (2013) studied lasso with LTS loss, which effectively relaxes the nonconvex sparsity constraint on β in the robust subsets problem (2.2) with an ℓ_1 penalty. In a slightly different line of work, Khan et al. (2007) robustified the LARS procedure by utilising resistant estimators of mean and covariance. Chen et al. (2013) proposed a related idea whereby the lasso objective is restated in terms of trimmed inner products.

The body of literature studying best subsets in the contaminated setting is relatively limited. Bertsimas et al. (2016) showed that their optimisation framework can incorporate subset selection with LAD loss, the ℓ_0 constrained analogy of (2.5). However, unlike the robust subsets problem, the LAD subsets problem does not result in an estimator resistant to contamination in the predictor matrix, which is arguably the most relevant case for contemporary applications involving large numbers of predictors. Heuristic algorithms for trimmed ℓ_0 constrained regression were developed and analysed in Bhatia et al. (2015) and Suggala et al. (2019) under particular conditions on the predictor matrix. Unfortunately, the algorithmic frameworks developed in those papers are relevant only for contamination in the response and do not apply to the problems we consider wherein X may also be contaminated. Liu et al. (2020) studied related heuristics based on the interesting idea of using a robust estimate of the gradient. A drawback of this approach is that each algorithmic iteration involves solving an optimisation problem afresh. After this work was first shared online, Insolia et al. (2021) established some theoretical guarantees in the form of oracle properties for robust subset selection.

Robust regression is an inherently nonconvex problem; see, e.g., the discussion in She and Owen (2011). Alfons et al. (2013) showed that the use of a convex loss function with lasso, such as in the LAD lasso problem, leads to a finite-sample breakdown point of $1/n$, the same as that from standard least squares loss. Accordingly, many sparse and robust estimators are defined in terms of nonconvex optimisation problems, nearly all of which rely solely on heuristics that are only capable of delivering approximate solutions (Alfons et al. 2013; Smucler and Yohai 2017; Chang et al. 2018). Although we also apply heuristics to a highly nonconvex problem, they form part of a broader framework that incorporates mixed-integer optimisation which is guaranteed to converge to a global minimiser, if one exists. In the nonsparse setting, mixed-integer optimisation has been used successfully in earlier works to find global minimisers for the problems of least trimmed squares (Zioutas et al. 2009) and its cousin least median of squares (Bertsimas and Mazumder 2014). See also Hofmann et al. (2010) for a tailored branch-and-bound method for least trimmed squares that comes with global convergence guarantees.

2.3 Computational methods

This section details computational methods for robust subset selection. We begin with a brief primer on mixed-integer optimisation and proceed to describe a mixed-integer program for robust subsets. Heuristics are presented, including a projected block-coordinate gradient descent method and a neighbourhood search method. The heuristics, together with mixed-integer optimisation, form a powerful computational toolkit for robust subsets. The section closes with practical guidance on parameter choices.

2.3.1 Mixed-integer optimisation

Primer

Recall that the general form for a mixed-integer program with a quadratic objective, linear constraints, and variable $\omega \in \mathbb{R}^p$ is

$$\begin{aligned} \min_{\omega} \quad & \omega^T Q \omega + q^T \omega \\ \text{s.t.} \quad & A\omega \leq b \\ & l_j \leq \omega_j \leq u_j, \quad j \in [p] \\ & \omega_j \in \mathbb{Z}, \qquad \text{for some } j \in [p], \end{aligned} \tag{2.6}$$

where positive semidefinite $Q \in \mathbb{R}^{p \times p}$ and $q \in \mathbb{R}^p$ form the objective, $A \in \mathbb{R}^{m \times p}$ is a constraint matrix and $b \in \mathbb{R}^m$ is a right-hand side vector, and $l \in \mathbb{R}^p$ and $u \in \mathbb{R}^p$ are lower and upper bound vectors. The program (2.6) is said to be a mixed-integer quadratic program, and the constraints $\omega_j \in \mathbb{Z}$ are said to be integrality constraints. It is these integrality constraints that render the feasible set nonconvex and lead to problems of the form (2.6) being NP-hard to solve in general (Aardal et al. 2005). State-of-the-art mixed-integer solvers such as `CPLEX`, `GLPK`, `Gurobi`, and `MOSEK` optimise (2.6) by applying branch-and-bound methods in combination with cutting plane generation techniques and elaborate heuristics. Roughly speaking, branch-and-bound methods operate by reducing the original problem to a series of subproblems represented as a search tree. Branches of the search tree are enumerated only if they can improve on the incumbent solution, as determined by the estimated lower and upper bounds on the optimal value of the objective function. Optimality of the solution is declared once the lower and upper bounds converge.

Mixed-integer programs

We turn our attention to formulating robust subset selection as a mixed-integer program. Towards this end, we begin with a formulation for best subset selection, itself a special case of robust subset selection. Letting $s \in \{0, 1\}^p$ be an auxiliary binary variable, the best subsets problem (2.1) has the following mixed-integer program representation:

$$\begin{aligned} \min_{\beta, s} \quad & \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \\ \text{s.t.} \quad & s_j \in \{0, 1\}, \qquad j \in [p] \\ & -s_j \mathcal{M}_\beta \leq \beta_j \leq s_j \mathcal{M}_\beta, \qquad j \in [p] \\ & \sum_{j=1}^{p} s_j \leq k, \end{aligned} \tag{2.7}$$

where $\mathcal{M}_\beta > 0$ is a problem-specific (fixed) parameter. The formulation (2.7) exploits the 'Big-M' constraints $-s_j\mathcal{M}_\beta \leq \beta_j \leq s_j\mathcal{M}_\beta$ to enforce sparsity on β. These Big-M constraints have the effect that

$$s_j = 0 \implies \beta_j = 0 \quad \text{and} \quad s_j = 1 \implies \beta_j \in [-\mathcal{M}_\beta, \mathcal{M}_\beta].$$

Hence, via $\mathcal{M}_\beta$, the s_j act as switches that control whether the β_j can take on nonzero values. The constraint $\sum_{j=1}^p s_j \leq k$ has the effect of upper bounding the ℓ_0-norm of β:

$$\sum_{j=1}^p s_j \leq k \implies \|\beta\|_0 \leq k,$$

thereby yielding the desired level of sparsity in the fitted model.

To generalise the program (2.7) to solve for the problem of interest, we exploit the following (equivalent) reformulation of the robust subsets problem (2.2):

$$\begin{aligned} \min_{\beta,\eta} \quad & \frac{1}{2}\sum_{i=1}^n (y_i - x_i^T\beta - \eta_i)^2 \\ \text{s.t.} \quad & \|\beta\|_0 \leq k \\ & \|\eta\|_0 \leq n - h, \end{aligned} \tag{2.8}$$

where we optimise over the continuous variable $\eta \in \mathbb{R}^n$ in place of the set-valued variable I. The indices of the nonzero elements in η correspond to the complement of the set of observation indices I. The trick of introducing auxiliary variables to achieve robustness has been used in several works previously (McCann and Welsch 2007; Menjoge and Welsch 2010; She and Owen 2011; Nguyen and Tran 2013; Suggala et al. 2019). In particular, it is known that constraining the ℓ_1-norm of η is equivalent to using Huber loss (She and Owen 2011), whereas constraining its ℓ_0-norm is equivalent to using LTS loss (Suggala et al. 2019). Though this trick has been used before, the algorithmic framework we develop and apply it in is itself novel.

To represent (2.8) as a mixed-integer program, we introduce the auxiliary binary variable $z \in \{0,1\}^n$ to construct the following formulation:

$$\begin{aligned} \min_{\beta,\eta,s,z} \quad & \frac{1}{2}\sum_{i=1}^n (y_i - x_i^T\beta - \eta_i)^2 \\ \text{s.t.} \quad & s_j \in \{0,1\}, \qquad j \in [p] \\ & -s_j\mathcal{M}_\beta \leq \beta_j \leq s_j\mathcal{M}_\beta, \qquad j \in [p] \\ & \sum_{j=1}^p s_j \leq k \\ & z_i \in \{0,1\}, \qquad i \in [n] \\ & -z_i\mathcal{M}_\eta \leq \eta_i \leq z_i\mathcal{M}_\eta, \qquad i \in [n] \\ & \sum_{i=1}^n z_i \leq n - h, \end{aligned} \tag{2.9}$$

where $\mathcal{M}_\eta > 0$ is a Big-M parameter for η. The robust subsets program (2.9) enforces sparsity on both β and η via the Big-M constraints.

For (2.9) to be a valid formulation of the robust subsets problem, insofar as its optimal solution is the same as that of (2.8), the Big-M parameters must be sufficiently large. More precisely, $\mathcal{M}_\beta$ and $\mathcal{M}_\eta$ should satisfy $\mathcal{M}_\beta \geq \|\beta^\star\|_\infty$ and $\mathcal{M}_\eta \geq \|\eta^\star\|_\infty$ for $\beta^\star$ and $\eta^\star$ that is an optimal solution to (2.8). Alternatively, the specification of either of these parameters can be avoided by replacing the Big-M constraints with indicator constraints or special ordered set (SOS) constraints. Such constraints do not require specification of any parameters but have the same effect as Big-M constraints. An SOS constraint (of type 1) has the effect that

$$(\beta_j, 1 - s_j) : \text{SOS-1} \implies \beta_j(1 - s_j) = 0.$$

Thus, replacing the Big-M constraints in (2.9) with SOS constraints does not change the optimal solution. However, it is our experience that the performance of the solver is generally superior when the problem is formulated with Big-M constraints. At the conclusion of this section, we show how the heuristics presented next can be used to estimate $\mathcal{M}_\beta$ and $\mathcal{M}_\eta$.

We make two remarks pertaining to the computation of a solution to (2.9):

- For a given problem instance, it might be the case that β is presumed to be fully dense (i.e., $k = p$), and it is thus desirable to remove the variable s and the corresponding Big-M constraints from the formulation. Likewise, if it is presumed that the data are uncontaminated (i.e., $h = n$), it is helpful to remove z from the problem, as well as η. The presolve routines used in most modern solvers are capable of identifying these situations and simplifying the formulation.

- The high-level solution strategy used by the solver can usually be tuned. Often, the competing goals of (a) finding a new feasible solution and (b) proving optimality of the incumbent solution are balanced. Still, work on either of these goals can also be prioritised. This capability is particularly useful for day-to-day data-analytic work in which obtaining high-quality solutions with low runtime is principally of interest.

Finally, several techniques are available that can improve the performance of the solver via simple modifications to the mixed-integer program (2.9). A brief discussion of these techniques is included in Appendix A.1.1.

2.3.2 Heuristics

Although modern mixed-integer solvers are capable of solving the mixed-integer programs that we have presented, they are not (in general) sufficiently quick to be of use by themselves for practical-sized problem instances (e.g., n and p in the hundreds). To this end, we propose tailored heuristic methods that complement the mixed-integer optimisation approach in the following ways:

- They can rapidly generate good feasible solutions to the robust subsets problem (2.8) that can be exploited by the solver as warm starts.

- Their solutions can be used to derive suitable values of the Big-M parameters (e.g., $\mathcal{M}_\beta$) required in the mixed-integer program.

- They can be applied to cross-validate the parameters k and h with low computational cost, which might otherwise require multiple expensive calls to the solver.

Projected block-coordinate gradient descent

In general, provably exact minimisers to the robust subsets problem (2.8) are unattainable without mixed-integer optimisation. However, first-order optimisation algorithms, namely, projected gradient descent methods, have been applied with great success to find good local minimisers for best subsets and related problems (Bertsimas et al. 2016; Kudo et al. 2020; Mazumder et al. 2023). Motivated by this success, we extend the projected gradient descent method developed in Bertsimas et al. (2016) for the best subsets problem (2.1) to the robust subsets problem (2.8). Their method involves a standard gradient descent update to the full set of coordinates followed by projection onto the feasible set of k-sparse solutions. Because the robust subsets problem is characterised by distinct blocks of coordinates, we adapt this scheme to perform block-coordinate updates. The resulting projected *block-coordinate* gradient descent method finds good feasible solutions that yield upper bounds to the optimal value of the robust subsets objective function.

For simplicity of exposition, $f(\beta, \eta)$ is used to denote the objective function in (2.8):

$$f(\beta, \eta) := \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta - \eta_i)^2 = \frac{1}{2}\|Y - X\beta - \eta\|_2^2. \tag{2.10}$$

The objective function (2.10) has the partial derivatives

$$\nabla_\beta f(\beta, \eta) = -X^T(Y - X\beta - \eta)$$

and

$$\nabla_\eta f(\beta, \eta) = -(Y - X\beta - \eta).$$

Observe that $\nabla_\beta f(\beta, \eta)$ and $\nabla_\eta f(\beta, \eta)$ are Lipschitz continuous, i.e., there exist real constants $L_\beta > 0$ and $L_\eta > 0$ such that

$$\|\nabla_\beta f(\beta, \eta) - \nabla_\beta f(\tilde{\beta}, \eta)\|_2 \leq L_\beta\|\beta - \tilde{\beta}\|_2 \quad \forall\, \beta, \tilde{\beta} \in \mathbb{R}^p,\ \eta \in \mathbb{R}^n \tag{2.11}$$

and

$$\|\nabla_\eta f(\beta, \eta) - \nabla_\eta f(\beta, \tilde{\eta})\|_2 \leq L_\eta\|\eta - \tilde{\eta}\|_2 \quad \forall\, \eta, \tilde{\eta} \in \mathbb{R}^n,\ \beta \in \mathbb{R}^p. \tag{2.12}$$

In particular, the Lipschitz constants $L_\beta = \|X^TX\|_2$ and $L_\eta = 1$, where $\|\cdot\|_2$ denotes the spectral norm of the matrix. The Lipschitz continuity of $\nabla_\beta f(\beta, \eta)$ and $\nabla_\eta f(\beta, \eta)$ leads to the block descent lemma (Beck 2015), whereby (2.10) can be upper bounded as follows.

Lemma 1. *Let $f(\beta, \eta)$ be the robust subset selection objective function (2.10). Then, for any $\bar{L}_\beta \geq L_\beta$ and any $\bar{L}_\eta \geq L_\eta$, it holds that*

$$f(\tilde{\beta}, \eta) \leq Q(\tilde{\beta}, \beta) := f(\beta, \eta) + \nabla_\beta f(\beta, \eta)^T(\tilde{\beta} - \beta) + \frac{1}{2}\bar{L}_\beta\|\tilde{\beta} - \beta\|_2^2 \quad \forall\, \beta, \tilde{\beta} \in \mathbb{R}^p,\ \eta \in \mathbb{R}^n$$

and

$$f(\beta, \tilde{\eta}) \leq R(\tilde{\eta}, \eta) := f(\beta, \eta) + \nabla_\eta f(\beta, \eta)^T(\tilde{\eta} - \eta) + \frac{1}{2}\bar{L}_\eta\|\tilde{\eta} - \eta\|_2^2 \quad \forall\, \eta, \tilde{\eta} \in \mathbb{R}^n,\ \beta \in \mathbb{R}^p.$$

The proposed projected block-coordinate gradient descent method performs cyclic updates by alternating between minimization of the upper bounds $Q(\tilde{\beta}, \beta)$ and $R(\tilde{\eta}, \eta)$. The hard-thresholding operator is pivotal to this minimization and is defined for a vector $c \in \mathbb{R}^p$ as

$$\mathrm{H}(c; k) \in \underset{\alpha \in \mathbb{R}^p : \|\alpha\|_0 \leq k}{\arg\min} \|\alpha - c\|_2^2.$$

Taking $\{(1), \ldots, (p)\}$ to denote an ordering of $\{1, \ldots, p\}$ such that $|c_{(1)}| \geq |c_{(2)}| \geq \cdots \geq |c_{(p)}|$, it is well-known that $\mathrm{H}(c; k)$ has the following analytic form:

$$\hat{\alpha}_j = \begin{cases} c_j & \text{if } j \in \{(1), \ldots, (k)\} \\ 0 & \text{otherwise} \end{cases}, \qquad j \in [p].$$

The operator $\mathrm{H}(c, k)$ retains the k largest elements of the vector c measured in absolute value and sets the remaining elements to zero. Observe that $\mathrm{H}(c, k)$ is a set-valued map because more than one valid permutation of the indices might exist. Using the hard-thresholding operator, the computation for a single update to β can be written as

$$\begin{aligned} \hat{\beta} &\in \underset{\tilde{\beta} \in \mathbb{R}^p : \|\tilde{\beta}\|_0 \leq k}{\arg\min} \; Q(\tilde{\beta}, \beta) \\ &= \underset{\tilde{\beta} \in \mathbb{R}^p : \|\tilde{\beta}\|_0 \leq k}{\arg\min} \left\| \tilde{\beta} - \left(\beta - \frac{1}{\bar{L}_\beta} \nabla_\beta f(\beta, \eta) \right) \right\|_2^2 \\ &= \mathrm{H}\left(\beta - \frac{1}{\bar{L}_\beta} \nabla_\beta f(\beta, \eta); k \right). \end{aligned}$$

Thus, with fixed η, an update to β is performed by taking a gradient descent step followed by a mapping to the nearest k-sparse subspace of $\mathbb{R}^p$. The second set of coordinates η can be updated similarly. In fact, with fixed β, such an update yields exact minimisation with respect to η. This result follows from the definition of the hard-thresholding operator (take $\alpha = \eta$ and $c = Y - X\beta$).

Using the above ingredients, Algorithm 1 presents the projected block-coordinate gradient descent method for optimisation of (2.8). Algorithm 1 first performs cyclic projected gradient descent updates until a convergence tolerance ϵ is satisfied. Upon convergence, the active set is fixed, and the coefficients are 'polished'. The polishing step can be performed by a simple least squares fit restricted to the predictors J and observations I. In the special case that $h = n$, the algorithm reduces to the projected gradient descent method in Bertsimas et al. (2016).

We now establish some convergence properties of Algorithm 1, extending results and proof techniques from Bertsimas et al. (2016, Proposition 6 and Theorem 3.1). To this end, we begin by stating the following definitions for points of (2.8) that are stationary and ϵ-optimal stationary.

Definition 1. *The point $(\hat{\beta}, \hat{\eta})$, with $\|\hat{\beta}\|_0 \leq k$ and $\|\hat{\eta}\|_0 \leq n - h$, is said to be a stationary point of the optimisation problem (2.8) if, for any $\bar{L}_\beta \geq L_\beta$ and any $\bar{L}_\eta \geq L_\eta$, it satisfies the fixed point equations*

$$\hat{\beta} \in \mathrm{H}\left(\hat{\beta} - \frac{1}{\bar{L}_\beta} \nabla_\beta f(\hat{\beta}, \hat{\eta}); k \right) \quad \text{and} \quad \hat{\eta} \in \mathrm{H}\left(\hat{\eta} - \frac{1}{\bar{L}_\eta} \nabla_\eta f(\hat{\beta}, \hat{\eta}); n - h \right).$$

Furthermore, $(\hat{\beta}, \hat{\eta})$ is said to be an ϵ-optimal stationary point if, for any $\epsilon > 0$, it satisfies the inequalities

$$\left\| \hat{\beta} - \mathrm{H}\left(\hat{\beta} - \frac{1}{\bar{L}_\beta} \nabla_\beta f(\hat{\beta}, \hat{\eta}); k \right) \right\|_2^2 \leq \epsilon \quad \text{and} \quad \left\| \hat{\eta} - \mathrm{H}\left(\hat{\eta} - \frac{1}{\bar{L}_\eta} \nabla_\eta f(\hat{\beta}, \hat{\eta}); n - h \right) \right\|_2^2 \leq \epsilon.$$

With these definitions in mind, the convergence properties of Algorithm 1 are given as follows.

Algorithm 1: Projected block-coordinate gradient descent

input : $\bar{L}_\beta \geq L_\beta$, $\bar{L}_\eta \geq L_\eta$, and $\epsilon > 0$.
initialise: $\beta^{(0)} \in \mathbb{R}^k \times \{0\}^{p-k}$ and $\eta^{(0)} \in \mathbb{R}^{n-h} \times \{0\}^h$.

1 For $m \geq 0$, repeat the following until $f(\beta^{(m)}, \eta^{(m)}) - f(\beta^{(m+1)}, \eta^{(m+1)}) \leq \epsilon$:
Update $\beta^{(m)}$ as

$$\beta^{(m+1)} \in \mathrm{H}\left(\beta^{(m)} - \frac{1}{\bar{L}_\beta}\nabla_\beta f(\beta^{(m)}, \eta^{(m)}); k\right).$$

Update $\eta^{(m)}$ as

$$\eta^{(m+1)} \in \mathrm{H}\left(\eta^{(m)} - \frac{1}{\bar{L}_\eta}\nabla_\eta f(\beta^{(m+1)}, \eta^{(m)}); n-h\right).$$

2 Fix the active sets

$$J = \left\{j \in [p] : \beta_j^{(m)} \neq 0\right\} \quad \text{and} \quad I = \left\{i \in [n] : \eta_i^{(m)} = 0\right\},$$

and solve the low-dimensional convex problem

$$\min_{\beta,\eta} \quad f(\beta, \eta) \quad \text{s. t.} \quad \beta_j = 0 \,\forall\, j \notin J,\, \eta_i = 0 \,\forall\, i \in I.$$

Proposition 1. *Let $\{(\beta^{(m)}, \eta^{(m)})\}$ be a sequence generated by Algorithm 1. Then, for any $\bar{L}_\beta \geq L_\beta$ and any $\bar{L}_\eta \geq L_\eta$, the sequence $\{f(\beta^{(m)}, \eta^{(m)})\}$ is decreasing, converges, and satisfies the inequality*

$$\begin{aligned} f(\beta^{(m)}, \eta^{(m)}) - f(\beta^{(m+1)}, \eta^{(m+1)}) \\ \geq \frac{1}{2}(\bar{L}_\beta - L_\beta)\|\beta^{(m+1)} - \beta^{(m)}\|_2^2 + \frac{1}{2}(\bar{L}_\eta - L_\eta)\|\eta^{(m+1)} - \eta^{(m)}\|_2^2. \end{aligned} \tag{2.13}$$

Furthermore, for any $\bar{L}_\beta > L_\beta$, any $\bar{L}_\eta > L_\eta$, and a stationary point $(\beta^\star, \eta^\star)$, the sequence $\{(\beta^{(m)}, \eta^{(m)})\}$ satisfies the following inequality after running Algorithm 1 for M iterations:

$$\min_{1 \leq m \leq M}\left(\|\beta^{(m+1)} - \beta^{(m)}\|_2^2 + \|\eta^{(m+1)} - \eta^{(m)}\|_2^2\right) \leq 2\frac{f(\beta^{(1)}, \eta^{(1)}) - f(\beta^\star, \eta^\star)}{M \min(\bar{L}_\beta - L_\beta, \bar{L}_\eta - L_\eta)}.$$

Proposition 1 establishes that Algorithm 1 generates a convergent sequence of objective values for the robust subsets problem. In particular, it follows from the second inequality that the algorithm arrives at an ϵ-optimal stationary point in $O(\frac{1}{\epsilon})$ iterations. We highlight that Proposition 1 does not require any special conditions on the predictor matrix X, which may be contaminated.

Neighbourhood search

Given an initial point $(\beta^{(0)}, \eta^{(0)})$ satisfying the sparsity constraints on β and η, Algorithm 1 is guaranteed to converge. However, as with most nonconvex optimisation problems, the

choice of the initial point can impact the quality of the solution produced. In general, setting $\beta^{(0)} = 0$ and $\eta^{(0)} = 0$ does not result in satisfactory solutions. To this end, we apply a neighbourhood search method that largely alleviates this issue. Such methods recently proved useful in Mazumder et al. (2023) for the ℓ_1 and ℓ_2 regularised best subsets problems. For reasons to be explained, the neighbourhood search method (as a byproduct of its design) also produces solutions to the robust subsets problem (2.8) for an entire grid of values of the parameters k and h. This set of fitted models produced by the method is useful in practice because the best predictive (k, h) is typically unknown and needs to be chosen from a set of parameters, say, $K \times H$ with $K = \{k_1, \ldots, k_q\}$ and $H = \{h_1, \ldots, h_r\}$. For instance, given data with $n = 100$ and $p = 20$ that might contain up to 25% contamination, it is natural to consider $K = \{0, \ldots, 20\}$ and $H = \{75, 80, \ldots, 100\}$.

The algorithm is conceptually simple but slightly cumbersome to write down. To assist in this effort, $\beta(k_i, h_j)$ and $\eta(k_i, h_j)$ are taken to denote variables in the robust subsets problem (2.8) with $k = k_i$ and $h = h_j$, and $\hat{\beta}(k_i, h_j)$ and $\hat{\eta}(k_i, h_j)$ as the corresponding solutions produced by Algorithm 1. We assume that $k_1 < k_2 < \cdots < k_q$ and $h_1 < h_2 < \cdots < h_r$. Algorithm 2 presents the neighbourhood search method. Algorithm 2 first computes an initial set of solutions corresponding to the parameter

Algorithm 2: Neighbourhood search

input : $K = \{k_1, \ldots, k_q\}$, $H = \{h_1, \ldots, h_r\}$, and $\epsilon > 0$.

1 For all $(i, j) \in [q] \times [r]$, run Algorithm 1 initialised with

$$\beta^{(0)}(k_i, h_j) = 0 \quad \text{and} \quad \eta^{(0)}(k_i, h_j) = 0.$$

2 Repeat the following step for all $(i, j) \in [q] \times [r]$:
Take the neighbourhood of (i, j) as

$$\mathcal{N}(i, j) = \{a \in [q], b \in [r] : |i - a| + |j - b| = 1\}.$$

For all $(a, b) \in \mathcal{N}(i, j)$, run Algorithm 1 initialised with

$$\beta^{(0)}(k_i, h_j) = \mathrm{H}\left(\hat{\beta}(k_a, h_b); k_i\right) \quad \text{and} \quad \eta^{(0)}(k_i, h_j) = \mathrm{H}\left(\hat{\eta}(k_a, h_b); n - h_j\right).$$

If the best solution obtained from the neighbourhood initialisations improves on the incumbent solution, update $\hat{\beta}(k_i, h_j)$ and $\hat{\eta}(k_i, h_j)$ with the best solution.

3 Repeat step 2 until successive changes in $\sum_{i=1}^{q} \sum_{j=1}^{r} f\left(\hat{\beta}(k_i, h_j), \hat{\eta}(k_i, h_j)\right)$ are ϵ small.

set $K \times H$ by running Algorithm 1 initialised with zero vectors. In the second step, it progresses through $K \times H$, at each stage fixing (k, h) at (k_i, h_j) and initialising Algorithm 1 with the solutions that neighbor (k_i, h_j). The neighbouring solutions are usually small perturbations to the support of the incumbent solution, which often leads to the discovery of new feasible solutions. The final step involves recursively iterating this update scheme until no further improvements can be made. It is our experience that approximately 10-20 rounds of updates are typically required to achieve convergence.

2.3.3 Parameter choices

Big-M parameters

To operationalise the mixed-integer program described in this section, it is necessary to choose suitable values of the Big-M parameters (e.g., $\mathcal{M}_\beta$). Large values of these parameters can lead to numerical issues and poor solver performance, and thus we wish to set them to values as small as reasonably possible. A simple approach is to take $\mathcal{M}_\beta = \tau\|\hat{\beta}\|_\infty$ for some $\tau \geq 1$, where $\hat{\beta}$ is a solution obtained from the heuristics. The same estimation process can be applied for other Big-M parameters.

By estimating the Big-M parameters using the process described here, there exists a possibility of excluding the true optimal solution from the feasible set. It remains an open research question how to estimate provably correct parameters in the absence of any assumptions on X. As an alternative to Big-M constraints, our implementation also supports SOS constraints. SOS constraints never exclude the true optimal solution but are less numerically efficient.

Sparsity and robustness parameters

The choice of k and h plays a critical role. Choosing a value of k that is too large leads the estimator to overfit to the data, and choosing a value of k that is too small leads the estimator to underfit. Similarly, choosing a large h makes the estimator susceptible to breakdown, and choosing a small h leads to a loss of efficiency (because it might lead to discarding good data). Given that we are interested in building predictive models, cross-validation can address these issues. However, because the data might be contaminated, standard cross-validation metrics such as mean square prediction error are inappropriate. A suitable alternative is trimmed (mean square) prediction error, which trims a portion of the largest square errors. Letting $e_{(1)}, \ldots, e_{(n)}$ denote prediction errors ordered by absolute value, the trimmed prediction error can be written as

$$\text{Trimmed prediction error} := \frac{\sum_{i=1}^{[(1-\alpha)n]} e_{(i)}^2}{[(1-\alpha)n]},$$

where the trimming parameter is typically taken conservatively as $\alpha \in \{0.25, 0.5\}$. This metric can be computed over a grid of candidate values of (k, h) using the cross-validation errors.

There are alternatives to cross-validating h. Alfons et al. (2013) applied a two-step procedure where h is initially set to a fixed proportion of n and then revised upward based on the magnitude of the residuals. However, this approach involves additional parametric assumptions and is numerically expensive like cross-validation. An avenue of future work is to investigate efficient procedures, e.g., estimating h and β jointly as can be done with Huber loss for σ (the scale parameter) and β (Huber 1981; Lambert-Lacroix and Zwald 2011).

2.4 Breakdown point

This section discusses the robustness of robust subset selection in terms of its finite-sample breakdown point. The notion of a finite-sample breakdown point originated with Hampel (1971) and Donoho and Huber (1983) and has since become a standard measure for robustness in the casewise contamination framework. Roughly speaking, the breakdown

point of an estimator is the minimum fraction of contaminated observations required to corrupt the estimator arbitrarily badly. A formal definition is given as follows.

Definition 2. *Let (X, Y) be an uncontaminated sample of size n, and let $(\tilde{X}, \tilde{Y})$ be the same sample with $1 \leq m \leq n$ observations replaced arbitrarily. Let $\Theta(X, Y)$ be an estimator given the sample (X, Y). Then the finite-sample breakdown point of $\Theta(X, Y)$ is defined as*

$$b(\Theta; X, Y) := \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : \sup_{(\tilde{X}, \tilde{Y})} \|\Theta(X, Y) - \Theta(\tilde{X}, \tilde{Y})\|_2 = \infty \right\}.$$

We take $\Theta(X, Y)$ to be the objective value of the robust subsets estimator. Thus, with the above definition in mind, the main result of this section is written as follows.

Theorem 1. *Let (X, Y) be a sample of size n, and let $\Theta(X, Y)$ be the optimal objective value to the robust subset selection problem (2.2) with $h \leq n$. Then $\Theta(X, Y)$ has the finite-sample breakdown point*

$$b(\Theta; X, Y) = \frac{n - h + 1}{n}.$$

The proof follows steps similar to those used in the proof of the breakdown point in Bertsimas and Mazumder (2014) for the objective value of the least quantile of squares estimator.

It follows from Theorem 1 that robust subset selection can withstand up to $n - h$ contaminated observations. Moreover, because fixing $h = n$ yields the best subsets estimator, it follows that the breakdown point of best subset selection is $1/n$, meaning it is not robust to any level of contamination in the data. These results are consistent with experimental evidence provided in the following section.

We close this section with three remarks pertaining to the parameter h:

- The performance of the solver is related to the choice of h. The most conservative choice $h = [0.5n]$ is also the most computationally cumbersome. Therefore, it is desirable in terms of computation to choose a value of h that is as large as reasonably possible.
- If the analyst is comfortable that the data contain no more than 25% contamination, taking $h = [0.75n]$ is generally accepted as a good compromise between efficiency and robustness (Rousseeuw and Van Driessen 2006).
- Using cross-validation as outlined in the previous section can alleviate the need to manually set h. Nonetheless, the trimming parameter α in the cross-validation metric remains to be chosen, but 25% again seems a judicious choice.

2.5 Experiments

We perform a series of numerical experiments on synthetic data to evaluate the performance of our estimator and algorithms in a variety of scenarios. The experiments of Section 2.5.1 compare robust subsets with existing estimators, while those of Section 2.5.2 compare different algorithmic approaches for its computation.

In support of these exercises, the methods described in Section 2.3 were implemented in the `R` package `robustsubsets`. Our package calls `Gurobi` as the mixed-integer solver

and implements Algorithms 1 and 2 in `C++`. The package first runs neighbourhood search over a specified parameter grid $K \times H$ and then runs mixed-integer optimisation using the warm starts and variable bounds from neighbourhood search. The data are standardised to have zero median and unit (normalised) median absolute deviation in advance of fitting the model. To obtain the best subsets estimator, the set H is taken as $\{n\}$, and the data are standardised to have zero mean and unit standard deviation. The final model fits are returned on the original scale of the data.

All experiments are carried out using `R` 4.1.0 and `Gurobi` 9.1.2.

2.5.1 Comparisons of estimators

Setup

We study the linear model

$$Y = X\beta^0 + \varepsilon, \quad \varepsilon \sim N(0, I\sigma^2),$$

with the entries of the coefficient vector β^0 drawn randomly from $\{-1, 0, 1\}$ and the number of nonzero coefficients $p_0 := \|\beta^0\|_0 \in \{5, 10\}$. The rows of the predictor matrix X are sampled iid as $x_i \sim N(0, \Sigma)$, where Σ has row i and column j constructed as $0.35^{|i-j|}$ for all $i, j \in [p]$. The noise variance σ^2 is chosen to yield the desired signal-to-noise ratio (SNR), where

$$\text{SNR} := \frac{\text{Var}(x^T\beta^0)}{\sigma^2} = \frac{(\beta^0)^T\Sigma\beta^0}{\sigma^2}.$$

We take $\text{SNR} \in \{1, 4\}$ when $p_0 = 5$ and $\text{SNR} \in \{4, 9\}$ when $p_0 = 10$, with $\text{SNR} = 1$ corresponding to 50% proportion of variance explained (PVE), where

$$\text{PVE} := \frac{\text{Var}(x^T\beta^0)}{\text{Var}(y)} = \frac{\text{SNR}}{\text{SNR} + 1}.$$

We study a low-dimensional setup in which $n = 500$ and $p = 100$, and a high-dimensional setup in which $n = 100$ and $p = 500$.

We consider four contamination settings:

1. No contamination - The response and predictors are both uncontaminated.
2. Contamination of Y - The response is contaminated by sampling the noise as a mixture of normal distributions: $\varepsilon_i \sim (1 - \delta)N(0, \sigma^2) + \delta N(10\sigma, \sigma^2)$, $i \in [n]$.
3. Contamination of X - The rows of the predictor matrix are first sampled as $N(0, \Sigma)$ and the response is generated. Each row of X is subsequently contaminated with probability δ by randomly selecting $0.1p$ predictors and replacing their values with independent draws from a $N(10, 1)$ distribution.
4. Contamination of Y and X - The response and predictors are both contaminated as described above.

The contamination probability $\delta = 0.1$. The expected number of contaminated observations in the sample (X, Y) is thus $0.1n$ under settings two and three and $0.19n$ under setting four. Settings one, two, and four are similar to those studied in Chang et al. (2018). Setting three is added to evaluate the effect of contamination in X alone.

To benchmark robust subsets, three contemporary sparse and robust estimators are also evaluated: lasso with Tukey's bisquare loss (MM lasso) via `pense` 1.2.9 (Smucler and

Yohai 2017; Cohen Freue et al. 2019), lasso with LTS loss (sparse LTS) via `robustHD` 0.6.1 (Alfons et al. 2013), and lasso with Huber loss (Huber lasso) via `hqreg` 1.4 (Rosset and Zhu 2007; Yi and Huang 2017). The availability of high-quality implementations of these estimators underscores their relevance to practitioners. We also evaluate the vanilla (least squares) lasso as implemented in `glmnet` 4.1-1.

For robust subsets, the tuning parameters k and h are swept over the grid $K \times H$ with $K = \{0, \dots, 20\}$ and $H = \{[0.75n], [0.80n], \dots, n\}$. For best subsets, k is swept over the same K. For the lasso estimators, we sweep the tuning parameter λ over 50 values linearly spaced on the log scale, with the maximum λ set according to the default of each package.

To measure expected *out-of-sample* prediction loss, we study the relative prediction error:

$$\begin{aligned}\text{Relative prediction error} &:= \frac{\mathrm{E}[(y - \hat{\mu} - x^T\hat{\beta})^2 \mid X, Y]}{\sigma^2} \\ &= \frac{(\beta^0 - \hat{\beta})^T\Sigma(\beta^0 - \hat{\beta}) + \hat{\mu}^2 + \sigma^2}{\sigma^2},\end{aligned}$$

where $\hat{\mu}$ is an estimate of the intercept (the true intercept is zero), and $\hat{\beta}$ is an estimate of β^0. The best attainable relative prediction error is 1, while the relative prediction error of the null model with $\hat{\beta} = 0$ is $\mathrm{SNR} + 1$. We also study the sparsity of the fitted model:

$$\text{Model sparsity} := \|\hat{\beta}\|_0,$$

and, to measure support recovery, the F1 score:

$$\text{F1 score} := \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}},$$

which is the harmonic average of recall (the true positive rate) and precision (the positive predictive value). The best attainable F1 score is 1, indicating that the support of $\hat{\beta}$ exactly matches that of β^0. Hastie et al. (2020) considered these three metrics (relative prediction error, model sparsity, and F1 score) in their comparisons of best subsets and lasso. The metrics are all evaluated with respect to tuning parameters chosen via 10-fold cross-validation.[2] The cross-validation metrics are (mean square) prediction error for the nonrobust estimators and trimmed prediction error with 25% trimming for the robust estimators. For best subsets and robust subsets, only the heuristics are used during cross-validation to maintain reasonable runtime. Mixed-integer optimisation is run on the k and h yielding the lowest cross-validation error with variable bounds estimated by the method of Section 2.3.3 using $\tau = 1.5$.

Results

We conduct 30 simulations for each set of simulation parameters and then aggregate the results. The simulations are performed in parallel, each running on a single core of an AMD Ryzen Threadripper 3970x. In the interest of space, we confine the results and discussion here to sparsity level $p_0 = 5$ and relegate those for $p_0 = 10$ to Appendix A.3.1.

[2]The parameter h in sparse LTS is not treated as a tuning parameter in `robustHD`, and is instead fixed at 75% of the sample size. After the model is initially fit, `robustHD` applies a reweighting step to improve efficiency. The reader is referred to Alfons et al. (2013) for details.

Figures 2.1, 2.2, and 2.3 report the relative prediction error, model sparsity, and F1 score, respectively. The vertical bars represent averages, and the error bars denote (one) standard errors. The dashed horizontal lines in Figure 2.1 indicate the relative prediction error from the null model, and those in Figure 2.2 indicate the sparsity of the true model.

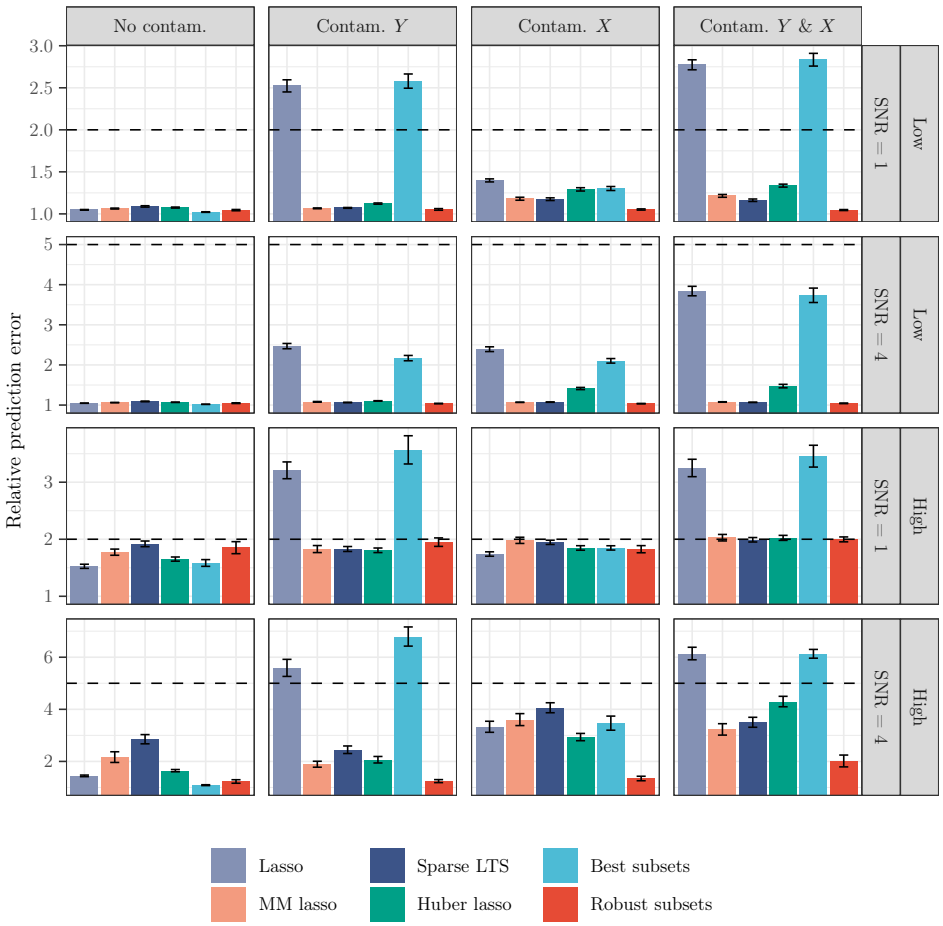

Figure 2.1: Relative prediction error estimated over 30 simulations with $p_0 = 5$. The vertical bars represent averages, and the error bars denote (one) standard errors. The dashed horizontal lines indicate the relative prediction error from the null model.

In the uncontaminated settings, best subsets exhibits excellent prediction accuracy and support recovery. Lasso, while also showing excellent prediction accuracy, has inferior support recovery compared with best subsets. The F1 scores of lasso are hindered by the large number of irrelevant predictors it picks up. When contamination is introduced, both best subsets and lasso display significant performance degradation across the board. Only in the Low-4 configuration (low-dimensional setup with SNR $= 4$), where there are relatively few predictors and the signal is strong, do they consistently outperform the null model in terms of prediction accuracy.

In the contaminated settings, robust subsets ameliorates the degradation in performance that occurs in best subsets. Robust subsets behaves in a manner similar to that of best subsets as if it were applied to a reduced set of 'good' observations. Similarly, MM lasso, sparse LTS, and Huber lasso largely retain the operational characteristics

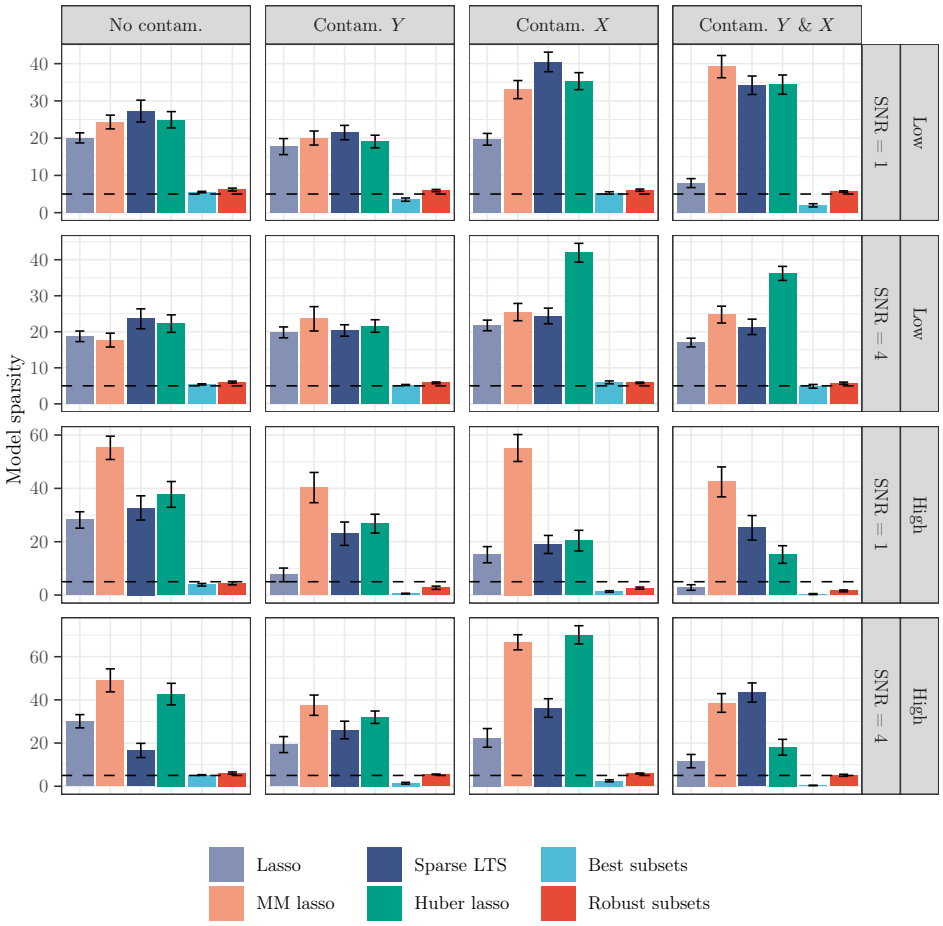

Figure 2.2: Model sparsity estimated over 30 simulations with $p_0 = 5$. The vertical bars represent averages, and the error bars denote (one) standard errors. The dashed horizontal lines indicate the true model sparsity.

of lasso, though the latter struggles with contamination in X. In terms of prediction accuracy, robust subsets produces fitted models that are competitive with and often superior to those from the robust lasso estimators. The main success story is the High-4 configuration, in which robust subsets improves markedly on its competitors across all contamination settings.

Robust subsets inherits the good support recovery qualities of best subsets. In almost all cases, robust subsets enjoys the highest F1 score among the robust estimators. Upon closer inspection, the robust lasso estimators perform slightly better in detecting true positives. However, they pay a steep price in the number of false positives they produce, leading to lower F1 scores overall. They also produce relatively dense models. For instance, when X is contaminated, MM lasso selects up to 60 predictors, 12 times the true number of nonzeros. On the other hand, robust subsets consistently delivers models that more closely reflect the true sparsity level.

Generally, the robust estimators struggle more with contamination in the predictors than in the response. In fact, in the High-1 configuration, when X is contaminated, all robust estimators fail to predict better than their nonrobust counterparts and offer little to no improvement on the null model. This result suggests that building good

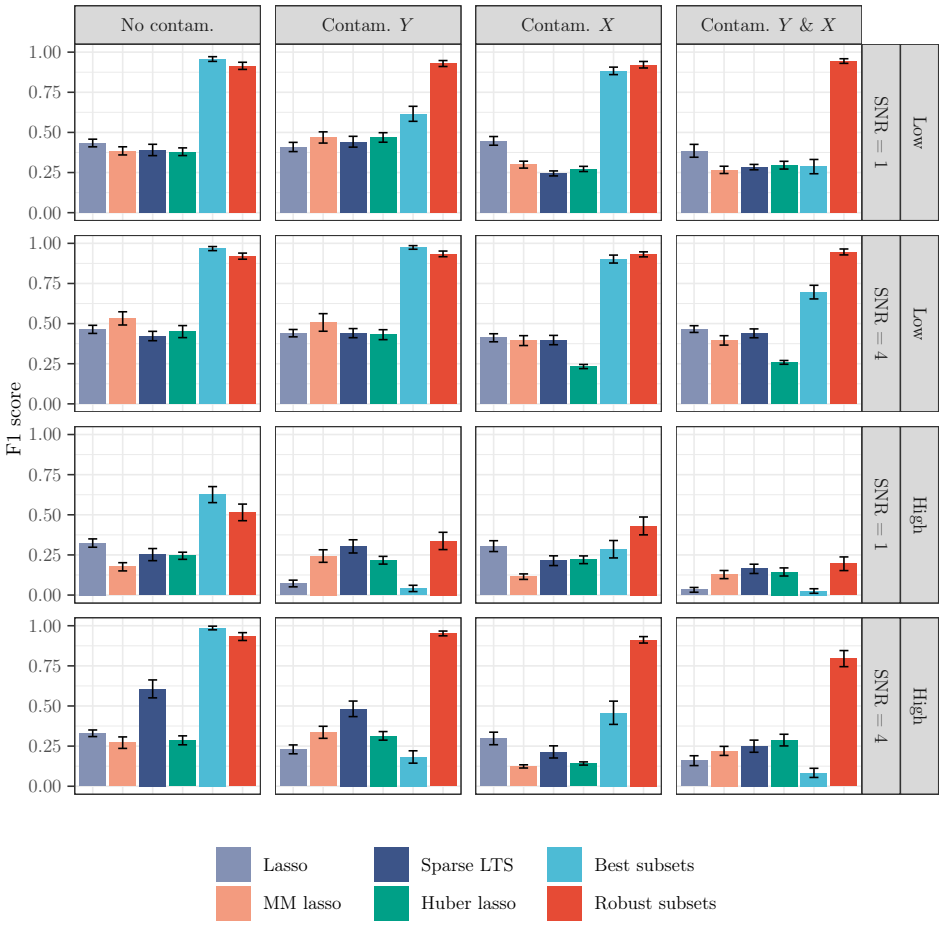

Figure 2.3: F1 score estimated over 30 simulations with $p_0 = 5$. The vertical bars represent averages, and the error bars denote (one) standard errors.

predictive models may be an unachievable goal when the signal is weak and the number of contaminated predictors is large. However, when the signal is strong, as in the High-4 configuration, robust subsets is able to offer improvement. In fact, robust subsets performs about as well with regard to support recovery as it does in the low-dimensional setups.

The results of this section follow from choosing the estimators' tuning parameters via 10-fold cross-validation. It would be interesting to evaluate other tuning mechanisms, e.g., Akaike's information criterion (AIC, Akaike 1973) or Schwarz's Bayesian information criterion (BIC, Schwarz 1978). However, the resulting experiments would be formidable and potentially distract from our focus on the core strengths and weaknesses of the different estimators. Roughly speaking, AIC should provide models with better in-sample fit than BIC, while BIC should yield better support recovery. It is unclear whether either criterion would change the relative ranking of the estimators compared with the finite-sample results here, though there is an asymptotic equivalence between (leave-one-out) cross-validation and AIC (Stone 1977). We refer the interested reader to Ding et al. (2018) for an extended discussion on this topic.

2.5.2 Comparisons of algorithms

Setup

The previous experiments compare robust subsets with existing estimators—the experiments below provide further insight into the algorithms underlying robust subsets. The simulation design remains as before. We focus on the high-dimensional setup, fixing $n = 100$ and taking $p \in \{500, 1,000\}$. The number of nonzeros $p_0 = 5$ and SNR $= 4$. We consider contamination setting two where Y is contaminated and setting four where X is also contaminated. The total proportion of contaminated observations in both settings is fixed at 10%, with contamination evenly split between Y and X in setting four.

To isolate the individual contributions of each algorithm in our framework, we consider several different computational approaches: neighbourhood search (K and H specified as before) without mixed-integer optimisation, mixed-integer optimisation without warm starts or variable bounds, and mixed-integer optimisation with warm starts and variable bounds from neighbourhood search. These approaches are respectively labeled 'heuristics', 'MIO', and 'MIO+heuristics'. The variable bounds for the last approach are again estimated using the method of Section 2.3.3 with $\tau \in \{1, 1.5\}$. The MIO approach uses SOS constraints in place of Big-M constraints since variable bounds are unavailable. The solver is run on an AMD Ryzen Threadripper 3970x with a 30 minute time limit.

To measure the quality of the solution produced, we study the number of true positive predictors selected and the relative objective gap:

$$\text{Relative objective gap} := \frac{\hat{f} - f^\star}{f^\star},$$

where $\hat{f}$ is the attained objective value, and $f^\star$ is the best objective value among all approaches considered. To measure progress towards proving optimality, we study the relative optimality gap:

$$\text{Relative optimality gap} := \frac{\hat{f} - \hat{f}_L}{\hat{f}},$$

where $\hat{f}_L$ is the lower bound on the optimal objective value delivered by the solver.[3] A relative optimality gap of zero indicates that the computed solution is provably optimal. Additionally, we measure whether the solver terminated within the time limit (i.e., had an optimality gap of zero) and the runtime. These metrics are evaluated at $k = 5$ and $h = 90$.

Results

Table 2.1 reports results from 30 simulations for $p = 500$. Results for $p = 1,000$ are detailed in Appendix A.3.2. Averages or proportions are reported with standard errors in parentheses.

When only Y is contaminated, all algorithmic approaches are equally effective at delivering high-quality solutions as measured by the number of true positive selections and the objective gap. There is, however, significant disparity in terms of the optimality gap. Without warm starts or variable bounds, the solver is never able to improve the lower bound in the 30 minute time limit. Yet, when guided by this information from the

[3] This definition is used in `Gurobi`.

	True pos.	Obj. gap (%)	Opt. gap (%)	Term. (%)	Time (mins.)
Contamination of Y					
Heuristics	5.0 (0.0)	0.0 (0.0)	-	-	0.5 (0.0)
MIO	5.0 (0.0)	0.0 (0.0)	100.0 (0.0)	0.0 (0.0)	30.1 (0.0)
MIO+heur. (1)	5.0 (0.0)	0.0 (0.0)	0.0 (0.0)	100.0 (0.0)	0.9 (0.1)
MIO+heur. (1.5)	5.0 (0.0)	0.0 (0.0)	25.5 (5.4)	53.3 (9.1)	19.1 (2.1)
Contamination of Y and X					
Heuristics	4.9 (0.1)	1.9 (1.6)	-	-	2.5 (0.1)
MIO	5.0 (0.0)	0.0 (0.0)	100.0 (0.0)	0.0 (0.0)	30.1 (0.0)
MIO+heur. (1)	5.0 (0.0)	0.0 (0.0)	30.5 (5.4)	33.3 (8.6)	26.3 (1.8)
MIO+heur. (1.5)	5.0 (0.0)	0.0 (0.0)	96.3 (1.2)	0.0 (0.0)	32.6 (0.1)

Table 2.1: True positive selections, relative objective gap, relative optimality gap, termination rate, and runtime estimated over 30 simulations with $n = 100$, $p = 500$, $p_0 = 5$, and SNR $= 4$. Averages or proportions are reported next to (one) standard errors in parentheses.

heuristics, the solver proves optimality in 100% of the simulation instances for $\tau = 1$, typically within a minute. These same figures are 53% and 19 minutes for $\tau = 1.5$, confirming that small values of τ play an important role in determining the speed at which the optimality gap is closed.

When Y and X are contaminated, the heuristic solutions are slightly lower-quality. The three mixed-integer approaches continue to generate excellent solutions. For $\tau = 1$, the solver attains a zero optimality gap in a third of the simulation instances. The optimality gap is much looser for $\tau = 1.5$. More generous time limits (roughly several hours to tens of hours) are required to close the optimality gap in most cases. The heuristics now take longer to converge than when Y is contaminated, 2 minutes more on average. Likewise, the solver takes longer to progress the lower bound regardless of the value for τ. These longer runtimes and weaker optimality gaps can be attributed to poorer conditioning of the gram matrix $X^T X$.

The speed at which the solver is able to close the optimality gap can also be impacted by the degree to which observations are outlying. This impact is a consequence of the fact that the Big-M parameter $\mathcal{M}_\eta$ (the outlier bound) must be of the same order of magnitude as the outliers. Another contributing factor is that the condition number of the gram matrix is affected by the size of the outliers in X. Overall, smaller outliers mean tighter optimality gaps and smaller runtimes. Larger outliers have the opposite effect.

Additional results in the appendices show that the proposed methods scale to $p = 1,000$ in reasonable time. It is difficult, however, to close the optimality gap in the 30 minute time limit. Nevertheless, when warm-started, the solver usually finds the optimal solution in the first few minutes, while the remaining time is spent proving optimality. Thus, if the problem instance appears intractable, it can be sufficient to run the solver for a short time to obtain a high-quality solution. Scaling to instances with p (or n) in the tens or hundreds of thousands while proving optimality is a direction for future research.

Besides the type of contamination and number of predictors, which we vary in these experiments, several other factors determine the benefit of running the solver after the heuristics. Though it is difficult to quantify exact gains in general, our experience is that mixed-integer optimisation is most useful when the number of contaminated observations, number of nonzero coefficients, or levels of correlation among predictors are high. The

SNR does not appear to be an important factor.

2.6 Archaeological glass vessels dataset

This section illustrates an application of robust subsets using the archaeological glass vessels dataset introduced in Janssens et al. (1998) and Lemberge et al. (2000) and studied vis-à-vis sparsity and robustness in Smucler and Yohai (2017) and Amato et al. (2021). The dataset was obtained from the supplemental material of Christidis et al. (2020). It consists of observations on 180 glass vessels from the 16th and 17th centuries that were uncovered in archaeological excavations in Antwerp, Belgium. To understand the origins of the vessels and the trade connections between producers, the chemical compositions of the vessels were studied. To determine these compositions, electron-probe X-ray microanalysis (EPXMA) was used to produce data containing EPXMA intensities for 1920 different frequencies on each of the glass vessels. The data were subsequently processed to yield the concentrations of chemical compounds present in the vessels. However, the processing step is time-consuming and challenging to automate, and thus interest exists in regression methods that can directly predict the concentrations using the EPXMA data. The majority of estimators do not perform well at this task because the data are high-dimensional and include multiple observations that constitute outliers (Serneels et al. 2005; Maronna 2011).

Following Smucler and Yohai (2017), the response is taken as the concentration of the chemical compound PbO and the predictors as the frequencies 15 through 500. Frequencies outside this range have little variation and are almost null. The resulting sample has $n = 180$ and $p = 486$. Each estimator is applied to this full sample. Lasso produces a fitted model containing 29 frequencies. MM lasso selects 27 frequencies, sparse LTS selects 11, and Huber lasso selects 2. Best subsets selects 10 frequencies, while robust subsets produces a model with 3. To glean insight into these fitted models, we present Figure 2.4, which indicates the nonzero coefficients in each model. It is apparent

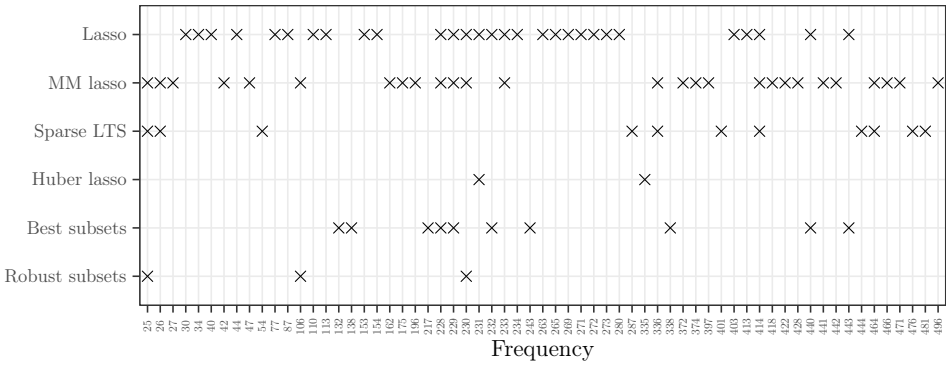

Figure 2.4: Selected frequencies (predictors) for the archaeological glass vessels dataset. The marks identify the frequencies with nonzero coefficients in the fitted models.

from this figure that the robust subsets and best subsets models do not overlap at all. The story is largely similar for lasso, with only 6 of its 29 predictors shared with its robust adaptions. While Huber lasso delivers a highly sparse model, its frequencies are not shared with the other robust estimators, possibly because it does not try to resist contamination in X. On the other hand, every frequency picked up by robust subsets is

also picked up by MM lasso.

To evaluate the prediction accuracy of the competing estimators, we use 10-fold cross-validation and record the trimmed prediction error. Figure 2.5 reports this metric across several levels of trimming to accommodate various severities of contamination. The vertical bars represent averages, and the error bars denote (one) standard errors. For

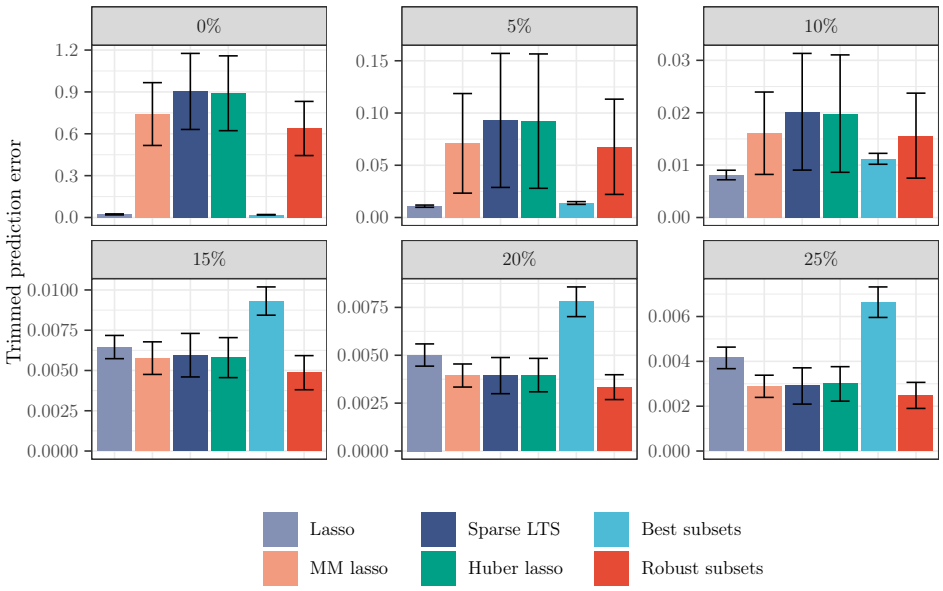

Figure 2.5: Trimmed prediction error, expressed as a function of the trimming level, estimated via 10-fold cross-validation for the archaeological glass vessels dataset. The vertical bars represent averages, and the error bars denote (one) standard errors.

low levels of trimming, the nonrobust estimators dominate in terms of prediction error. This result is not particularly surprising because any outliers will inflate the prediction error. The transition point at which all estimators fare similarly occurs between the 10% and 15% trimming levels. At higher levels of trimming, the robust estimators outperform their nonrobust counterparts. In particular, robust subsets improves substantially on best subsets and yields the smallest prediction error on average among the competing estimators.

2.7 Concluding remarks

Best subset selection is a classic tool for sparse regression, and recent developments in mathematical optimisation have paved the way for exciting new research into this estimator. Inspired by these developments, this chapter proposes robust subset selection, an adaption of best subset selection that it is resistant to casewise contaminated data. The combinatorial problem that defines robust subsets is shown to be amenable to mixed-integer optimisation, a technology that continues to display tremendous improvements. To speed up runtime, heuristic methods that complement the mixed-integer optimisation approach are developed. Central to the heuristics is a projected block-coordinate gradient descent method, for which we derive convergence properties. As a statistical guarantee, the objective value of the robust subsets estimator is shown to resist a specifiable level of contamination in finite samples. Numerical experiments on synthetic and real data

yield findings consistent with this result. The ability of best subsets to recover the true support and produce good predictions is observed to deteriorate significantly if the data are contaminated. In contrast, robust subsets resists contamination by excluding from the model fit the subset of observations that induce the most substantial losses. Compared with robust adaptions of continuous shrinkage estimators, robust subsets does well to closely recover the underlying sparsity pattern. This property makes robust subsets a promising tool for applications in which the fitted models themselves are of interest and not just their prediction accuracy alone.

Our implementation `robustsubsets` is available as an `R` package at

https://github.com/ryan-thompson/robustsubsets.

Chapter 3

Group selection and shrinkage: Structured sparsity for semiparametric models

3.1 Introduction

Sparsity over group structures arises in connection with a myriad of statistical and machine learning problems, e.g., multitask learning (Obozinski et al. 2006), sparse additive modelling (Ravikumar et al. 2009), and hierarchical selection (Lim and Hastie 2015). Even sparse linear modelling can involve structured sparsity, such as when a categorical predictor is represented as a sequence of dummy variables. In certain domains, groups may emerge naturally, e.g., disaggregates of the same macroeconomic series or genes of the same biological path. The prevalence of such problems motivates principled estimation procedures capable of encoding structure into the fitted models they produce.

Given response $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, predictors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times p}$, and nonoverlapping groups $\mathcal{G}_1, \ldots, \mathcal{G}_g \subseteq \{1, \ldots, p\}$, group lasso (Yuan and Lin 2006; Meier et al. 2008) solves

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \ell\left(\mathbf{x}_i^\top \boldsymbol{\beta}, y_i\right) + \sum_{k=1}^{g} \lambda_k \|\boldsymbol{\beta}_k\|,$$

where $\ell : \mathbb{R}^2 \to \mathbb{R}_+$ is a loss function (e.g., square loss for regression or logistic loss for classification), $\lambda_1, \ldots, \lambda_g$ are nonnegative tuning parameters, and $\boldsymbol{\beta}_k \in \mathbb{R}^{p_k}$ are the coefficients $\boldsymbol{\beta}$ indexed by $\mathcal{G}_k$.[1] Group lasso couples coefficients via their l_2-norm so that all predictors in a group are selected together.

Just as lasso (Tibshirani 1996) is the continuous relaxation of the combinatorially-hard problem of best subset selection ('best subset'), so is group lasso the relaxation of the combinatorial problem of group subset selection ('group subset'):

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \ell\left(\mathbf{x}_i^\top \boldsymbol{\beta}, y_i\right) + \sum_{k=1}^{g} \lambda_k 1(\|\boldsymbol{\beta}_k\| \neq 0).$$

Unlike group lasso, which promotes group sparsity implicitly by nondifferentiability of the l_2-norm at the null vector, group subset explicitly penalises the number of nonzero groups. Consequently, one might interpret group lasso as a compromise made in the

[1] Here and throughout, the intercept term is omitted to facilitate exposition.

interest of computation. However, group lasso has a trick up its sleeve that group subset does not: shrinkage. Shrinkage estimators such as lasso are more resilient than best subset to high-levels of noise (Breiman 1996; Hastie et al. 2020). This phenomenon comes down to a classic bias-variance trade-off. The lasso achieves lower variance but biases its coefficients to zero, while best subset's unbiased coefficients come at the price of higher variance. If the noise is high, a reduction in variance can be well-worth an increase in bias.

The above consideration motivates one to shrink the group subset estimator:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \ell\left(\mathbf{x}_i^\top \boldsymbol{\beta}, y_i\right) + \sum_{k=1}^{g} \lambda_{0k} 1(\|\boldsymbol{\beta}_k\| \neq 0) + \sum_{k=1}^{g} \lambda_{1k} \|\boldsymbol{\beta}_k\|. \tag{3.1}$$

In contrast to group lasso and group subset, (3.1) directly controls both group sparsity and shrinkage via separate penalties—selection via group subset and shrinkage via group lasso. The combination of best subset and lasso in the unstructured setting results in good predictive models with low false positive selection rates across a range of noise levels (Mazumder et al. 2023). We demonstrate that the same benefit is realised in the structured setting. Specifically, our theoretical and experimental analyses confirm that shrinkage is helpful when the noise is high but harmful when it is low. The advantage of our estimator is that it can vary the shrinkage level independently of the sparsity level and hence adapts to both high and low noise regimes.

Unfortunately, the estimators (3.1), including group lasso and group subset as special cases, do not accommodate overlap among groups. Specifically, if two groups overlap, one cannot be selected independently of the other. To encode sophisticated structures, such as hierarchies or graphs, groups must often overlap. To address this issue, one can introduce group-specific vectors $\bar{\boldsymbol{\nu}}_k \in \mathbb{R}^p$ $(k = 1, \ldots, g)$ that are zero everywhere except at the positions indexed by $\mathcal{G}_k$. Letting $\mathcal{V}$ be the set of all tuples $\bar{\boldsymbol{\nu}} := (\bar{\boldsymbol{\nu}}_1, \ldots, \bar{\boldsymbol{\nu}}_g)$ with elements satisfying this property, group subset with shrinkage becomes

$$\min_{\substack{\boldsymbol{\beta}\in\mathbb{R}^p,\, \bar{\boldsymbol{\nu}}\in\mathcal{V} \\ \boldsymbol{\beta}=\sum_k \bar{\boldsymbol{\nu}}_k}} \sum_{i=1}^{n} \ell\left(\mathbf{x}_i^\top \boldsymbol{\beta}, y_i\right) + \sum_{k=1}^{g} \lambda_{0k} 1(\|\bar{\boldsymbol{\nu}}_k\| \neq 0) + \sum_{k=1}^{g} \lambda_{1k} \|\bar{\boldsymbol{\nu}}_k\|. \tag{3.2}$$

The vectors $\bar{\boldsymbol{\nu}}_1, \ldots, \bar{\boldsymbol{\nu}}_g$ are a decomposition of $\boldsymbol{\beta}$ into a sum of latent coefficients that facilitate selection of overlapping groups. For instance, if three predictors, x_1, x_2, and x_3, are spread across two groups, $\mathcal{G}_1 = \{1, 2\}$ and $\mathcal{G}_2 = \{2, 3\}$, then $\beta_1 = \bar{\nu}_{1,1}$, $\beta_2 = \bar{\nu}_{2,1} + \bar{\nu}_{2,2}$, and $\beta_3 = \bar{\nu}_{3,2}$. Since β_2 has a separate latent coefficient for each group, $\mathcal{G}_1$ or $\mathcal{G}_2$ can be selected independently of the other. This latent coefficient approach originated for group lasso with Jacob et al. (2009) and Obozinski et al. (2011). When all groups are disjoint, (3.2) reduces exactly to (3.1).

This chapter develops computational methods and statistical theory for group subset with and without shrinkage. Via the formulation (3.2), our work accommodates the general overlapping groups setting. On the computational side, we develop algorithms that scale to compute quality (approximate) solutions of the combinatorial optimisation problem. Our framework comprises coordinate descent and local search and applies to general smooth convex loss functions (i.e., regression and classification), building on recent advances for best subset (Hazimeh and Mazumder 2020; Dedieu et al. 2021). In contrast to existing computational methods for group subset (Guo et al. 2014; Bertsimas and King 2016), which rely on branch-and-bound or commercial mixed-integer optimisers, our methods scale to instances with millions of predictors or groups. We implement

our framework in the publicly available `R` package `grpsel`. On the statistical side, we establish new error bounds for group subset with and without shrinkage. The bounds apply in the overlapping setting and allow for model misspecification. The analysis sheds light on the advantages of structured sparsity and the benefits of shrinkage.

The new estimators have application to a broad range of statistical problems. We focus on sparse semiparametric modelling, a procedure wherein y is modelled via a sum of functions $\sum_j f_j(x_j)$, and f_j can be zero, linear, or nonlinear. Chouldechova and Hastie (2015) and Lou et al. (2016) estimate these flexible models using group lasso with overlapping groups and regression splines. After conducting synthetic experiments on the efficacy of our estimators in fitting these models, we carry out two empirical studies. The first study involves modelling supermarket foot traffic using sales volumes on different products. Only a fraction of supermarket products are traded in volume, necessitating sparsity. The second study involves modelling recessionary periods in the economy using macroeconomic series. The macroeconomic literature contains many examples of sparse linear modelling (De Mol et al. 2008; Li and Chen 2014), yet theory does not dictate linearity. Together these studies suggest semiparametric models are an excellent compromise between fully linear and fully nonparametric alternatives.

Independently and concurrently to this work, Hazimeh et al. (2023) study computation and theory for group subset with nonoverlapping groups. Their algorithms likewise build on Hazimeh and Mazumder (2020) but apply only to square loss regression. Also related is Zhang et al. (2023) who propose a computational 'splicing' technique for group subset that appears promising, though they do not consider overlapping groups or shrinkage.

3.1.1 Organisation

The chapter is structured as follows. Section 3.2 presents computational methods. Section 3.3 provides statistical theory. Section 3.4 describes simulation experiments. Section 3.5 reports data analyses. Section 3.6 closes the chapter. All proofs are relegated to the appendices.

3.2 Computation

This section introduces our optimisation framework and its key components: coordinate descent and local search. The framework applies to any smooth loss function $\ell(z, y)$ convex in z. The discussion below addresses the specific cases of square loss $\ell(z, y) = (y - z)^2/2$, which is suitable for regression, and logistic loss $\ell(z, y) = -y \log(z) - (1 - y) \log(1 - z)$, which is suitable for classification. Both loss functions are implemented in `grpsel`. Throughout this section, the predictor matrix $\mathbf{X}$ is assumed to have columns with mean zero and unit l_2-norm.

3.2.1 Problem reformulation

From a computational perspective, it helps to reformulate the group subset problem (3.2) as an unconstrained minimisation problem involving only the latent coefficients $\bar{\boldsymbol{\nu}}$. For this task, we denote by $\boldsymbol{\nu}_k \in \mathbb{R}^{p_k}$ the restriction of $\bar{\boldsymbol{\nu}}_k$ to the coordinates indexed by group k. No information is lost in this restriction since all elements not indexed by $\mathcal{G}_k$ are zero. We also introduce the vector $\boldsymbol{\nu} := (\boldsymbol{\nu}_1^\top, \dots, \boldsymbol{\nu}_g^\top)^\top \in \mathbb{R}^{\sum_{k=1}^g p_k}$ formed by vertically concatenating the vectors $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_g$. Consider now the unconstrained minimisation

problem

$$\min_{\boldsymbol{\nu}\in\mathbb{R}^{\sum_{k=1}^{g} p_k}} F(\boldsymbol{\nu}) := L(\boldsymbol{\nu}) + R(\boldsymbol{\nu}).$$

Here, the function $L(\boldsymbol{\nu})$ is the loss term:

$$L(\boldsymbol{\nu}) := \sum_{i=1}^{n} \ell\left(\sum_{k=1}^{g} \mathbf{x}_{ik}^{\top}\boldsymbol{\nu}_k, y_i\right),$$

where $\mathbf{x}_{ik}$ is the ith row of the matrix $\mathbf{X}_k$, with $\mathbf{X}_k$ the restriction of $\mathbf{X}$ to the columns indexed by group k. The function $R(\boldsymbol{\nu})$ is the regulariser term:

$$R(\boldsymbol{\nu}) := \sum_{k=1}^{g} \left(\lambda_{0k} 1(\|\boldsymbol{\nu}_k\| \neq 0) + \lambda_{1k}\|\boldsymbol{\nu}_k\|\right).$$

Observe that the loss $L(\boldsymbol{\nu})$ is exactly equivalent to that in (3.2) since

$$\sum_{i=1}^{n} \ell\left(\sum_{k=1}^{g} \mathbf{x}_{ik}^{\top}\boldsymbol{\nu}_k, y_i\right) = \sum_{i=1}^{n} \ell\left(\mathbf{x}_i^{\top}\sum_{k=1}^{g} \bar{\boldsymbol{\nu}}_k, y_i\right) = \sum_{i=1}^{n} \ell\left(\mathbf{x}_i^{\top}\boldsymbol{\beta}, y_i\right).$$

The regulariser $R(\boldsymbol{\nu})$ is likewise equivalent because $\|\boldsymbol{\nu}_k\| = \|\bar{\boldsymbol{\nu}}_k\|$. As the equalities immediately above suggest, it is straightforward to recover $\boldsymbol{\beta}$ from $\boldsymbol{\nu}$.

3.2.2 Coordinate descent

Coordinate descent algorithms are optimisation routines that minimise along successive coordinate hyperplanes. The coordinate descent scheme developed here iteratively fixes all but one group of coordinates (a coordinate group) and minimises in the directions of these coordinates.

The objective function $F(\boldsymbol{\nu})$ is a sum of smooth convex and discontinuous nonconvex functions and is hence discontinuous nonconvex. The minimisation problem with respect to group k is

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^{p_k}} F(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}, \boldsymbol{\xi}, \boldsymbol{\nu}_{k+1}, \ldots, \boldsymbol{\nu}_g). \tag{3.3}$$

The complexity of this coordinate-wise minimisation depends on the type of loss function and the properties of the group matrix $\mathbf{X}_k$. In the case of square loss, the minimisation involves a least-squares fit in p_k coordinates, taking $O(p_k^2 n)$ operations. To bypass these involved computations, a partial minimisation scheme is adopted whereby each coordinate group is updated using a single gradient descent step taken with respect to that group. This scheme results from a standard technique of replacing the objective function with a surrogate function that is an upper bound. To this end, we require Lemma 2.

Lemma 2. *Let $L : \mathbb{R}^{\sum_{k=1}^{g} p_k} \to \mathbb{R}$ be a continuously differentiable function. Suppose there exists a $c_k > 0$ such that the gradient of L with respect to the kth coordinate group satisfies the Lipschitz property*

$$\|\nabla_k L(\boldsymbol{\nu}) - \nabla_k L(\tilde{\boldsymbol{\nu}})\| \leq c_k\|\boldsymbol{\nu}_k - \tilde{\boldsymbol{\nu}}_k\|,$$

for all $\boldsymbol{\nu} \in \mathbb{R}^{\sum_{k=1}^{g} p_k}$ and $\tilde{\boldsymbol{\nu}} \in \mathbb{R}^{\sum_{k=1}^{g} p_k}$ that differ only in group k. Then it holds

$$L(\boldsymbol{\nu}) \leq \bar{L}_{\bar{c}_k}(\boldsymbol{\nu}; \tilde{\boldsymbol{\nu}}) := L(\tilde{\boldsymbol{\nu}}) + \nabla_k L(\tilde{\boldsymbol{\nu}})^{\top}(\boldsymbol{\nu}_k - \tilde{\boldsymbol{\nu}}_k) + \frac{\bar{c}_k}{2}\|\boldsymbol{\nu}_k - \tilde{\boldsymbol{\nu}}_k\|^2, \tag{3.4}$$

for any $\bar{c}_k \geq c_k$.

Lemma 2 is the block descent lemma of Beck and Tetruashvili (2013), which holds under a Lipschitz condition on the coordinate-wise gradients of $L(\boldsymbol{\nu})$. This condition is satisfied for square loss with $c_k = \gamma_k^2$ and for logistic loss with $c_k = \gamma_k^2/4$, where γ_k is the maximal eigenvalue of $\mathbf{X}_k^\top \mathbf{X}_k$. Using the result of Lemma 2, an upper bound of $F(\boldsymbol{\nu})$, treated as a function in group k, is given by

$$\bar{F}_{\bar{c}_k}(\boldsymbol{\nu}; \tilde{\boldsymbol{\nu}}) := \bar{L}_{\bar{c}_k}(\boldsymbol{\nu}; \tilde{\boldsymbol{\nu}}) + R(\boldsymbol{\nu}). \tag{3.5}$$

Thus, in place of the minimisation (3.3), we use the minimisation

$$\min_{\boldsymbol{\xi} \in \mathbb{R}^{p_k}} \bar{F}_{\bar{c}_k}(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1}, \boldsymbol{\xi}, \boldsymbol{\nu}_{k+1}, \ldots, \boldsymbol{\nu}_g; \tilde{\boldsymbol{\nu}}). \tag{3.6}$$

This new problem admits a simple analytical solution, given by Proposition 2.

Proposition 2. *Define the thresholding function*

$$T_c(\boldsymbol{\xi}; \lambda_0, \lambda_1) := \begin{cases} \left(1 - \dfrac{\lambda_1}{c\|\boldsymbol{\xi}\|}\right)_+ \boldsymbol{\xi} & \text{if } \left(1 - \frac{\lambda_1}{c\|\boldsymbol{\xi}\|}\right)_+ \|\boldsymbol{\xi}\| \geq \sqrt{\dfrac{2\lambda_0}{c}} \\ \mathbf{0} & \text{otherwise,} \end{cases} \tag{3.7}$$

where $(x)_+$ *is shorthand for* $\max(x, 0)$. *Then the coordinate-wise minimisation problem* (3.6) *is solved by*

$$\boldsymbol{\nu}_k^\star = T_{\bar{c}_k}\left(\tilde{\boldsymbol{\nu}}_k - \frac{1}{\bar{c}_k} \nabla_k L(\tilde{\boldsymbol{\nu}}); \lambda_{0k}, \lambda_{1k}\right).$$

Proposition 2 states that a minimiser is given by appropriately thresholding a gradient descent update to coordinate group k. For both square and logistic loss, the gradient $\nabla_k L(\tilde{\boldsymbol{\nu}})$ can be expressed as

$$\nabla_k L(\tilde{\boldsymbol{\nu}}) = -\mathbf{X}_k^\top \mathbf{r},$$

where $\mathbf{r} = \mathbf{y} - \sum_{j=1}^g \mathbf{X}_j \tilde{\boldsymbol{\nu}}_j$ for square loss and $\mathbf{r} = \mathbf{y} - (1 + \exp(-\sum_{j=1}^g \mathbf{X}_j \tilde{\boldsymbol{\nu}}_j))^{-1}$ for logistic loss. Hence, a solution to (3.6) can be computed in as few as $O(p_k n)$ operations.

Algorithm 3 now presents the coordinate descent scheme. Several algorithmic optimi-

Algorithm 3: Coordinate descent

input : $\boldsymbol{\nu}^{(0)} \in \mathbb{R}^{\sum_{k=1}^g p_k}$
for $m = 1, 2, \ldots$ **do**
  $\boldsymbol{\nu}^{(m)} \leftarrow \boldsymbol{\nu}^{(m-1)}$
  for $k = 1, \ldots, g$ **do**
    $\boldsymbol{\nu}_k^{(m)} \leftarrow \underset{\boldsymbol{\xi} \in \mathbb{R}^{p_k}}{\arg\min}\ \bar{F}_{\bar{c}_k}(\boldsymbol{\nu}_1^{(m)}, \ldots, \boldsymbol{\nu}_{k-1}^{(m)}, \boldsymbol{\xi}, \boldsymbol{\nu}_{k+1}^{(m)}, \ldots, \boldsymbol{\nu}_g^{(m)}; \boldsymbol{\nu}^{(m)})$
  end
  if *converged* **then break**
end
return $\boldsymbol{\nu}^{(m)}$

sations and heuristics can improve performance; these are discussed in Appendix B.1.5.

While Algorithm 3 may appear as an otherwise standard coordinate descent algorithm, the presence of the group subset penalty complicates the analysis of its convergence properties. No standard convergence results directly apply; e.g., Tseng (2001) that applies

to group lasso cannot be invoked immediately here. Hence, we work towards establishing some properties tailored to Algorithm 3. Two results are presented. Lemma 3 establishes convergence of the sequence of objective values. Theorem 2 establishes convergence of the sequence of iterates to a stationary point of $F(\boldsymbol{\nu})$ that satisfies a certain coordinate-wise property.

Lemma 3. *Let $\bar{c}_k \geq c_k$ for all $k = 1, \ldots, g$. Then the sequence of objective values $\{F(\boldsymbol{\nu}^{(m)})\}_{m \in \mathbb{N}}$ produced by Algorithm 3 is decreasing, convergent, and satisfies the inequality*

$$F(\boldsymbol{\nu}^{(m)}) - F(\boldsymbol{\nu}^{(m+1)}) \geq \sum_{k=1}^{g} \frac{\bar{c}_k - c_k}{2} \|\boldsymbol{\nu}_k^{(m+1)} - \boldsymbol{\nu}_k^{(m)}\|^2.$$

Lemma 3 is derived from the result of Lemma 2.

Theorem 2. *Let $\bar{c}_k > c_k$ for all $k = 1, \ldots, g$. Suppose $\lambda_{1k} > 0$ for all $k = 1, \ldots, g$, or no elements of $\boldsymbol{\nu}$ tend to $\pm\infty$. Then the sequence of iterates $\{\boldsymbol{\nu}^{(m)}\}_{m \in \mathbb{N}}$ produced by Algorithm 3 converge to a solution $\boldsymbol{\nu}^\star$ that is a stationary point of $F(\boldsymbol{\nu})$ satisfying the fixed point equations*

$$\boldsymbol{\nu}_k^\star = T_{\bar{c}_k}\left(\boldsymbol{\nu}_k^\star - \frac{1}{\bar{c}_k}\nabla_k L(\boldsymbol{\nu}^\star); \lambda_{0k}, \lambda_{1k}\right), \quad k = 1, \ldots, g. \tag{3.8}$$

To prove Theorem 2, Lemma 3 is used to show that the active set stabilises during coordinate descent. This property allows the group subset penalty to be treated as a fixed quantity after some finite number of iterations, and in turn, opens up the results of Tseng (2001). The fixed point equations in Theorem 2 have the interpretation that the limit point of the iterates of Algorithm 3 cannot be improved by partially minimising in the directions of any coordinate group. This notion is stronger than stationarity alone because all points satisfying (3.8) are stationary, but not all stationary points satisfy (3.8).

3.2.3 Local search

Local search methods have a long history in combinatorial optimisation. Here we present a local search method tailored specifically to the group subset problem. The proposed method generalises an algorithm that first appeared in Beck and Eldar (2013) for solving instances of unstructured sparse optimisation. Hazimeh and Mazumder (2020) and Dedieu et al. (2021) adapt it to best subset with promising results. The core idea is simple: given an incumbent solution, search a neighbourhood local to that solution for a minimiser with lower objective value by discretely optimising over a small set of coordinate groups. This scheme turns out to be useful when the predictors are strongly correlated, a situation in which coordinate descent alone may produce a poor solution.

Define the group sparsity pattern of the vector $\boldsymbol{\nu}$ to be the set of nonzero group indices:

$$\mathrm{gs}(\boldsymbol{\nu}) := \{k \in \{1, \ldots, g\} : \|\boldsymbol{\nu}_k\| \neq 0\},$$

and define the constraints sets

$$C_s^1(\boldsymbol{\nu}) := \left\{ \mathbf{z} \in \{0,1\}^{\sum_{k=1}^g p_k} : \mathrm{gs}(\mathbf{z}) \subseteq \mathrm{gs}(\boldsymbol{\nu}), \sum_{k=1}^{g} 1(\|\mathbf{z}_k\| \neq 0) \leq s \right\}$$

and

$$C_s^2(\boldsymbol{\nu}) := \left\{ \mathbf{z} \in \{0,1\}^{\sum_{k=1}^g p_k} : \mathrm{gs}(\mathbf{z}) \not\subseteq \mathrm{gs}(\boldsymbol{\nu}), \sum_{k=1}^g 1(\|\mathbf{z}_k\| \neq 0) \leq s \right\}.$$

Now, consider the optimisation problem

$$\min_{\substack{\boldsymbol{\xi} \in \mathbb{R}^{\sum_{k=1}^g p_k} \\ \mathbf{z}^1 \in C_s^1(\boldsymbol{\nu}),\, \mathbf{z}^2 \in C_s^2(\boldsymbol{\nu})}} F(\boldsymbol{\nu} - \mathbf{z}^1 \circ \boldsymbol{\nu} + \mathbf{z}^2 \circ \boldsymbol{\xi}), \tag{3.9}$$

where $\circ$ notates element-wise multiplication. Given a fixed vector $\boldsymbol{\nu}$, a solution to (3.9) is a minimiser among all ways of replacing a subset of active coordinate groups in $\boldsymbol{\nu}$ with a new subset. The complexity of the problem is dictated by s, which controls the size of these subsets. When $s = g$, the full combinatorial problem whose solution is a global minimiser of the group subset problem is recovered. For $s \ll g$, a reduced combinatorial problem is obtained whose solution space is usually orders of magnitude smaller than that of the full problem.

The limiting case $s = 1$ admits an efficient computational scheme, given in Algorithm 4, referred to hereafter as local search. Algorithm 4 comprises two low-complexity loops:

Algorithm 4: Local search

input : $\boldsymbol{\nu} \in \mathbb{R}^{\sum_{k=1}^g p_k}$
$\mathcal{A} \leftarrow \mathrm{gs}(\boldsymbol{\nu})$
for $k \in \mathcal{A}$ **do**
  for $j \notin \mathcal{A}$ **do**
    $\boldsymbol{\nu}^{(j)} \leftarrow \boldsymbol{\nu}$
    $\boldsymbol{\nu}_k^{(j)} \leftarrow \mathbf{0}$
    $\boldsymbol{\nu}_j^{(j)} \leftarrow \underset{\boldsymbol{\xi} \in \mathbb{R}^{p_k}}{\arg\min}\, F(\boldsymbol{\nu}_1^{(j)}, \ldots, \boldsymbol{\nu}_{j-1}^{(j)}, \boldsymbol{\xi}, \boldsymbol{\nu}_{j+1}^{(j)}, \ldots, \boldsymbol{\nu}_g^{(j)})$
  end
  $j^\star \leftarrow \underset{j \notin \mathcal{A}}{\arg\min}\, F(\boldsymbol{\nu}^{(j)})$
  if $F(\boldsymbol{\nu}^{(j^\star)}) < F(\boldsymbol{\nu})$ **then**
    $\boldsymbol{\nu} \leftarrow \boldsymbol{\nu}^{(j^\star)}$
    break
  end
end
return $\boldsymbol{\nu}$

an outer loop over the active set and an inner loop over the inactive set. Within the inner loop, an active coordinate group is removed, and the objective is minimised in the directions of an inactive coordinate group. This minimisation problem can be solved by iterating the thresholding operator (3.7).

Since Algorithm 4 involves optimising in one coordinate group only, it need not produce a stationary point even if initialised at one. Algorithm 5 thus combines local search with coordinate descent. The combined algorithm first produces a candidate solution using coordinate descent and then follows up with local search. This scheme is iterated until the solution cannot be improved. Compared with coordinate descent alone, coordinate descent with local search can yield significantly lower objective values in high-correlation scenarios. Algorithm 5 is guaranteed to converge because Algorithm 4

Algorithm 5: Coordinate descent with local search

input : $\hat{\boldsymbol{\nu}}^{(0)} \in \mathbb{R}^{\sum_{k=1}^{g} p_k}$
for $m = 1, 2, \ldots$ **do**
  $\boldsymbol{\nu}^{(m)} \leftarrow$ output of Algorithm 3 initialised with $\hat{\boldsymbol{\nu}}^{(m-1)}$
  $\hat{\boldsymbol{\nu}}^{(m)} \leftarrow$ output of Algorithm 4 initialised with $\boldsymbol{\nu}^{(m)}$
  if $\boldsymbol{\nu}^{(m)} = \hat{\boldsymbol{\nu}}^{(m)}$ **then break**
end
return $\boldsymbol{\nu}^{(m)}$

never increases the objective value (by construction), Algorithm 3 is convergent, and the objective function is bounded below.

3.2.4 Regularisation sequence

To ensure larger groups are not unfairly penalised more strongly than smaller groups, the parameters λ_{0k} and λ_{1k} are configured to reflect the group size p_k. Suitable default choices are $\lambda_{0k} = p_k \lambda_0$ and $\lambda_{1k} = \sqrt{p}_k \lambda_1$, where λ_0 and λ_1 are nonnegative. For fixed λ_1, we take a sequence $\{\lambda_0^{(t)}\}_{t=1}^{T}$ such that $\lambda_0^{(1)}$ yields $\hat{\boldsymbol{\nu}} = \mathbf{0}$, and sequentially warm start the algorithms. That is, the solution for $\lambda_0^{(t+1)}$ is obtained by using the solution from $\lambda_0^{(t)}$ as an initialisation point. The sequence $\{\lambda_0^{(t)}\}_{t=1}^{T}$ is computed in such a way that the active set of groups corresponding to $\lambda_0^{(t+1)}$ is always different to that corresponding to $\lambda_0^{(t)}$. Proposition 3 presents the details of this method, extending an idea of Hazimeh and Mazumder (2020) for best subset.

Proposition 3. *Suppose that* $\hat{\boldsymbol{\nu}}^{(t)}$ *is the result of running Algorithm 3 with* $\lambda_0 = \lambda_0^{(t)}$. *Let* $\mathcal{A}^{(t)}$ *be the active set of groups. Then running Algorithm 3 initialised to* $\hat{\boldsymbol{\nu}}^{(t)}$ *and using* $\lambda_0 = \lambda_0^{(t+1)}$ *where*

$$\lambda_0^{(t+1)} = \alpha \cdot \max_{k \notin \mathcal{A}^{(t)}} \left(\frac{\left(\|\nabla L(\hat{\boldsymbol{\nu}}^{(t)})\| - \lambda_{1k} \right)_+^2}{2 p_k \bar{c}_k} \right)$$

produces a solution $\hat{\boldsymbol{\nu}}^{(t+1)}$ *such that* $\hat{\boldsymbol{\nu}}^{(t+1)} \neq \hat{\boldsymbol{\nu}}^{(t)}$ *for any* $\alpha \in [0, 1)$.

3.3 Error bounds

This section presents a finite-sample analysis of the proposed estimators. In particular, we state probabilistic upper bounds for the error of estimating the underlying regression function. These bounds accommodate overlapping groups and model misspecification. The role of structure and shrinkage is discussed, and comparisons are made with known bounds for other estimators.

3.3.1 Setup

The data is assumed to be generated according to the regression model

$$y_i = f^0(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $f^0 : \mathbb{R}^p \to \mathbb{R}$ is a regression function, $\mathbf{x}_i \in \mathbb{R}^p$ are fixed predictors, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is iid stochastic noise. This flexible specification encompasses the semiparametric model $f^0(\mathbf{x}) = \sum_{j=1}^g f_j(x_j)$ (with f_j zero, linear, or nonlinear) that is the focus of our empirical studies, and the linear model $f^0(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}^0$. Let $\mathbf{f}^0 := (f^0(\mathbf{x}_1), \dots, f^0(\mathbf{x}_n))^\top$ be the vector of function evaluations at the sample points. The goal of this section is to place probabilistic upper bounds on $\|\mathbf{f}^0 - \hat{\mathbf{f}}\|^2/n$, the estimation error of $\hat{\mathbf{f}} := \mathbf{X}\hat{\boldsymbol{\beta}}$.

The objects of our analysis are the group subset estimators (3.2). We allow the predictor groups $\mathcal{G}_1, \dots, \mathcal{G}_g$ to overlap. To facilitate comparisons against existing results, we constrain the number of nonzero groups rather than penalise them. To this end, let $\mathcal{V}(s)$ be the set of all $\bar{\boldsymbol{\nu}}$ such that at most s groups are nonzero:[2]

$$\mathcal{V}(s) := \left\{ \bar{\boldsymbol{\nu}} \in \mathcal{V} : \sum_{k=1}^g 1(\|\bar{\boldsymbol{\nu}}_k\| \neq 0) \leq s \right\}.$$

We consider the regular group subset estimator:

$$\min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p,\, \bar{\boldsymbol{\nu}} \in \mathcal{V}(s) \\ \boldsymbol{\beta} = \sum_{k=1}^g \bar{\boldsymbol{\nu}}_k}} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{3.10}$$

and the shrinkage estimator:

$$\min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p,\, \bar{\boldsymbol{\nu}} \in \mathcal{V}(s) \\ \boldsymbol{\beta} = \sum_{k=1}^g \bar{\boldsymbol{\nu}}_k}} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2 \sum_{k=1}^g \lambda_k \|\bar{\boldsymbol{\nu}}_k\|. \tag{3.11}$$

The results derived below apply to global minimisers of these nonconvex problems. The algorithms of the preceding section cannot guarantee such minimisers in general.[3] If global optimality is of foremost concern, the output of our algorithms can be used to initialise a mixed-integer optimiser which can guarantee a global solution at additional computational expense.

3.3.2 Bound for group subset selection

We begin with Theorem 3, which characterises an upper bound for group subset with no shrinkage. The notation $p_{\max} := \max_k p_k$ represents the maximal group size. As is customary, we absorb numerical constants into the term $C > 0$.

Theorem 3. *Let $\delta \in (0, 1]$ and $\alpha \in (0, 1)$. Then, for some numerical constant $C > 0$, the group subset estimator (3.10) satisfies*

$$\frac{1}{n}\|\mathbf{f}^0 - \hat{\mathbf{f}}\|^2 \leq \min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p,\, \bar{\boldsymbol{\nu}} \in \mathcal{V}(s) \\ \boldsymbol{\beta} = \sum_{k=1}^g \bar{\boldsymbol{\nu}}_k}} \frac{1+\alpha}{(1-\alpha)n} \|\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{C\sigma^2}{\alpha(1-\alpha)n} \left[s p_{\max} + s \log\left(\frac{g}{s}\right) + \log(\delta^{-1}) \right] \tag{3.12}$$

with probability at least $1 - \delta$.

[2] For all values of the group subset penalty parameter λ_0, there exists a constraint parameter s which yields an identical solution.

[3] In recent work, Guo et al. (2021) show that statistical properties of best subset remain valid when the attained minimum is within a certain neighbourhood of the global minimum. We expect their analysis extends to structured settings.

The first term on the right-hand side of (3.12) is the error incurred by the oracle in approximating $\mathbf{f}^0$ as $\mathbf{X}\boldsymbol{\beta}$. In general, this error is unavoidable in finite-dimensional settings. The three terms inside the brackets have the following interpretations. The first term is the cost of estimating $\boldsymbol{\beta}$; with s active groups, there are at most $s \times p_{\max}$ parameters to estimate. The second term is the price of selection; it follows from an upper bound on the total number of group subsets. The third term controls the trade-off between the tightness of the bound and the probability it is satisfied. Finally, the scalar α appears in the bound due to the proof technique (as in, e.g., Rigollet 2015). When $\mathbf{f}^0 = \mathbf{X}\boldsymbol{\beta}^0$, α need not appear. Hazimeh et al. (2023) obtain a similar bound for $\mathbf{f}^0 = \mathbf{X}\boldsymbol{\beta}^0$ in the case of equisized nonoverlapping groups. In the special case that all groups are singletons, (3.12) matches the well-known bound for best subset (Raskutti et al. 2011).

Theorem 3 confirms that group subset is preferable to best subset in structured settings. Consider the following example. Suppose we have g groups each of size p_0 so that the total number of predictors is $p = g \times p_0$. It follows for group sparsity level s that the ungrouped selection problem involves choosing $s \times p_0$ predictors. Accordingly, the ungrouped bound scales as $sp_0 + sp_0 \log(p/(sp_0)) = sp_0 + sp_0 \log(g/s)$. On the other hand, the grouped bound scales as $sp_0 + s\log(g/s)$, i.e., it improves by a factor p_0 of the logarithm term.

3.3.3 Bounds for group subset selection with shrinkage

We now establish bounds for group subset with shrinkage. The results are analogous to those established in Mazumder et al. (2023) for best subset with shrinkage. Two results are given, a bound where the error decays as $1/\sqrt{n}$, and another where the error decays as $1/n$. Adopting standard terminology (e.g., Hastie et al. 2015), the former bound is referred to as a 'slow rate' and the latter bound as a 'fast rate'.

The slow rate is presented in Theorem 4.

Theorem 4. *Let $\delta \in (0, 1]$. Let γ_k be the maximal eigenvalue of the matrix $\mathbf{X}_k^\top \mathbf{X}_k/n$ and*

$$\lambda_k \geq \frac{\sqrt{\gamma_k}\sigma}{\sqrt{n}} \sqrt{p_k + 2\sqrt{p_k \log(g) + p_k \log(\delta^{-1})} + 2\log(g) + 2\log(\delta^{-1})}, \quad k = 1, \ldots, g.$$

Then the group subset estimator (3.11) satisfies

$$\frac{1}{n}\|\mathbf{f}^0 - \hat{\mathbf{f}}\|^2 \leq \min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p,\, \bar{\boldsymbol{\nu}} \in \mathcal{V}(s) \\ \boldsymbol{\beta} = \sum_{k=1}^g \bar{\boldsymbol{\nu}}_k}} \frac{1}{n}\|\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}\|^2 + 4\sum_{k=1}^g \lambda_k \|\bar{\boldsymbol{\nu}}_k\| \tag{3.13}$$

with probability at least $1 - \delta$.

In the case of no overlap, Theorem 4 demonstrates that the shrinkage estimator satisfies the same slow rate as group lasso (Lounici et al. 2011, Theorem 3.1). An identical expression to Lounici et al. (2011) for λ_k can be stated here using a more intricate chi-squared tail bound in the proof. In the case of overlap, the same slow rate can be obtained for group lasso from Percival (2012, Lemma 4).

The following assumption is required to establish the fast rate.

Assumption 1. *Let $s < \min(n/p_{\max}, g)/2$. Then there exists a $\phi(2s) > 0$ such that*

$$\min_{\substack{\boldsymbol{\theta} \in \mathbb{R}^p,\, \bar{\boldsymbol{\nu}} \in \mathcal{V}(2s) \\ \boldsymbol{\theta} = \sum_{k=1}^g \bar{\boldsymbol{\nu}}_k \neq \mathbf{0}}} \frac{\|\mathbf{X}\boldsymbol{\theta}\|}{\sqrt{n}\sum_{k=1}^g \|\bar{\boldsymbol{\nu}}_k\|} \geq \phi(2s).$$

Assumption 1 is satisfied provided no collection of $2s$ groups have linearly dependent columns in $\mathbf{X}$. This condition is a weaker version of the restricted eigenvalue condition used in Lounici et al. (2011) and Percival (2012) for group lasso, which (loosely speaking) places additional restrictions on the correlations of the columns in $\mathbf{X}$.

The fast rate is presented in Theorem 5. The notation $\lambda_{\max} := \max_k \lambda_k$ represents the maximal shrinkage parameter.

Theorem 5. *Let Assumption 1 hold. Let $\delta \in (0, 1]$ and $\alpha \in (0, 1)$. Let $\lambda_1, \dots, \lambda_g \geq 0$. Then, for some numerical constant $C > 0$, the group subset estimator (3.11) satisfies*

$$\frac{1}{n}\|\mathbf{f}^0 - \hat{\mathbf{f}}\|^2 \leq \min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p, \tilde{\boldsymbol{\nu}} \in \mathcal{V}(s) \\ \boldsymbol{\beta} = \sum_{k=1}^g \tilde{\boldsymbol{\nu}}_k}} \frac{1+\alpha}{(1-\alpha)n}\|\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{C\sigma^2}{\alpha(1-\alpha)n}\left[sp_{\max} + s\log\left(\frac{g}{s}\right) + \log(\delta^{-1})\right] + \frac{C\lambda_{\max}^2}{\alpha(1-\alpha)\phi(2s)^2} \tag{3.14}$$

with probability at least $1 - \delta$.

Theorem 5 establishes that the shrinkage estimator achieves the bound of the regular estimator up to an additional term that depends on $\lambda_{\max}$ and $\phi(2s)$. By setting the shrinkage parameters to zero, the dependence on these terms vanishes, and the bounds are identical.

Theorems 4 and 5 together show that group subset with shrinkage does no worse than group lasso or group subset. This property is helpful because group lasso tends to outperform when the noise is high or the sample size is small, while group subset tends to outperform in the opposite situation. This empirical observation is consistent with the above bounds since the slow rate (3.13) depends on $\sigma/\sqrt{n}$ while the fast rate (3.14) depends on σ^2/n. Hence, (3.13) is typically the tighter of the two bounds for large σ or small n.

3.4 Simulations

This section investigates the statistical and computational performance of the proposed estimators for sparse semiparametric modelling on synthetic data. They are compared against group lasso, group SCAD, and group MCP, the latter two estimators being group versions of the smoothly clipped absolute deviation penalty (Fan and Li 2001) and the minimax concave penalty (Zhang 2010). The group subset estimators are fit using `grpsel`, our `R` implementation of the algorithms presented in Section 3.2. Group lasso, group SCAD, and group MCP are fit using the popular `R` package `grpreg` (Breheny and Huang 2015).

3.4.1 Tuning parameters and implementation

The range of tuning parameters for each estimator is:

- Group subset: a grid of λ_0 chosen adaptively using the method of Proposition 3, where the first λ_0 sets all coefficients to zero;
- Group subset+lasso: a grid of λ_1 containing logarithmically spaced points between $\lambda_1^{\max}$ and $\lambda_1^{\min} = 10^{-4}\lambda_1^{\max}$, where $\lambda_1^{\max}$ is the smallest value that sets all coefficients to zero, and for each value of λ_1, a grid of λ_0 chosen as above;

- Group lasso: a grid of λ containing logarithmically spaced points between $\lambda^{\max}$ and $\lambda^{\min} = 10^{-4}\lambda^{\max}$, where $\lambda^{\max}$ is the smallest value that sets all coefficients to zero;
- Group SCAD: the same grid of λ as above, and for each value of λ, a grid of the nonconvexity parameter γ containing logarithmically spaced points between $\gamma^{\max} = 100$ and $\gamma^{\min} = 2 + 10^{-4}$; and
- Group MCP: the same grid of λ as above, and for each value of λ, a grid of the nonconvexity parameter γ containing logarithmically spaced points between $\gamma^{\max} = 100$ and $\gamma^{\min} = 1 + 10^{-4}$.

Grids of 100 points are used for the primary tuning parameters (λ_0, λ) and grids of 30 points for the secondary tuning parameters (λ_1, γ).

Unlike `grpsel`, `grpreg` does not have native support for overlapping groups. To this end, we use the approach proposed in Jacob et al. (2009) of expanding the predictor matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ by replicating a predictor each time it appears in a new group to get $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times \sum_{k=1}^{g} p_k} = (\mathbf{X}_1, \ldots, \mathbf{X}_g)$. `grpreg` is then run on the expanded predictor matrix $\tilde{\mathbf{X}}$ with nonoverlapping groups. The downside of this approach compared with that of `grpsel` is the additional memory required to store $\tilde{\mathbf{X}}$, which can be much wider than $\mathbf{X}$.

3.4.2 Sparse semiparametric modelling

Recall that in a sparse semiparametric model, the response y is modelled via a sum of univariate functions $\sum_j f_j(x_j)$, where f_j can be zero, linear, or nonlinear. We follow the approach of Chouldechova and Hastie (2015) in using overlapping groups and regression splines to fit this model. Briefly, for every x_j, an orthogonal spline is computed and two groups are formed: a linear group containing the first term of the spline (assumed equal to x_j) and a nonlinear group containing all terms of the spline. Due to the linear and nonlinear groups overlapping, the fit $\hat{f}_j$ is nonlinear whenever the nonlinear group is selected regardless of whether the linear group is also selected. The group penalty parameters are scaled to control the trade-off between fitting f_j as linear or nonlinear. For group subset penalty parameter λ, we set $\lambda_k = \lambda$ for k a linear group and $\lambda_k = 2\lambda$ for k a nonlinear group. For group lasso penalty parameter λ, we set $\lambda_k = \sqrt{2}\lambda$ for nonlinear groups to achieve an equivalent penalisation.

3.4.3 Simulation design

We study regression and classification. For regression, the response is generated according to

$$y_i = f^0(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

while, for classification, it is generated as

$$y_i = \begin{cases} 0 & \text{if } f^0(\mathbf{x}_i) + \varepsilon_i < 0 \\ 1 & \text{if } f^0(\mathbf{x}_i) + \varepsilon_i \geq 0, \end{cases} \quad i = 1, \ldots, n.$$

In both cases, $f^0(\mathbf{x}) = \sum_{j=1}^{g} f_j^0(x_j)$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The covariates $\mathbf{x}_i$ are treated fixed and constructed by (1) drawing samples iid from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, (2) applying the standard normal distribution function to produce uniformly distributed variables that conform to $\mathbf{\Sigma}$ (see Falk 1999), and (3) min-max scaling to the interval $[-1, 1]$. The correlation matrix $\mathbf{\Sigma}$ is defined elementwise as $\Sigma_{i,j} = \rho^{|i-j|}$. The number of covariates is 2,500. For regression,

50 of the these covariates are selected at random to be nonzero: 40 relate to the response linearly as $f(x) = x$ and 10 relate nonlinearly as $f(x) = \cos(\pi x)$, $f(x) = \sin(\pi x)$, or $f(x) = \exp(10x)$. For classification, support recovery is more difficult, so for that task 10 covariates are linear and 5 are nonlinear. The function evaluations are scaled to mean zero and variance one so that all functions are on the same footing. Each of the 2,500 covariates are expanded using a cubic regression spline containing three knots at equispaced quantiles. Four terms are in each spline so that the number of predictors $p = 10,000$. The number of groups $g = 5,000$, consisting of 2,500 linear groups and 2,500 nonlinear groups. The sample size $n = 1,000$ is fixed and the noise parameter σ is varied to alter the signal-to-noise ratio (SNR).

3.4.4 Statistical performance

For regression, we measure out-of-sample test loss by the mean square loss on a testing set and report it relative to that of the best-performing estimator:

$$\text{Relative test loss} := \frac{\text{Mean square loss} - \text{Mean square loss}^\star}{\text{Mean square loss}^\star},$$

where $\star$ indicates the minimal mean square (test) loss among the five estimators considered. The best possible value of this metric is zero. For classification, we report the same metric but measured in terms of mean logistic loss. As a measure of sparsity, we report the number of fitted functions that are nonzero:

$$\text{Sparsity} := \sum_{j=1}^{g} \hat{f}_j \not\equiv 0.$$

Finally, as a measure of support recovery, we report the micro F1-score for the classification of linear and nonlinear functions:

$$\text{Support recovery} := \frac{2 \cdot \text{True positives}}{2 \cdot \text{True positives} + \text{False positives} + \text{False negatives}}.$$

The best possible value of this metric is one and the null value is zero. These metrics are all evaluated using tuning parameters that minimise mean square loss or mean logistic loss over a separate validation set of size n.

The metrics under consideration are aggregated over 30 simulations. The regression results are reported in Figure 3.1 and the classification results in Figure 3.2. The solid points are averages and the error bars are standard errors.

Consider first the regression results. In line with our theory, group subset exhibits excellent performance when the signal is strong but fares poorly when the signal is weak. Group lasso behaves contrarily, performing capably at low SNRs but poorly at high SNRs. The transition between the two estimators in terms of relative test loss occurs at SNR ≈ 2 in the moderate-correlation scenario ($\rho = 0.5$) and later at SNR ≈ 3 in the high-correlation scenario ($\rho = 0.9$). Group subset+lasso achieves the best of both worlds. It improves the performance of group subset when the signal is weak and, by tapering off shrinkage, eventually behaves like group subset when the signal is strong. Group SCAD and group MCP also try to unwind shrinkage via their nonconvexity parameters. Even so, there remains a gap between these estimators and group subset+lasso. The latter converges earlier to the right sparsity level and the correct model. The gap is most stark in the high-correlation scenario, where their support recovery is roughly half that of the group subset estimators at SNR $= 10$.

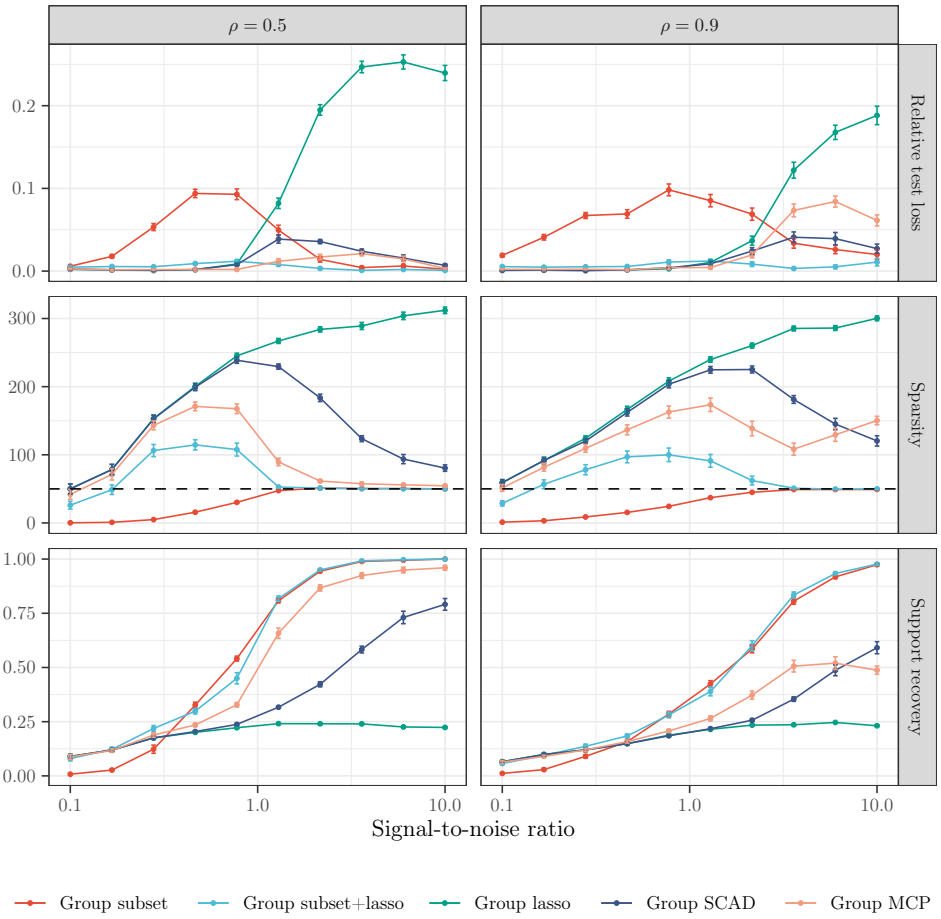

Figure 3.1: Comparisons of estimators for sparse semiparametric regression. Metrics are aggregated over 30 synthetic datasets generated with $n = 1,000$, $p = 10,000$, and $g = 5,000$. Solid points represent averages and error bars denote (one) standard errors. Dashed lines indicate the true number of nonzero functions.

Turning our attention to the classification results, we again see that group lasso maintains an edge in prediction over group subset when the signal is weak and vice-versa when the signal is strong. As before, the gap between group lasso and group subset at low SNRs is closed once shrinkage is introduced. The group subset estimators also continue to have a clear upper hand in support recovery. Closer inspection reveals this upper hand is because they make fewer false positive selections than the competing estimators, similar to the regression setting. There is, however, one interesting difference to the regression setting: group subset+lasso has an advantage over group subset when the SNR is high. In this regime, it has noticeably lower relative test loss and marginally better support recovery. This result corresponds to a well-known phenomenon where the maximum likelihood estimator diverges as the classes become separable (see Hastie et al. 2015). Shrinkage has the desirable side-effect of preventing this divergence.

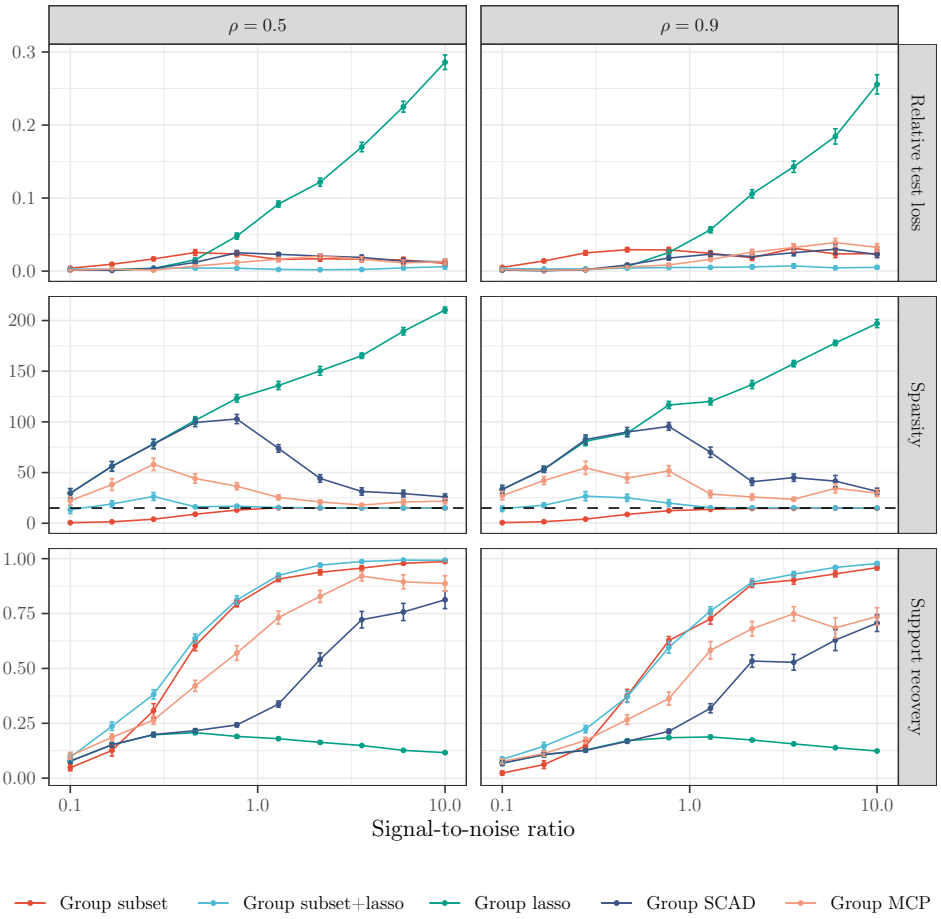

Figure 3.2: Comparisons of estimators for sparse semiparametric classification. Metrics are aggregated over 30 synthetic datasets generated with $n = 1,000$, $p = 10,000$, and $g = 5,000$. Solid points represent averages and error bars denote (one) standard errors. Dashed lines indicate the true number of nonzero functions.

3.4.5 Computational performance

We now compare the computational performance of `grpsel` against `grpreg` for regression. Our estimators—group subset and group subset+lasso—and those of `grpreg`—group lasso, group SCAD, and group MCP—each solve different optimisation problems, so it does not make sense to ask whether one implementation is faster than another for the same problem. Rather, the purpose of these comparisons is to provide indications of run time and computational complexity for alternative approaches to structured sparsity. Both `grpsel` and `grpreg` are set with a convergence tolerance of 10^{-4}. All run times and iteration counts are measured with reference to the coordinate descent algorithms of each package over a grid of the primary tuning parameter. Where there is a secondary tuning parameter, the figures reported are averaged over the secondary parameter, e.g., the total time taken to evaluate a 100×30 grid of parameters divided by 30. The simulation design is as before, but we now fix the SNR and vary the number of covariates to measure scalability.

The results as aggregated over 30 synthetic datasets are reported in Figure 3.3. The

vertical bars are averages and the error bars are standard errors. For $p = 10,000$, `grpsel`

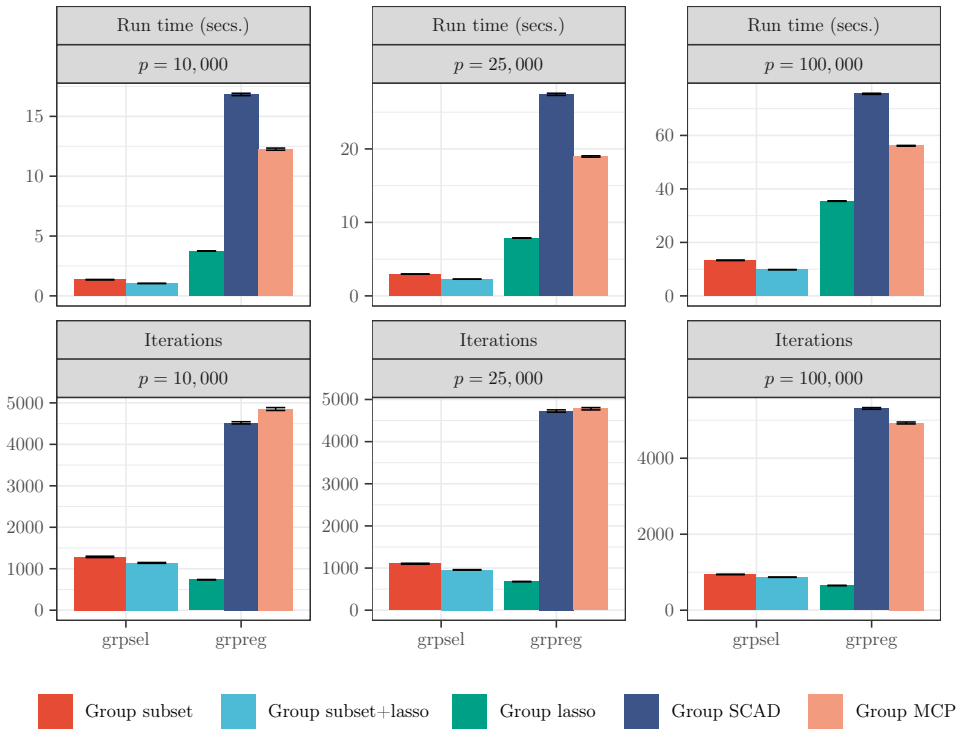

Figure 3.3: Comparisons of packages and estimators. Metrics are aggregated over 30 synthetic datasets generated with SNR $= 1$, $\rho = 0.5$, and $n = 1,000$. Vertical bars represent averages and error bars denote (one) standard errors.

can compute an entire solution path in one to two seconds. Group subset+lasso achieves marginally lower run times than group subset, requiring slightly fewer iterations to converge thanks to the additional regularisation from shrinkage. For $p = 25,000$ the run times from `grpsel` are still less than five seconds to fit a path, while for $p = 100,000$ the times are around 10 seconds. `grpreg` is also impressive in these scenarios, though relative to `grpsel` it is slower. Group lasso converges in the fewest iterations, followed by group subset+lasso. Group SCAD and group MCP always take several thousand iterations to converge.

3.5 Data analyses

This section studies two contemporary problems: modelling foot traffic in major supermarkets and modelling recessions in the business cycle. Both problems are characterised by the availability of many candidate predictors and the possibility for misspecification of linear models. These characteristics motivate consideration of sparse semiparametric models.

3.5.1 Supermarket foot traffic

The first dataset contains anonymised data on foot traffic and sales volumes for a major Chinese supermarket (see Wang 2009).[4] The task is to model foot traffic using the sales volumes of different products. To facilitate managerial decision-making, the fitted model should identify a subset of products that well-predict foot traffic (i.e., it should be sparse).

The sample contains $n = 464$ days. We randomly hold out 10% of the data as a testing set and use the remaining data as a training set. Sales volumes are available for 6,398 products. To fit sparse semiparametric models, the approach described in Section 3.4.2 is applied. A four-term cubic regression spline is used for each product, resulting in $p = 25,592$ predictors and $g = 12,796$ groups (6,398 linear groups plus 6,398 nonlinear groups). As a measure of predictive accuracy, we report mean square loss on the testing set. This metric is reported in absolute terms, not relative to a benchmark. We also report the number of fitted functions that are nonzero. As benchmarks, we include random forest and lasso, which respectively produce dense nonparametric models and sparse linear models. In addition, the (unconditional) mean is evaluated as a predictive method to assess the value added by the predictors. Ten-fold cross-validation is used to choose tuning parameters.

The metrics under consideration are aggregated over 30 training-testing set splits and reported in Table 3.1. Averages are reported with standard errors in parentheses. Group

		Sparsity		
	Mean square loss	Total	Linear	Nonlinear
Group subset+lasso	0.650 (0.026)	178.6 (2.3)	115.7 (1.2)	62.9 (2.0)
Group lasso	0.651 (0.025)	267.0 (12.9)	132.9 (2.3)	134.1 (11.0)
Group MCP	0.658 (0.026)	215.1 (6.1)	120.9 (1.2)	94.1 (5.5)
Lasso	0.712 (0.028)	165.2 (5.0)	165.2 (5.0)	-
Random forest	1.330 (0.057)	-	-	-
Mean	9.856 (0.367)	-	-	-

Table 3.1: Comparisons of methods for modelling supermarket foot traffic. Metrics are aggregated over 30 splits of the data into training and testing sets. Averages are reported next to (one) standard errors in parentheses.

subset+lasso used to fit sparse semiparametric models leads to the lowest mean square loss, though within statistical precision of group lasso and group MCP. Nonetheless, group subset+lasso has the edge of selecting fewer products than either of these estimators, which is advantageous for decision-making. Lasso yields models that are slightly more sparse but at the same time substantially less predictive, signifying linearity might be too restrictive. Random forest is markedly worse than any sparse method. It appears only a small fraction of products explain foot traffic, around 2–4%.

Since this dataset is a time series, the noise may exhibit temporal dependence. Such dependence could weaken predictive performance and invalidate our statistical theory. An avenue for future work is to extend the group subset estimators and the accompanying theory to accommodate dependence, perhaps by placing an autoregressive structure on the noise.

[4] Available at https://personal.psu.edu/ril4/DataScience.

3.5.2 Economic recessions

The second dataset contains monthly data on macroeconomic series for the United States (see McCracken and Ng 2016).[5] The dataset is augmented with the National Bureau of Economic Research recession indicator, a dummy variable that indicates whether a given month is a period of recession or expansion.[6] The task is to model the recession indicator using the macroeconomic series. Such models are useful for scenario analysis and for assessing economic conditions in the absence of low-frequency variables such as quarterly gross domestic product growth.

The sample contains $n = 746$ months, ending in October 2021. It includes the COVID-19 recession. We again randomly hold out 10% of the data as a testing set and use the remaining data as a training set. Because there are relatively few recessionary periods, a stratified split is applied so that the proportion of recessions in the testing and training sets are equal; see Kohavi (1995) for the advantages of stratified splits. The dataset has 127 macroeconomic series. We add to this dataset six lags of each series, leading to a total set of 889 series. Applying a four-term spline to each series yields $p = 3556$ predictors and $g = 1778$ groups (889 linear groups plus 889 nonlinear groups). To evaluate predictive accuracy, we report mean logistic loss on the testing set. The remaining metrics and methods are as before.

The results as aggregated over 30 training-testing set splits are reported in Table 3.2. The sparse semiparametric models from group subset+lasso predict recessionary periods

		Sparsity		
	Mean logistic loss	Total	Linear	Nonlinear
Group subset+lasso	0.968 (0.103)	44.0 (2.5)	16.2 (0.9)	27.8 (1.8)
Group lasso	1.063 (0.102)	57.5 (1.5)	30.7 (0.7)	26.8 (1.0)
Group MCP	1.141 (0.092)	31.5 (0.7)	18.7 (0.5)	12.8 (0.4)
Lasso	1.093 (0.107)	56.8 (1.1)	56.8 (1.1)	-
Random forest	1.294 (0.039)	-	-	-
Mean	3.703 (0.000)	-	-	-

Table 3.2: Comparisons of methods for modelling economic recessions. Metrics are aggregated over 30 splits of the data into training and testing sets. Averages are reported next to (one) standard errors in parentheses.

well. They perform better than those from group lasso and, at the same time, depend on fewer macroeconomic series. Group MCP is sparser still but also less predictive. Lasso also yields models worse than those from group lasso and group subset+lasso, highlighting the value in allowing for nonlinearity. As with the supermarket dataset, the dense models from random forest underperform relative to the sparse models from other methods. All methods improve on the mean, illustrating the predictive content of monthly series.

As before, there is a possibility for temporal dependence in the response here—a recessionary period is likely to follow a recessionary period. It is also well-known that there are extreme values in some macroeconomic series, particularly during the COVID-19 recession. Though we preprocessed the datasets to remove outlying data points using standard univariate trimming techniques (see Appendix B.3.1), a more principled

[5] The January 2022 vintage is used, available at https://research.stlouisfed.org/econ/mccracken. See Appendix B.3.1 for preprocessing steps.

[6] Available at https://fred.stlouisfed.org/series/USREC.

approach could consider a robust version of group subset. See, e.g., the robust variant of best subset recently proposed by Thompson (2022).

3.6 Concluding remarks

Despite a broad array of applications, structured sparsity via group subset selection is not well-studied, especially in high dimensions where it is computationally taxing. This chapter represents an effort to close the gap. Our optimisation framework consists of low complexity algorithms that come with convergence results. A theoretical analysis of the proposed estimators illuminates some of their finite-sample properties. The estimators behave favourably in simulation, exhibiting excellent support recovery when fitting sparse semiparametric models. In real-world modelling tasks, they improve on popular benchmarks.

Our implementation `grpsel` is available on the `R` repository `CRAN`.

Chapter 4

Familial inference

4.1 Introduction

Hypothesis testing is one of statistics' most important contributions to the scientific method. Testing helps advance diverse lines of inquiry, from evaluating the efficacy of experimental drugs to assessing the validity of psychological theories. Researchers working on these problems often characterise their questions as competing statements about a centre μ of one or more distributions. In the simplest one-sample setting, these statements take the form

$$\mathrm{H}_0 : \mu \in \mathcal{M}_0 \quad \text{vs.} \quad \mathrm{H}_1 : \mu \in \mathcal{M}_1,$$

where $\mathcal{M}_0$ and $\mathcal{M}_1$ are a partition of the support $\mathcal{M}$ of μ. There are myriads of classical tests for one- and two-sample hypotheses of centre. When μ is the mean, the most well-known of these is the t test (Student 1908), and its extension to independent samples from populations with differing variances (Welch 1947). When μ is the median, the sign test (Fisher 1925) is available, as is the median test for independent samples (Mood 1950). The signed-rank test, or rank-sum test for independent samples, are also tests of medians under certain assumptions (Wilcoxon 1945; Mann and Whitney 1947).

The possibility to test different centres such as the mean and median raises the question of what qualifies as a 'centre'. We posit that a centre of a random variable X should satisfy at least two criteria: (1) a reflection of X about the centre should preserve the centre, and (2) a shift in X by a constant should move the centre by that same constant. This definition is purposefully broad to accommodate the many notions of centre used throughout statistics. The mean and median trivially satisfy these criteria, as do other popular notions such as the mode, symmetric trimmed mean, and symmetric Winsorised mean. Quantiles other than the median, and by extension order statistics such as the minimum and maximum, are not centres under these criteria as they are not preserved by reflection in general. Still, the fact that there are many possibilities for centre can complicate hypothesis testing in science.

In certain applied areas (e.g., psychology and medicine), *scientific hypotheses* are often silent about a specific centre and instead tend to be statistically vague, e.g., 'treatment A is more efficacious than treatment B'. This ambiguity makes translation to *statistical hypotheses* inherently subjective and can leave researchers questioning which centre to use. See Blakely and Kawachi (2001), Ben-Aharon et al. (2019), and Rousselet and Wilcox (2020) for discussions of this issue in epidemiology, medicine, and psychology. Moreover, ambiguity about the correct (or best) centre leaves the possibility for a gap

between scientific theory and statistical practice that can lead to rejection of a true null, threatening the validity of findings. Sometimes H_0 can be rejected just by switching from one centre to another, say from the mean to the median. Choosing a specific centre to attain a certain result is both statistically invalid and scientifically unethical. In the face of replicability crises in various disciplines (see, e.g., Ioannidis 2005; Open Science Collaboration 2015; Christensen and Miguel 2018), the possibility for significant results of this sort is concerning. Transparent statistical tools are needed to instil confidence in scientific claims.

Motivated by the preceding discussion, this chapter proposes a new approach to hypothesis testing: familial inference. Unlike existing inferential methods, which test hypotheses about a single centre, methods for familial inference test hypotheses about a *family* of plausible centres, with the ultimate goal of strengthening any claims of significance. More specifically, consider a family of centres $\{\mu(\lambda) : \lambda \in \Lambda\}$ where λ indexes each member (centre). The familial testing problem is to decide which hypothesis concerning this family is correct:

$$H_0 : \mu(\lambda) \in \mathcal{M}_0 \text{ for some } \lambda \in \Lambda \quad \text{vs.} \quad H_1 : \mu(\lambda) \in \mathcal{M}_1 \text{ for all } \lambda \in \Lambda.$$

The familial null hypothesis states that at least one member (centre) of the family is contained in the null set $\mathcal{M}_0$. The alternative hypothesis is that no member is in $\mathcal{M}_0$. The problem of performing inference on $\mu(\lambda)$ is related to testing nonparametric regression curves (see, e.g., Hall and Hart 1990; Neumeyer and Dette 2003), though our work is distinct in motivation and approach.[1] This chapter studies the family of centres induced by the Huber loss function (Huber 1964). The Huber function is comprised piecewise of square and absolute loss, where λ controls the transition point. By sweeping λ between 0 and infinity, one obtains a family of centres that includes the mean and median as limit points. All members of this 'Huber family' satisfy our criteria for centre.

Familial inference is more sophisticated than inference for a single centre and requires new tools developed in this chapter. Our first methodological development is a Bayesian nonparametric procedure for one- and two-sample testing. The procedure is based on the limit of a Dirichlet process prior (Ferguson 1973), sometimes referred to as the Bayesian bootstrap (Rubin 1981). Bayesian tests have several advantages over frequentist tests, including that they measure the probability of H_0. Frequentist approaches only deliver p-values that are at best a proxy for this probability. We refer the reader to Kruschke (2013) and Benavoli et al. (2017) for discussions on the merits of Bayesian testing. Besides the advantages of a Bayesian approach, the nonparametric nature of our test ameliorates concern about model misspecification. Though numerous existing works address Bayesian nonparametric testing (Ma and Wong 2011; Benavoli et al. 2014; Huang and Ghosh 2014; Benavoli et al. 2015; Holmes et al. 2015; Filippi and Holmes 2017; Gutiérrez et al. 2019; Pereira et al. 2020), these treat hypotheses about single statistical parameters or entire distributions, distinct from the familial hypotheses treated in this chapter.

Our second methodological development is an algorithm for fitting the Huber family, necessary to implement the new test. The algorithm is a pathwise optimisation routine that exploits piecewise linearity of the Huber solution path to fit the family (containing infinitely many centres) in a single pass over the data. It has low computational complexity and terminates in at most $n - 1$ steps, where n is the sample size. We elucidate the connection between our algorithm and least angle regression (Efron et al. 2004; Rosset

[1] We are motivated from the perspective of scientific reproducibility and our approach involves new statistical and computational developments.

and Zhu 2007), popularly used for fitting the lasso regularisation path (Tibshirani 1996). The algorithms devised in this chapter are made available in the open-source `R` package `familial`, designed with a standard interface similar to that of existing tests in the `stats` package. Methods for visualising the posterior family via functional boxplots (Sun and Genton 2011) are provided. `familial` is publicly available on the `R` repository `CRAN`.

To illustrate our proposal, we consider data from a study of mammalian sleep patterns in Savage and West (2007). The data contains sleep times for $n = 83$ species of mammals. A histogram of the ratio of sleeping hours to waking hours is plotted in Figure 4.1. The data are heavily right-skewed, suggesting the mean and median are probably far separated. Suppose we ask whether mammals tend to spend as much time sleeping as they do awake,

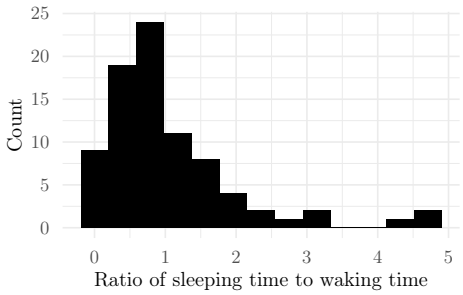

Figure 4.1: Histogram of the mammalian sleep data.

i.e., whether $\mu = 1$. A t test that the mean is one yields a p-value of 0.698. A bootstrap test that the mean is one, conducted as a robustness check, produces only a marginally smaller p-value of 0.668. A sign test that the median is one gives a p-value of 0.028. At a conventional 0.05 significance level, these tests do not yield the same answer to our scientific question. This inconsistency raises the question of how exactly to proceed in the absence of a guiding scientific theory.

Using our procedure, we estimate the posterior Huber family via 1,000 Bayesian bootstraps, summarised in Figure 4.2 by a functional boxplot. As the Huber parameter

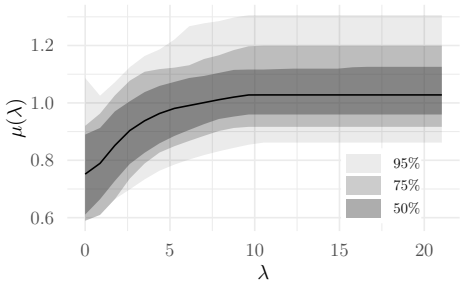

Figure 4.2: Functional boxplot of the posterior density of the Huber family for the mammalian sleep data. Shading indicates different central regions of the posterior.

$\lambda \to \infty$, the 50% central region of the posterior encloses the null value (recall the mean is attained in the limit). By querying the posterior, we find a probability of 0.633 that at least one centre in the family equals one.[2] Under zero-one loss configured analogously to using a 0.05 frequentist significance level (detailed later), the familial test finds insufficient

[2]See equation (4.1) for the formula for computing the null probability from the posterior.

evidence to reject the null in favour of the alternative. Because no specific choice was made about the centre, the problem of choosing between conflicting tests does not arise. Most importantly, we do not arrive at a result that would hold only under a certain centre.

4.1.1 Organisation

This chapter is structured as follows. Section 4.2 describes the Bayesian nonparametric testing procedure. Section 4.3 details the pathwise algorithm for fitting the Huber family. Section 4.4 addresses the two-sample problem. Section 4.5 discusses the relation between familial testing and intersection-union testing. Section 4.6 presents results from numerical simulations. Section 4.7 illustrates the new test in two real-world case studies. Section 4.8 closes the chapter. Proofs are available in the appendices.

4.2 Bayesian nonparametric test

This section presents our Bayesian nonparametric procedure for familial testing.

4.2.1 Inference problem

Let $X_1, \ldots, X_n$ be an iid sample according to a distribution P_0. Our goal is to carry out inference on the set $\{\mu_0(\lambda) : \lambda \in \Lambda\}$, where

$$\mu_0(\lambda) := \arg\min_{\mu \in \mathcal{M}} \mathrm{E}\left[\ell_\lambda\left(\frac{X-\mu}{\sigma}\right)\right] = \arg\min_{\mu \in \mathcal{M}} \int \ell_\lambda\left(\frac{x-\mu}{\sigma}\right) dP_0(x).$$

Here, $\ell_\lambda : \mathbb{R} \to \mathbb{R}_+$ is a loss function controlled by the parameter λ. The constant $\sigma > 0$ is necessary in certain loss functions to make λ invariant to the spread of X.[3] The population centre $\mu_0(\lambda)$ minimises the expectation of the loss configured by λ under P_0. To maintain generality throughout this section, we do not specify a particular loss function. However, to give a concrete example that will be the focus of subsequent sections, one may consider the Huber function

$$\ell_\lambda(z) = \begin{cases} \frac{1}{2}z^2 & \text{if } |z| < \lambda \\ \lambda|z| - \frac{1}{2}\lambda^2 & \text{if } |z| \geq \lambda. \end{cases}$$

The support of λ is $\Lambda = (0, \infty)$. The mean of P_0 is the limiting solution as $\lambda \to \infty$. The median is the limiting solution in the other direction. The continuum of centres therebetween comprises the Huber family. Though our focus is the full Huber family corresponding to $\Lambda = (0, \infty)$, the approach we propose can accommodate the restriction to any subset of the family given by $\Lambda = [a, b]$ for $0 < a < b < \infty$.

If the true generative model P_0 were known, we would immediately have access to the family $\{\mu_0(\lambda) : \lambda \in \Lambda\}$. Of course, this is not the case in practice—P_0 is unknown. The traditional parametric Bayesian approach to this problem proceeds by means of a prior on parameters for a class of models for P. A valid criticism of this approach is the implicit assumption that P_0 is contained in the model class. Misspecified models can lead to false conclusions, which is troubling in the context of hypothesis testing. To this end, the Bayesian nonparametric approach is an appealing alternative. Rather than placing

[3] Invariance of λ to spread is relevant for testing independent samples, addressed later.

a prior on the parameters governing a distribution for P, one places a prior directly on the distribution itself. The Dirichlet process—a probability distribution on the space of probability distributions—is a natural candidate for this task. Since Dirichlet processes have support on a large class of distributions, they are a popular prior in Bayesian nonparametrics. The reader is referred to MacEachern (2016) for a recent and accessible overview of their properties.

4.2.2 Bayesian bootstrap

We denote by $\mathrm{DP}(cP_\pi)$ a Dirichlet process with base distribution P_π and concentration parameter $c > 0$. The concentration parameter is used to impart confidence in P_π. With a Dirichlet process as a prior on P, our Bayesian model is

$$X_1, \ldots, X_n \,|\, P \overset{\text{iid}}{\sim} P, \quad P \sim \mathrm{DP}(cP_\pi).$$

Ferguson (1973) shows the posterior corresponding to this model is also a Dirichlet process:

$$P \,|\, X_1 = x_1, \ldots, X_n = x_n \sim \mathrm{DP}\left(cP_\pi + \sum_{i=1}^{n} \delta_{x_i}\right),$$

where δ_{x_i} is the Dirac measure at x_i. The Dirichlet process is a conjugate prior for iid sampling under P, and the posterior is the base distribution P_π with added point masses at the sample realisations $x_1, \ldots, x_n$. A base distribution P_π must be chosen to operationalise this model. If one wishes to minimise the impact of the choice of P_π, it is sensible to consider the limiting case where the concentration parameter $c \to 0$, which leads to the posterior

$$P \,|\, X_1 = x_1, \ldots, X_n = x_n \sim \mathrm{DP}\left(\sum_{i=1}^{n} \delta_{x_i}\right).$$

Gasparini (1995) shows that this posterior exactly matches the Bayesian bootstrap, proposed by Rubin (1981) as the Bayesian analog of the frequentist bootstrap (Efron 1979). MacEachern (1993) also establishes a unique connection of this posterior to the empirical distribution of the data. The Bayesian bootstrap places support only on the observed data and is equivalent to

$$P(\cdot) = \sum_{i=1}^{n} w_i \delta_{x_i}(\cdot), \quad (w_1, \ldots, w_n) \sim \mathrm{Dirichlet}(1, \ldots, 1),$$

where $\mathrm{Dirichlet}(1, \ldots, 1)$ is the n-dimensional Dirichlet distribution with all concentration parameters equal to one. Sometimes this distribution is referred to as 'flat' or 'uniform'. The first- and second-order asymptotic properties of the Bayesian bootstrap are described in Lo (1987) and Weng (1989). As well as being theoretically well-understood, the Bayesian bootstrap admits scalable sampling algorithms that are trivially parallelisable, making posterior exploration highly tractable. See Fong et al. (2019), Lyddon et al. (2019), and Barrientos and Peña (2020) for recent applications of the Bayesian bootstrap to complex models and data. As with those applications, tractability is key here.

We now have a posterior for P, and consequently also a posterior on any summaries of P (see, e.g., Lee and MacEachern 2014), including those of interest—families of centres. To estimate the posterior for a given family we propose Algorithm 6. Simulating random

Algorithm 6: Bayesian bootstrap for familial inference

input : $(x_1, \ldots, x_n)$

for $b = 1, \ldots, B$ **do**

1 Sample $(w_1^{(b)}, \ldots, w_n^{(b)})$ from Dirichlet$(1, \ldots, 1)$

2 Compute $\mu^{(b)}(\lambda) = \arg\min_{\mu \in \mathcal{M}} \sum_{i=1}^{n} w_i^{(b)} \ell_\lambda([x_i - \mu]/\sigma^{(b)})$ for all $\lambda \in \Lambda$

end

output : $\{\mu^{(b)}(\lambda) : \lambda \in \Lambda\}_{b=1}^{B}$

numbers from Dirichlet$(1, \ldots, 1)$ in step one is straightforward: take n iid draws from an exponential distribution with rate parameter one and rescale these draws such that their sum is one. Solving the minimisation problem in step two for all $\lambda \in \Lambda$ is more complex, with the exact complexity depending on the loss function. In the next section, we present a numerical routine that addresses the case where the loss function is the Huber function. Since λ in the Huber function is sensitive to changes in spread, we configure $\sigma^{(b)}$ to be the median absolute deviation of the bootstrap sample (i.e., the weighted median absolute deviation with weights $w_1^{(b)}, \ldots, w_n^{(b)}$). The standard deviation of the bootstrap sample could also be used.

From the output of Algorithm 6, the posterior probabilities $p_{\mathrm{H}_0} := \mathrm{P}(\mathrm{H}_0 \,|\, x_1, \ldots, x_n)$ of $\mathrm{H}_0 : \mu \in \mathcal{M}_0$ and $p_{\mathrm{H}_1} := \mathrm{P}(\mathrm{H}_1 \,|\, x_1, \ldots, x_n)$ of $\mathrm{H}_1 : \mu \in \mathcal{M}_1$ are estimable as

$$\hat{p}_{\mathrm{H}_0} := \frac{1}{B} \sum_{b=1}^{B} 1(\exists\, \lambda \in \Lambda : \mu^{(b)}(\lambda) \in \mathcal{M}_0) \tag{4.1}$$

and

$$\hat{p}_{\mathrm{H}_1} := \frac{1}{B} \sum_{b=1}^{B} 1(\forall\, \lambda \in \Lambda : \mu^{(b)}(\lambda) \in \mathcal{M}_1).$$

Since H_0 and H_1 are mutually exclusive and collectively exhaustive, $p_{\mathrm{H}_0} + p_{\mathrm{H}_1} = 1$ and, for any B, $\hat{p}_{\mathrm{H}_0} + \hat{p}_{\mathrm{H}_1} = 1$.

4.2.3 Decision rule

To map the estimated posterior probabilities $\hat{p}_{\mathrm{H}_0}$ and $\hat{p}_{\mathrm{H}_1}$ to a decision, we assign a loss to each possible decision. Specifically, given the posterior probability vector $\hat{p} = (\hat{p}_{\mathrm{H}_0}, \hat{p}_{\mathrm{H}_1})^\top$, we make the decision giving lowest posterior expected loss $L\hat{p}$, where L is loss matrix with rows corresponding to the decision to accept H_0, accept H_1, or accept neither (an *indeterminate* decision). We use

$$L := \begin{pmatrix} l_{\mathrm{H}_0|\mathrm{H}_0} & l_{\mathrm{H}_0|\mathrm{H}_1} \\ l_{\mathrm{H}_1|\mathrm{H}_0} & l_{\mathrm{H}_1|\mathrm{H}_1} \\ l_{\mathrm{I}|\mathrm{H}_0} & l_{\mathrm{I}|\mathrm{H}_1} \end{pmatrix} \begin{matrix} \\ = \\ \\ \end{matrix} \begin{matrix} \mathrm{H}_0 \\ \mathrm{H}_1 \\ \mathrm{I} \end{matrix} \begin{matrix} \mathrm{H}_0 & \mathrm{H}_1 \\ \end{matrix} \begin{pmatrix} 0 & 20 \\ 20 & 0 \\ 1 & 1 \end{pmatrix}, \tag{4.2}$$

where $l_{\mathrm{H}_j|\mathrm{H}_k}$ denotes the loss incurred in accepting H_j when H_k is true for $j, k = 0, 1$, and where $l_{\mathrm{I}|\mathrm{H}_k}$ denotes the loss from an indeterminate decision for $k = 0, 1$. Under the above configuration of L, either H_0 or H_1 is accepted depending on whether $\hat{p}_{\mathrm{H}_0}$ or $\hat{p}_{\mathrm{H}_1}$ is greater than 0.95, analogous to a 0.05 level frequentist test. When both probabilities are less than 0.95 the decision is indeterminate.

4.3 Huber family

To implement the testing procedure of the preceding section, we require a method for fitting the family of centres to each distribution drawn from the posterior—i.e., for solving the optimisation problems in step two of Algorithm 6 given fixed bootstrap weights $w_1^{(b)}, \ldots, w_n^{(b)}$. This section develops a method for optimisation with the Huber function.

4.3.1 Optimisation problem

For simplicity of exposition, we drop the bootstrap iteration superscript (b) and fix $\sigma = 1$ without loss of generality. The Huber function as a function of the residual $x - \mu$ can then be expressed as

$$\ell_\lambda(x - \mu) = \begin{cases} \frac{1}{2}(x-\mu)^2 & \text{if } |x-\mu| < \lambda \\ \lambda|x-\mu| - \frac{1}{2}\lambda^2 & \text{if } |x-\mu| \geq \lambda. \end{cases}$$

We denote the loss over the weighted (bootstrap) sample by

$$\mathcal{L}_\lambda(\mu) := \sum_{i=1}^n w_i \ell_\lambda(x_i - \mu).$$

Our goal is to devise an algorithm for computing the set $\{\mu(\lambda) : \lambda \in \Lambda\}$, where

$$\mu(\lambda) := \arg\min_{\mu \in \mathbb{R}} \mathcal{L}_\lambda(\mu) \tag{4.3}$$

and $\Lambda = (0, \infty)$. For an equally weighted sample, (4.3) includes as limiting cases the sample mean and sample median. When the weights are unequal, the limit points become the *weighted mean* and *weighted median*, interpretable as the mean and median of the bootstrap sample. The weighted mean is defined by

$$\bar{\mu} := \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^n w_i (x_i - \mu)^2 = \sum_{i=1}^n w_i x_i,$$

and the weighted median by

$$\tilde{\mu} := \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^n w_i |x_i - \mu|.$$

There is no analytical solution for the weighted median. In fact, the weighted mean is the only Huber centre that admits an analytical solution in general.

If Λ were a finite set, it would be possible to solve the optimisation problem (4.3) for each of its elements. For given λ, the one-dimensional problem (4.3) is convex, and although it does not admit an analytical solution, it is amenable to simple numerical routines (Huber and Ronchetti 2009). Even if Λ is not finite, one might try approximating it using a fine grid and then proceed to solve each minimisation individually. Recall though each set of minimisation problems needs to be solved B times in the Bayesian bootstrap, where B might be 1,000, 10,000, or larger. Thus, even with an efficient algorithm, total cumulative runtime can be prohibitive. Notwithstanding runtime considerations, such an approach still only yields an approximation. Instead of an approximation, we propose a fast and exact pathwise algorithm that optimises (4.3) for all values of λ.

4.3.2 Pathwise optimisation routine

Our approach exploits piecewise linearity of the solution path $\mu(\lambda)$ for $\lambda \in (0, \infty)$, a property we now demonstrate. The gradient of the Huber function with respect to μ is

$$\frac{\partial \ell_\lambda(x-\mu)}{\partial \mu} = \begin{cases} -(x-\mu) & \text{if } |x-\mu| < \lambda \\ -\lambda \operatorname{sign}(x_i - \mu) & \text{if } |x-\mu| \geq \lambda. \end{cases}$$

Hence, the gradient of the loss over the weighted sample is

$$\frac{\partial \mathcal{L}_\lambda(\mu)}{\partial \mu} = \sum_{i=1}^{n} w_i \frac{\partial \ell_\lambda(x_i - \mu)}{\partial \mu} = - \sum_{i:|x_i-\mu|<\lambda} w_i (x_i - \mu) - \sum_{i:|x_i-\mu|\geq\lambda} w_i \lambda \operatorname{sign}(x_i - \mu).$$

We denote the above gradient by $\mathcal{L}'(\mu)$, suppressing the explicit dependency on λ. The chain rule gives

$$\frac{\partial \mathcal{L}'(\mu(\lambda))}{\partial \lambda} = \frac{\partial \mathcal{L}'(\mu)}{\partial \mu}\bigg|_{\mu=\mu(\lambda)} \frac{\partial \mu(\lambda)}{\partial \lambda},$$

which, after evaluating gradients and rearranging terms, leads to

$$\frac{\partial \mu(\lambda)}{\partial \lambda} = -\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i}. \tag{4.4}$$

Observe that the gradient of the solution path $\partial\mu(\lambda)/\partial\lambda$ is piecewise constant as a function of λ, implying that $\mu(\lambda)$ is piecewise linear. It follows that $\mu(\lambda)$ is also piecewise continuous with left and right limits. It can be verified that the left and right limits at any knot $\lambda^\star$ equal $\mu(\lambda^\star)$, and hence that $\mu(\lambda)$ is continuous.

Since the solution path is piecewise linear, it is composed of a sequence of knots, i.e., certain values of λ at which $|x_i - \mu(\lambda)| = \lambda$ for one or more sample points. These knots correspond to crossing events, where sample points transition between the square and absolute pieces of the Huber function. Lemma 4 characterises a useful property in relation to these crossing events.

Lemma 4. *Suppose sample point x_0 satisfies $|x_0 - \mu(\lambda^\star)| \geq \lambda^\star$ for some $\lambda^\star > 0$. Then, for all $0 < \lambda < \lambda^\star$, there holds $|x_0 - \mu(\lambda)| \geq \lambda$.*

Lemma 4 implies that, for a decreasing sequence of λ, once a sample point has crossed to the absolute piece of the Huber function, it remains there. This property guarantees the existence of at most n knots along the solution path.

To trace out the solution path, we need only fit μ at each λ in the sequence of knots since any solution between knots is linearly interpolable. A method to efficiently determine the location and solution at each knot is required. Suppose we are at an arbitrary point (λ, μ) along the solution path. Then, thanks to piecewise linearity, the closest knot point (λ^+, μ^+) to the left of (λ, μ) is computable by taking a step $\gamma > 0$ (of a certain size) as follows:

$$\lambda^+ = \lambda - \gamma \tag{4.5}$$

and

$$\mu^+ = \mu + \gamma \frac{\partial \mu(\lambda)}{\partial \lambda}. \tag{4.6}$$

Equation (4.4) provides an analytical expression for the gradient $\partial\mu(\lambda)/\partial\lambda$. An expression for the required step size γ is still needed. To this end, we present Proposition 4.

Proposition 4. *Let (λ, μ) be any point along the solution path such that $|x_i - \mu| < \lambda$ for at least one $i = 1, \ldots, n$. Then the largest positive step size before the solution path reaches a knot point (λ^+, μ^+) to the left of (λ, μ) is*

$$\gamma = \min_{i:|x_i-\mu|<\lambda} \left(\frac{\lambda - s_i(x_i - \mu)}{1 - s_i \partial\mu(\lambda)/\partial\lambda} \right),$$

where $s_i = \operatorname{sign}(x_i - \tilde{\mu})$ and $\tilde{\mu}$ is the weighted median.

The requirement $|x_i - \mu| < \lambda$ for at least one $i = 1, \ldots, n$ guarantees the existence of at least one more unexplored knot along the solution path. Beyond the first and last knots, the solution path is flat.

Putting together the above ingredients and letting $\lambda = \lambda^{(m)}$, $\lambda^+ = \lambda^{(m+1)}$, $\mu = \mu^{(m)}$, and $\mu^+ = \mu^{(m+1)}$ we arrive at Algorithm 7. Starting at the rightmost knot point

Algorithm 7: Pathwise optimisation for the Huber family

input : $(x_1, \ldots, x_n)$ and $(w_1, \ldots, w_n)$

initialise : $\mu^{(1)} = \sum_{i=1}^n w_i x_i$ and $\lambda^{(1)} = \max_i(|x_i - \mu^{(1)}|)$

Calculate the sign $s_i = \operatorname{sign}(x_i - \tilde{\mu})$ for $i = 1, \ldots, n$

for $m = 1, \ldots, n-1$ **do**

  if $\{i : |x_i - \mu^{(m)}| < \lambda^{(m)}\} = \emptyset$ **then** $m = m - 1$ **break**

  Calculate the gradient

$$\eta = -\frac{\sum_{i:|x_i-\mu^{(m)}|\geq\lambda^{(m)}} w_i \operatorname{sign}(x_i - \mu^{(m)})}{\sum_{i:|x_i-\mu^{(m)}|<\lambda^{(m)}} w_i}$$

  Calculate the step size

$$\gamma = \min_{i:|x_i-\mu^{(m)}|<\lambda^{(m)}} \left(\frac{\lambda^{(m)} - s_i(x_i - \mu^{(m)})}{1 - s_i\eta} \right)$$

  Perform the updates $\lambda^{(m+1)} = \lambda^{(m)} - \gamma$ and $\mu^{(m+1)} = \mu^{(m)} + \gamma\eta$

end

output : $(\lambda^{(1)}, \ldots, \lambda^{(m+1)})$ and $(\mu^{(1)}, \ldots, \mu^{(m+1)})$

$(\lambda^{(1)}, \mu^{(1)})$, which corresponds to the weighted mean, the algorithm forges a path step-by-step to the leftmost knot point, which corresponds to the weighted median. Figure 4.3 illustrates this process on $n = 30$ iid draws from a standard normal distribution. The algorithm begins at a value of λ large enough to induce the weighted mean as the centre and then iteratively decreases λ. The final λ in this sequence of iterates is sufficiently small to induce the weighted median as the centre. Observe that the path is piecewise linear and continuous.

Thus far the spread has been fixed at $\sigma = 1$. To recover the solution path for $\sigma \neq 1$, we scale the output $(\lambda^{(1)}, \ldots, \lambda^{(m+1)})$ from Algorithm 7 by multiplying it by σ. The centres $(\mu^{(1)}, \ldots, \mu^{(m+1)})$ do not change. This scaling has the intended effect of using the scaled residual $(x - \mu)/\sigma$ in the Huber function instead of $x - \mu$. We remind the reader that this scaling makes the solution path scale-free, relevant for testing independent samples (addressed in Section 4.4).

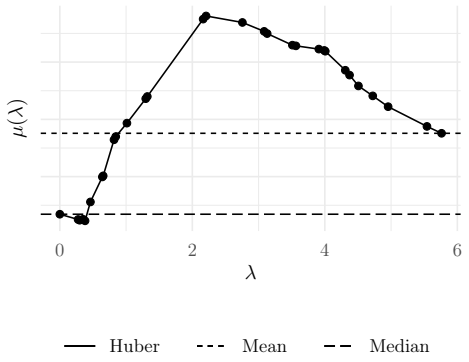

Figure 4.3: Algorithm 7 applied with $x_1, \ldots, x_n$ drawn from a standard normal distribution and $w_1, \ldots, w_n$ drawn from a flat Dirichlet distribution with $n = 30$. The solid points are iterates (knots) from the algorithm. Centres between iterates are linearly interpolated.

4.3.3 Relation to least angle regression

Algorithm 7 bears similarity to least angle regression (Efron et al. 2004; Rosset and Zhu 2007), a pathwise optimisation routine that traces the solution path of lasso regression coefficients. To clarify this similarity, first recall the Moreau envelope $f_\lambda(z)$ of a real-valued function $f(z)$, which is the infimal convolution of $f(z)$ and $g(z) = 1/(2\lambda)z^2$ (see, e.g., Polson et al. 2015). When $f(z) = |z|$, there is a precise relation between the Huber function $\ell_\lambda(z)$ and the Moreau envelope:

$$f_\lambda(z) := \inf_{\beta \in \mathbb{R}} \left(|\beta| + \frac{1}{2\lambda}(z - \beta)^2 \right) = \begin{cases} \frac{1}{2\lambda} z^2 & \text{if } |z| < \lambda \\ |z| - \frac{1}{2}\lambda & \text{if } |z| \geq \lambda. \end{cases}$$

The right-hand side is equal to $\ell_\lambda(z)/\lambda$. In words, multiplying the Moreau envelope of the absolute value function by λ yields the Huber function, a known result from convex analysis (Beck 2017). Hence, we have the chain of equalities

$$\begin{aligned} \min_{\mu \in \mathbb{R}} \sum_{i=1}^n w_i \ell_\lambda(x_i - \mu) &= \min_{\mu \in \mathbb{R}} \sum_{i=1}^n w_i \lambda f_\lambda(x_i - \mu) \\ &= \min_{\mu \in \mathbb{R}} \sum_{i=1}^n w_i \inf_{\beta_i \in \mathbb{R}} \left(\frac{1}{2}(x_i - \mu - \beta_i)^2 + \lambda |\beta_i| \right) \\ &= \min_{\mu, \beta_1, \ldots, \beta_n \in \mathbb{R}} \sum_{i=1}^n w_i \left(\frac{1}{2}(x_i - \mu - \beta_i)^2 + \lambda |\beta_i| \right). \end{aligned}$$

The infimum can be written as a minimum since the absolute value function is closed convex. The final line is a weighted lasso regression of $x_1, \ldots, x_n$ on an identity design matrix of dimensions $n \times n$, showing that the Huber problem (4.3) can be recast as a weighted lasso problem. Thus, applying least angle regression (configured with weights) to an identity design matrix yields a path identical to that produced by Algorithm 7. Despite this equivalence, the development of Algorithm 7 remains essential. Least angle regression is designed for general design matrices and, as such, does not exploit the structure of regression with an identity design (i.e., the Huber problem). Algorithm 7,

on the other hand, takes full advantage of this structure. In numerical experimentation, we observed that Algorithm 7 is typically an order of magnitude faster than least angle regression. Without this speedup, the Bayesian bootstrap would remain computationally burdensome.

4.4 Two-sample problem

The discussion up to now has focused on the one-sample setting. This section addresses the two-sample setting with samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$. Both paired samples and independent samples are covered.

4.4.1 Paired samples

The two samples are paired if X_i and Y_i are meaningfully coupled together (e.g., measurements on the same subject before and after treatment), in which case the sample sizes n_1 and n_2 are equal. Define the random variable Z_i as the difference $X_i - Y_i$. Then the familial hypotheses are

$$\mathrm{H}_0 : \mu_Z(\lambda) \in \mathcal{M}_0 \text{ for some } \lambda \in \Lambda \quad \text{vs.} \quad \mathrm{H}_1 : \mu_Z(\lambda) \in \mathcal{M}_1 \text{ for all } \lambda \in \Lambda,$$

where $\mu_Z(\lambda)$ is a centre of Z. The test and algorithms of the previous sections apply directly to the sample $Z_1, \ldots, Z_n$ with $n = n_1 = n_2$.

4.4.2 Independent samples

At least two different types of hypotheses are possible with independent samples. The first type is

$$\mathrm{H}_0 : \begin{array}{c} \mu_X(\lambda) - \mu_Y(\lambda) \in \mathcal{M}_0 \\ \text{for some } \lambda \in \Lambda \end{array} \quad \text{vs.} \quad \mathrm{H}_1 : \begin{array}{c} \mu_X(\lambda) - \mu_Y(\lambda) \in \mathcal{M}_1 \\ \text{for all } \lambda \in \Lambda. \end{array} \tag{4.7}$$

Here, the same centre of X is compared with the same centre of Y, i.e., the mean is compared with the mean, the median with the median, and so on. For the majority of situations we envisage, these hypotheses are a sensible choice. Another type of hypotheses is

$$\mathrm{H}_0 : \begin{array}{c} \mu_X(\lambda_1) - \mu_Y(\lambda_2) \in \mathcal{M}_0 \\ \text{for some } (\lambda_1, \lambda_2) \in \Lambda^2 \end{array} \quad \text{vs.} \quad \mathrm{H}_1 : \begin{array}{c} \mu_X(\lambda_1) - \mu_Y(\lambda_2) \in \mathcal{M}_1 \\ \text{for all } (\lambda_1, \lambda_2) \in \Lambda^2. \end{array} \tag{4.8}$$

With this type, every centre of X is compared with every centre of Y, i.e., the mean is compared with the median, the mean with the mean, etc. A test of hypotheses (4.8) is necessarily more conservative than a test of (4.7) since the former null encompasses the latter. We do not pursue (4.8) further in this chapter and leave it as the subject of future work.

Testing either of these hypotheses requires bootstrapping the families of X and Y with independently drawn weights. For a test of (4.7), each centre of Y is subtracted from the same centre of X. These differences are recorded within each bootstrap iteration. The posterior probability of H_0 is estimated by the proportion of times across bootstrap iterations that the set of differences intersects the null set $\mathcal{M}_0$.

4.5 Relation to intersection-union testing

The familial test we propose may be considered to have an *intersection-union* test format. Introduced by Berger (1982), an intersection-union test for a parameter $\theta \in \Theta$ is a test involving a null hypothesis that is a union of sets and an alternative hypothesis that is an intersection of sets.[4] Specifically, letting Θ_j denote a subset of Θ for $j = 1, 2, \ldots, k$, an intersection-union test evaluates the hypotheses

$$\mathrm{H}_0 : \theta \in \cup_{j=1}^{k} \Theta_j \quad \text{vs.} \quad \mathrm{H}_1 : \theta \in \cap_{j=1}^{k} \Theta_j^c, \tag{4.9}$$

where Θ_j^c is the complement of Θ_j. If H_0 is true, θ must be contained in at least one of the Θ_j subsets. Hence, to conduct an intersection-union test, it suffices to perform k separate tests of

$$\mathrm{H}_{0j} : \theta \in \Theta_j \quad \text{vs.} \quad \mathrm{H}_{1j} : \theta \in \Theta_j^c$$

and then reject the overall null hypothesis H_0 if and only if all k individual null hypotheses H_{0j} are rejected. Berger (1982) proves the overall type I error rate of this procedure is no bigger than α if the individual tests are conducted with level α. Berger (1982) also states conditions under which intersection-union tests have size exactly equal to α, since they are generally conservative with type I error rate less than α. Berger and Hsu (1996) generalise these conditions and also provide an example where an initially conservative intersection-union test can be modified to improve its frequentist power characteristics. Li et al. (2020) and Yin et al. (2021) contain some recent applications.

The connection between intersection-union tests and our familial test arises from the fact that the familial null and alternative can be broken down into a collection of individual hypotheses, each concerning a different centre indexed by λ:

$$\mathrm{H}_{0\lambda} : \mu(\lambda) \in \mathcal{M}_0 \quad \text{vs.} \quad \mathrm{H}_{1\lambda} : \mu(\lambda) \in \mathcal{M}_1.$$

Recall that $\mathcal{M}_0$ and $\mathcal{M}_1$ are a partition of the parameter space, so $\mathcal{M}_1 = \mathcal{M}_0^c$. Similar to an intersection-union test, the familial test rejects if and only if the individual null hypotheses $\mathrm{H}_{0\lambda}$ are rejected for all $\lambda \in \Lambda$. Consequently, the overall hypotheses can be expressed using a union and intersection:

$$\mathrm{H}_0 : \cup_{\lambda \in \Lambda} \{\mu(\lambda) \in \mathcal{M}_0\} \quad \text{vs.} \quad \mathrm{H}_1 : \cap_{\lambda \in \Lambda} \{\mu(\lambda) \in \mathcal{M}_1\}.$$

Here, the union and intersection are with respect to the individual events $\{\mu(\lambda) \in \mathcal{M}_0\}$ and $\{\mu(\lambda) \in \mathcal{M}_1\}$, rather than subsets of the parameter spaces as with the intersection-union test. Of course, the intersection-union hypotheses (4.9) can also be expressed in terms of individual events as $\mathrm{H}_0 : \cup_{j=1}^{k} \{\theta \in \Theta_j\}$ vs. $\mathrm{H}_1 : \cap_{j=1}^{k} \{\theta \in \Theta_j^c\}$. A key difference between the tests, however, is that the familial test involves an uncountable number of events, whereas the number of events k is typically finite in an intersection-union test. Though the Bayesian nonparametric procedure outlined in Section 4.2 does not formally control the size of the test, it is insightful to consider its size and power properties in repeated sampling experiments, an exercise undertaken next.

4.6 Simulations

This section reports numerical simulations designed to evaluate the finite sample properties of our test. To enable these exercises, the test and algorithms described in the preceding

[4] Intersection-union tests may be considered the opposite of union-intersection tests (Roy 1953), where the null is an intersection of sets and the alternative is a union of sets.

sections are implemented in the `R` package `familial`. For a sample of size $n = 200$, `familial` takes about half a second to perform 1,000 bootstraps for a single sample on one core of a modern processor. Parallelism is also supported. Run time scales linearly with the sample size, number of bootstraps, and if parallelised, number of processor cores.

4.6.1 One-sample and paired samples

We first study the one-sample setting with $X_1, \ldots, X_n$. This setting can also be interpreted as the paired samples setting where X_i is the difference of random variables. The distributions analysed are:

- $X \sim \text{Normal}(0, 1)$;
- $X \sim \text{Exponential}(1)$;
- $X \sim \text{Lognormal}(0, 1)$; and
- $X \sim \text{Poisson}(1)$.

These distributions cover different support types, skewness levels, and tail behaviors. Figure 4.4 visualises the distributions and their Huber families. For the normal, the family is a singleton. For the exponential and lognormal, the family is an interval with the mean and median as its endpoints. As the Poisson demonstrates, the family need not be bounded by the mean and median.

Table 4.1 summarises the tests evaluated and their associated hypotheses. A Bayesian

Test	Null hypothesis (H_0)	Alternative hypothesis (H_1)	Centre
Huber familial	$\exists\, \lambda \in \Lambda : \mu(\lambda) = \mu_0$	$\forall\, \lambda \in \Lambda : \mu(\lambda) \neq \mu_0$	Huber
Student t	$\mu = \mu_0$	$\mu \neq \mu_0$	Mean
Fisher sign	$\mu = \mu_0$	$\mu \neq \mu_0$	Median
Wilcoxon signed-rank	$\mu = \mu_0$	$\mu \neq \mu_0$	Median*

* Provided X is symmetric

Table 4.1: Tests evaluated in the one-sample (paired samples) setting.

adaptation of the signed-rank test developed by Benavoli et al. (2014) is also evaluated. The results from this Bayesian test are not included as they are practically indistinguishable from those for the regular signed-rank test. The t and signed-rank tests are performed using `t.test` and `wilcox.test` from the `stats` package in `R`. The sign test is a special case of the binomial test, performed using `binom.test` from the same package. To handle data points equal to the null value μ_0 (so-called ties) that can arise in testing discrete distributions, a modification to the sign test due to Fong et al. (2003) is used. The `wilcox.test` function uses a normal approximation, which is capable of dealing with ties. The Bayesian bootstrap used in the familial test has the advantage of being insensitive to ties. The number of Bayesian bootstraps is fixed at $B = 1,000$.

Figure 4.5 reports rejection frequencies for different values of μ_0 as averaged over 1,000 simulations. The sample size is fixed at $n = 200$. The shaded region indicates values of μ_0 for which the familial (Huber) null is true. Rejection frequency inside this region indicates the size of a test according to the familial null. Power of a test according to the familial alternative is the rejection frequency outside this region. The frequentist tests are carried out at the 0.05 level. The familial test is conducted using loss matrix (4.2), which rejects when the null has posterior probability less than 0.05.

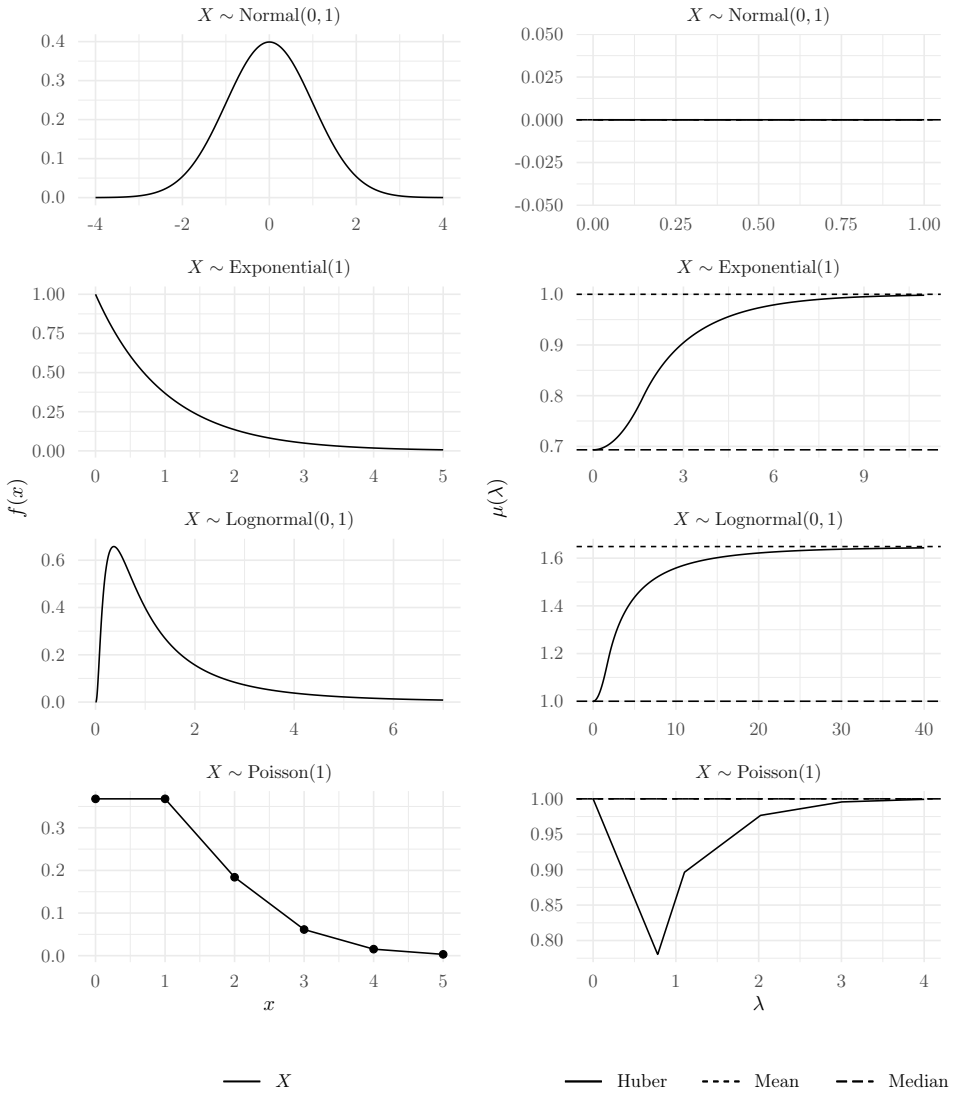

Figure 4.4: Distributions analysed in the one-sample (paired samples) setting. The plots in the left column depict the density or mass function for the population. The plots in the right column depict the corresponding Huber family.

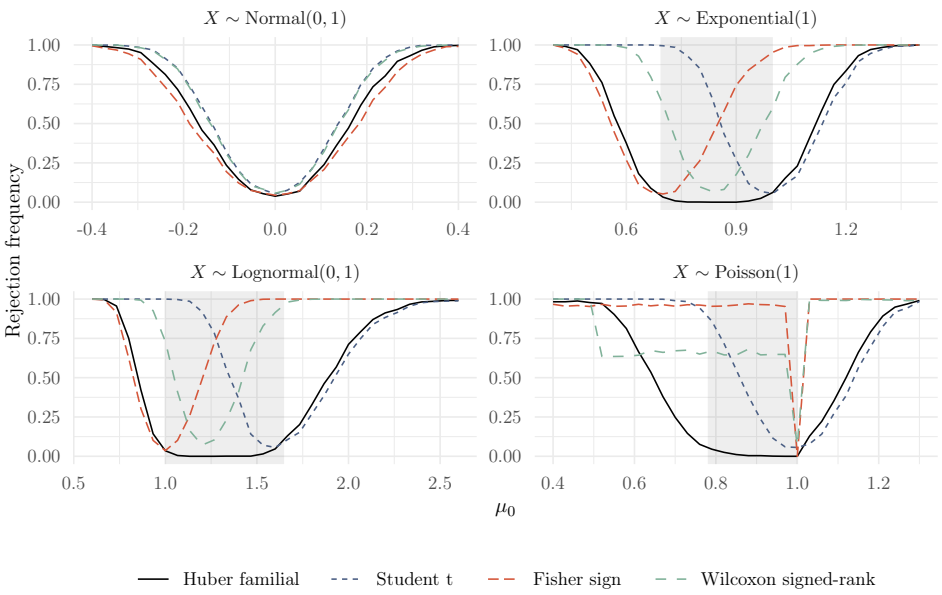

Figure 4.5: Rejection frequency as a function of the null value μ_0 in the one-sample (paired samples) setting. The sample size $n = 200$. The shaded region indicates values of μ_0 consistent with the familial null. Rejection frequency inside this region is size according to the familial null, and rejection frequency outside this region is power according to the familial alternative.

For the normal distribution, the familial test behaves similarly to the other tests. It has size no greater than 0.05 at $\mu_0 = 0$ and rejects sufficiently large departures from zero with high probability. Its power curve sits between those of the sign test and the signed-rank and t tests. The t test is well known to have optimal power here.

The story is more interesting for the exponential and lognormal distributions. Here, the curves for the sign and t tests attain their minima at different values of μ_0 since the null of each test is true at different locations. The signed-rank test fails as a test of the median due to X being asymmetric. The familial test behaves more conservatively than all three of these tests. It respects the familial null by rejecting with probability at most 0.05 in regions where some Huber centre is equal to μ_0. In regions with no Huber centre equal to μ_0, the familial test can be more powerful than the t or sign tests. For instance, it is more powerful than the t test for the exponential distribution when $\mu_0 > 1$. It is also more powerful than the sign test when $\mu_0 < 0.7$.

The Poisson distribution also tells an intriguing story. Since the Poisson is discrete, the power curves of the sign and signed-rank tests are step functions. In contrast to the other distributions, the curve of the sign test does not straddle the lower boundary of the familial null—due to the lower boundary being some centre other than the median. The familial test respects its null and has good power for $\mu_0 > 1$ compared with the t test.

4.6.2 Independent samples

We now consider the independent samples setting with $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$. The distributions analysed are:

- $X \sim \text{Normal}(0, 1)$, $Y \sim \text{Normal}(1, 1)$;

- $X \sim$ Exponential(1), $Y \sim$ Exponential(2);
- $X \sim$ Lognormal$(0, 1)$, $Y \sim$ Lognormal$(0, 0.5)$; and
- $X \sim$ Poisson(1), $Y \sim$ Poisson(1.2).

The distributions for X are the same as those in the one-sample setting. Figure 4.6 plots the distributions and corresponding differences in Huber families. For the normal, Y is a location shift on X, so the difference in families is a singleton. For the remaining distributions, Y has different skew (and tailedness) than X, so the difference in families are intervals. The Poisson is an example where the lower endpoint of the interval is not equal to the difference of means or medians.

We evaluate independent sample versions of the tests studied previously, summarised in Table 4.2. A Bayesian version of the rank-sum test by Benavoli et al. (2015) is also

Test	Null hypothesis (H_0)	Alternative hypothesis (H_1)	Centre
Huber familial	$\exists\, \lambda \in \Lambda :$ $\mu_X(\lambda) - \mu_Y(\lambda) = \mu_0$	$\forall\, \lambda \in \Lambda :$ $\mu_X(\lambda) - \mu_Y(\lambda) \neq \mu_0$	Huber
Welch t	$\mu_X - \mu_Y = \mu_0$	$\mu_X - \mu_Y \neq \mu_0$	Mean
Mood median	$\mu_X - \mu_Y = \mu_0$	$\mu_X - \mu_Y \neq \mu_0$	Median
Wilcoxon rank-sum	$\mu_X - \mu_Y = \mu_0$	$\mu_X - \mu_Y \neq \mu_0$	Median*

* Provided X and Y only differ in location

Table 4.2: Tests evaluated in the independent samples setting.

evaluated. The results from that test are not materially different from those for the regular rank-sum test, so they are not reported. The t and rank-sum tests are performed using `t.test` and `wilcox.test`. The median test is a special case of the chi-square test, performed using `chisq.test` from `stats`. Ties are again handled by `wilcox.test` via the normal approximation. For the median test, ties are discarded when calculating the test statistic.

Results from 1,000 simulations are reported in Figure 4.7. The sample sizes are fixed at $n_1 = n_2 = 200$. The shaded region again represents values of μ_0 consistent with the familial null.

The power curves for the normal distribution are not too different from the one-sample setting. The Huber centre for Y in the population is a point, so an independent samples test is not substantially different from a one-sample test with a point null.

For the exponential distribution, the rank-sum test fails as a test of medians since X and Y differ in scale, though curiously, it does not fail as a test of medians for the lognormal distribution. The t and median tests reject at rates above 0.05 in the middle of the familial null region, where the difference in means and difference in medians are both far from μ_0. There remains another Huber centre, not equal to the mean or median, for which the difference in centres is equal to μ_0. The familial test accounts for this centre and correctly accepts the null with high probability.

The median test applied to the Poisson distribution does not have the correct size at $\mu_0 = 0$ due to ties in the data. Likewise, the rank-sum test fails as a test of medians for the Poisson due to X and Y differing in shape and scale. The power curve of the t test does not straddle a boundary of the familial null. Unlike the median and rank-sum tests, the familial test succeeds as a test of medians, having zero size at $\mu_0 = 0$.

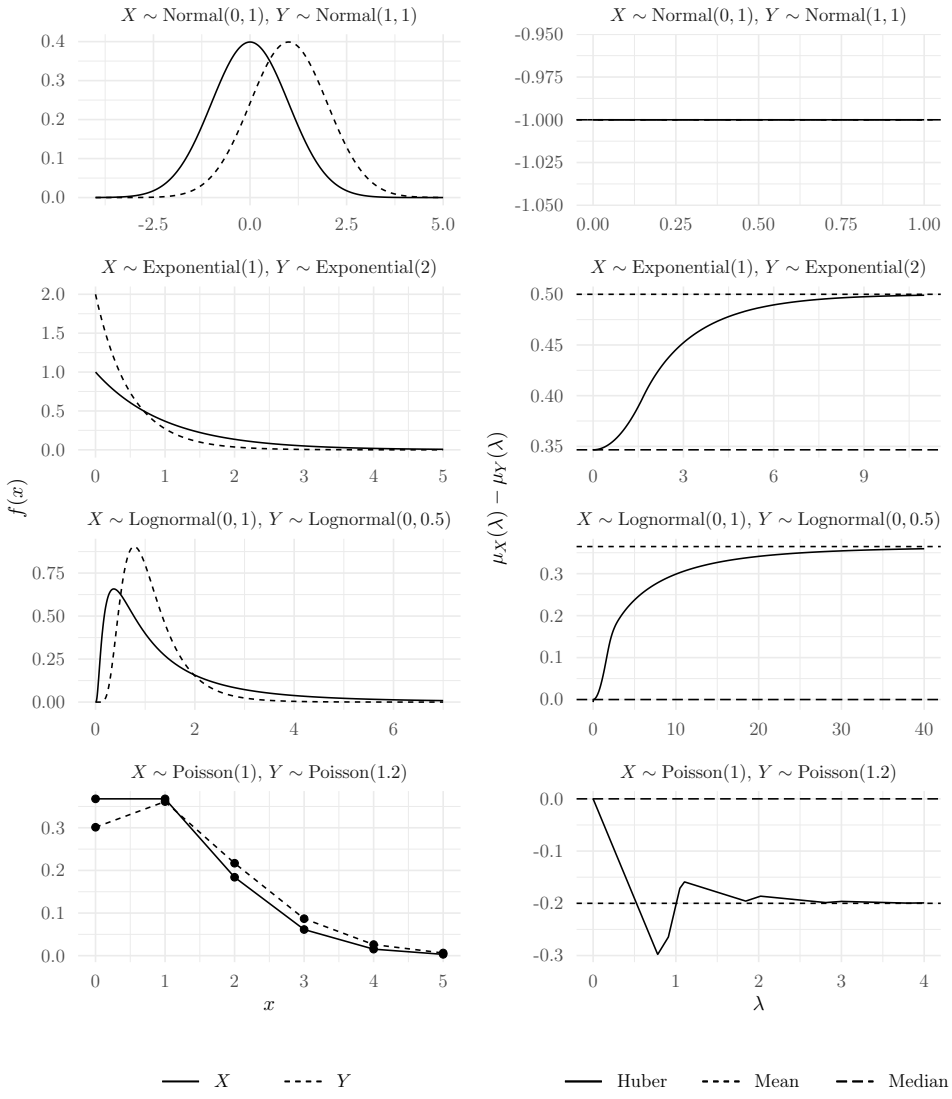

Figure 4.6: Distributions analysed in the independent samples setting. The plots in the left column depict the density or mass function for the populations. The plots in the right column depict the corresponding difference in Huber families.

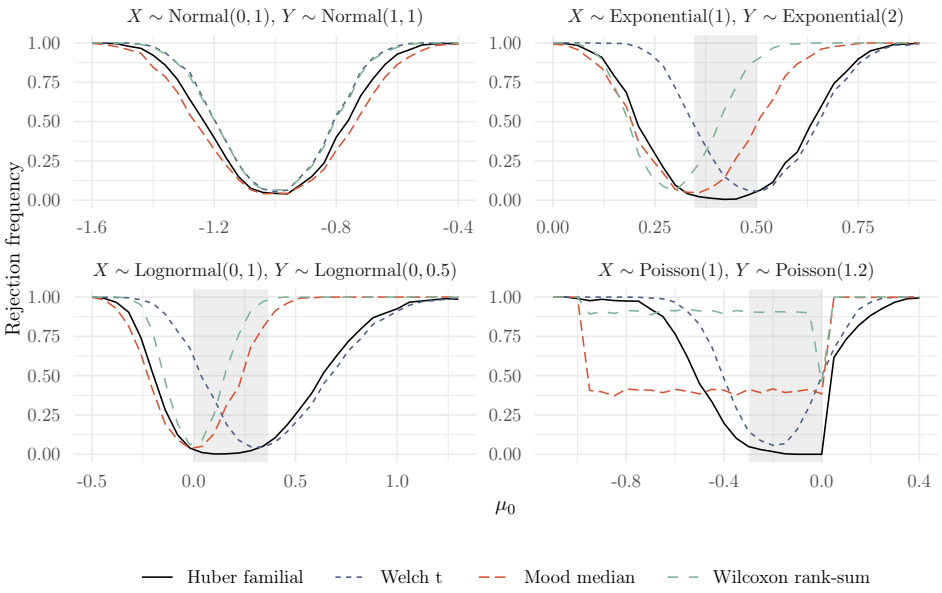

Figure 4.7: Rejection frequency as a function of the null value μ_0 in the independent samples setting. The sample sizes $n_1 = n_2 = 200$. The shaded region indicates values of μ_0 consistent with the familial null. Rejection frequency inside this region is size according to the familial null, and rejection frequency outside this region is power according to the familial alternative.

4.7 Case studies

This section illustrates the application of familial inference to two psychology experiments. The first illustration concerns paired samples. The second illustration addresses independent samples. All tests are conducted with the same configurations and rejection criteria used in the simulations.

4.7.1 Body posture study

Rosenbaum et al. (2017) conducted an experiment to ascertain the effect of body posture on selective attention.[5] The experiment employed the Stroop test, where subjects are asked to announce colours of a sequence of words and not the words themselves (e.g., announce 'blue' when the word 'red' is printed in blue). The difference in response times between congruent word-colour pairs and incongruent pairs is the Stroop effect. Experimental subjects took the test once while sitting and once while standing. The study found standing lowered the Stroop effect compared with sitting, indicating improved selective attention while standing.

The dataset contains paired observations on response times of $n = 50$ subjects. Figure 4.8 presents a histogram of differences in response time alongside a functional boxplot of the posterior Huber family. The response times do not deviate markedly from a normal distribution, though they are slightly left-skewed. The posterior concentrates well below zero, suggesting standing might reduce the Stroop effect.

[5] Refer to Experiment 3 in that paper.

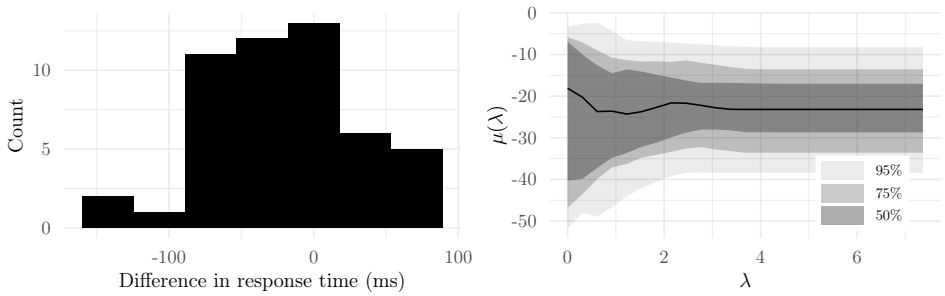

Figure 4.8: Body posture data. The left plot is a histogram of the data. The right plot is a functional boxplot of the posterior density of the Huber family. Shading indicates different central regions of the posterior.

The study reported a p-value of 0.004 from an F test of the interaction between congruency and posture in a repeated-measures ANOVA, equivalent to a Student t test that the mean difference in response times is zero. The Fisher sign test and Wilcoxon signed-rank test produce p-values of 0.007 and 0.006, respectively. The Huber familial test finds that the probability of the null is 0.005. All tests reject the null that body posture does not affect the Stroop effect. This result confirms that the original finding is not sensitive to the centre tested.

4.7.2 Multi-task perception study

Srna et al. (2018) ran an experiment to investigate if human performance at certain activities is affected by whether the activity is perceived as multi-tasking.[6] Experimental subjects were required to watch a video and transcribe the audio. This activity was framed as multi-tasking to a treatment group and single-tasking to a control group. Assignment to either group was random. The study found that subjects in the treatment group transcribed more words than those in the control group and the accuracy of their transcriptions was higher, suggesting perceiving an activity as multi-tasking improves performance at that activity.

We focus on the number of words transcribed. The dataset contains $n_1 = 82$ subjects in the treatment group and $n_2 = 80$ in the control group; see Figure 4.9. The groups are dissimilar in distribution, with the multi-task group being unimodal and the single-task group being multimodal. For small values of λ, the 50% central region of the posterior includes zero, indicating the null might be plausible.

The study reported a p-value of 0.033 from an F test of the multi-task condition in a one-way ANOVA, identical to a two-sample Student t test (with equal variance) that the mean number of words transcribed is equal between groups. The Mood median test yields a p-value of 0.271. The p-value from a Wilcoxon rank-sum test is 0.072. The Huber familial test returns 0.170 as the probability of the null. In contrast to the t test, the familial, sign, and rank-sum tests do not find the multi-task condition to affect performance. In particular, the familial test fails to find sufficient support for either hypothesis and returns an indeterminate result. Whether this is a meaningful discrepancy remains up to subject-matter experts to decide.

[6] Refer to Study 1a in that paper.

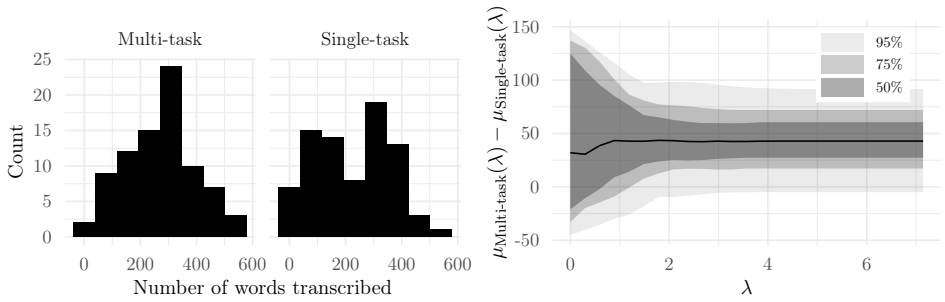

Figure 4.9: Multi-task perception data. The left plot is a histogram of the data by control and treatment. The right plot is a functional boxplot of the posterior density of the difference in Huber families. Shading indicates different central regions of the posterior.

4.8 Concluding remarks

It has become standard practice to translate scientific hypotheses into statistical hypotheses about a specific centre for the underlying distribution(s). Despite the ubiquity of this approach, there can be a lack of consensus about which centre bests reflects the original scientific hypotheses. When there is ambiguity, we argue one should adopt familial inference, which formulates hypotheses via a family of plausible centres. The contribution of this chapter is to study familial inference for centres belonging to the Huber family. A natural next step in this line of work is to develop familial inference for other statistical parameters such as conditional centres (regressions). Frequentist tests can be developed along these lines as well.

Our package `familial` implements the tools developed in this chapter and is publicly available on `CRAN`.

Chapter 5

Conclusion

The data analytic landscape of the 21st century has brought to the fore a profusion of challenging statistical problems. In this landscape, high- and infinite-dimensional statistical problems—the focus of this thesis—are among the most germane.

In Chapter 2, we consider sparse high-dimensional regression from the perspective of robustness. Our central contribution is a robust adaption of best subset selection, for which we devise a powerful algorithmic framework for computation and establish a theoretical guarantee for robustness. We use simulated and real data to illustrate the new estimator's positive qualities relative to existing continuous shrinkage estimators.

In Chapter 3, we study structured sparsity for high-dimensional regression and classification. The key contribution of this chapter are structured sparse estimators. We develop scalable computational algorithms for the new estimators and provide theoretical insights into their estimation error. In application to sparse semiparametric modelling, we demonstrate their improvement over existing convex and nonconvex estimators.

Finally, in Chapter 4, we shift our focus to hypothesis testing for centres. Our core contribution is the familial test, motivated by the idea of evaluating a family of centres rather than a single centre. For this infinite-dimensional testing problem, we propose a Bayesian nonparametric procedure enabled by a novel optimisation routine. Extensive empirical comparisons verify the favourable properties of the new test.

5.1 Future directions

Computation for best subset selection and its cousins, such as robust subset selection in Chapter 2, remains an open topic of research. Besides the custom projected gradient descent heuristics and off-the-shelf mixed-integer programming solvers used in this thesis, an intriguing possibility is to solve subset selection problems using neural networks. A growing body of research is exploring the application of graph neural networks—neural nets that input and output graphs—for the computation of solutions to combinatorial optimisation problems (Cappart et al. 2021; Nair et al. 2021). The potential for their application to subset selection is twofold. First, neural nets could produce new heuristics for finding near-optimal subsets. Unlike projected gradient heuristics, which are manually derived and hand-coded, the heuristics from neural nets are learned automatically by training over many different problem instances. These learned heuristics can be arbitrarily elaborate. Second, neural nets could produce new strategies for proving optimality of a subset. For instance, branching strategies—which guide the exploration of a branch-and-bound tree—can be learned using neural nets. These strategies influence how fast the

optimality gap is closed and thus the time to determine a provably optimally subset.

The structured sparsity penalties studied in Chapter 3 are versatile tools. A small but emerging line of research is applying structured penalties to induce predictor sparsity in neural nets (Feng and Simon 2019; Lemhadri et al. 2021). For this task, a set of groups are constructed, each group containing all the weights associated with a predictor at the network's input layer. Only if a predictor's group is selected is the predictor included in the network. Among other benefits, predictor-sparse neural nets are feasible for high-dimensional data—a setting where traditional neural nets fare poorly. Existing research on this topic has focused exclusively on lasso penalties. Given the positive results of Chapter 3, it is natural to ask what subset selection-type penalties have to offer in this domain. Setting aside the question of lasso vs. subset selection, another interesting yet seemingly unexplored line of research is to apply structured penalties to fit partially linear neural nets. Similar to the semiparametric models of Chapter 3, these neural nets would allow a subset of predictors to affect the response linearly. Besides possible gains in predictive performance, partially linear neural nets would have an edge in interpretability. They can be viewed as a more elaborate take on our semiparametric modelling approach.

The familial inference framework presented in Chapter 4 can accommodate statistical models more sophisticated than the unconditional centres we study as a starting point in this thesis. Regressions (i.e., conditional centres) are one important class of models, particularly given their ubiquity throughout the applied sciences. We expect our tools to extend gracefully to regression models. With minor modifications, one can apply our Bayesian nonparametric testing procedure to perform familial inference on one or more model coefficients. The key task remains to query the posterior for the probability that the targets of inference intersect the null set. The question of optimisation for the Huber family is somewhat more complicated. Importantly though, the inclusion of regressors does not violate the piecewise linearity of the Huber solution path. For this reason, we are confident that the optimisation algorithm developed in Chapter 4 can be generalised to accommodate regressors. Beyond developing the methodological tools themselves, we are interested in establishing asymptotic consistency results for familial tests of regressions. As regressions include unconditional centres as special cases, any consistency results would also provide theoretical guarantees for the familial test devised in Chapter 4.

Bibliography

Aardal, K., Nemhauser, G. L., and Weismantel, R., eds. (2005). *Handbooks in operations research and management science: Discrete optimization*. 1st ed. Vol. 12. Amsterdam, The Netherlands: Elsevier.

Akaike, H. (1973). 'Information theory and an extension of the maximum likelihood principle'. *Proceedings of the 2nd International Symposium on Information Theory*. Vol. 108, pp. 267–281.

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

Amato, U., Antoniadis, A., De Feis, I., and Gijbels, I. (2021). 'Penalised robust estimators for sparse and high-dimensional linear models'. *Statistical Methods and Applications* 30.1, pp. 1–48.

Barrientos, A. F. and Peña, V. (2020). 'Bayesian bootstraps for massive data'. *Bayesian Analysis* 15.2, pp. 363–388.

Beale, E. M. L., Kendall, M. G., and Mann, D. W. (1967). 'The discarding of variables in multivariate analysis'. *Biometrika* 54.3/4, pp. 357–366.

Beck, A. (2015). 'On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes'. *SIAM Journal on Optimization* 25.1, pp. 185–209.

— (2017). *First-order methods in optimization*. MOS-SIAM Series on Optimization. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics and Mathematical Optimization Society.

Beck, A. and Eldar, Y. C. (2013). 'Sparsity constrained nonlinear optimization: Optimality conditions and algorithms'. *SIAM Journal on Optimization* 23.3, pp. 1480–1509.

Beck, A. and Tetruashvili, L. (2013). 'On the convergence of block coordinate descent type methods'. *SIAM Journal on Optimization* 23.4, pp. 2037–2060.

Ben-Aharon, O., Magnezi, R., Leshno, M., and Goldstein, D. A. (2019). 'Median survival or mean survival: Which measure is the most appropriate for patients, physicians, and policymakers?' *Oncologist* 24.11, pp. 1469–1478.

Benavoli, A., Mangili, F., Corani, G., Zaffalon, M., and Ruggeri, F. (2014). 'A Bayesian Wilcoxon signed-rank test based on the Dirichlet process'. *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32, pp. 1026–1034.

Benavoli, A., Corani, G., Demšar, J., and Zaffalon, M. (2017). 'Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis'. *Journal of Machine Learning Research* 18, pp. 1–36.

Benavoli, A., Mangili, F., Ruggeri, F., and Zaffalon, M. (2015). 'Imprecise Dirichlet process with application to the hypothesis test on the probability that $X \leq Y$'. *Journal of Statistical Theory and Practice* 9.3, pp. 658–684.

Berger, R. L. (1982). 'Multiparameter hypothesis testing and acceptance sampling'. *Technometrics* 24.4, pp. 295–300.

Berger, R. L. and Hsu, J. C. (1996). 'Bioequivalence trials, intersection–union tests and equivalence confidence sets'. *Statistical Science* 11.4, pp. 283–319.

Bertsimas, D. and King, A. (2016). 'OR Forum—An algorithmic approach to linear regression'. *Operations Research* 64.1, pp. 2–16.

Bertsimas, D., King, A., and Mazumder, R. (2016). 'Best subset selection via a modern optimization lens'. *Annals of Statistics* 44.2, pp. 813–852.

Bertsimas, D. and Mazumder, R. (2014). 'Least quantile regression via modern optimization'. *Annals of Statistics* 42.6, pp. 2494–2525.

Bertsimas, D., Pauphilet, J., and Van Parys, B. (2020). 'Sparse regression: Scalable algorithms and empirical performance'. *Statistical Science* 35.4, pp. 555–578.

Bertsimas, D. and Van Parys, B. (2020). 'Sparse high-dimensional regression: Exact scalable algorithms and phase transitions'. *Annals of Statistics* 48.1, pp. 300–323.

Bhatia, K., Jain, P., and Kar, P. (2015). 'Robust regression via hard thresholding'. *Advances in Neural Information Processing Systems*. Vol. 28, pp. 721–729.

Blakely, T. A. and Kawachi, I. (2001). 'What is the difference between controlling for mean versus median income in analyses of income inequality?' *Journal of Epidemiology and Community Health* 55.5, pp. 352–353.

Breheny, P. and Huang, J. (2011). 'Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection'. *Annals of Applied Statistics* 5.1, pp. 232–253.

— (2015). 'Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors'. *Statistics and Computing* 25.2, pp. 173–187.

Breiman, L. (1996). 'Heuristics of instability and stabilization in model selection'. *Annals of Statistics* 24.6, pp. 2350–2383.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). 'Aggregation for Gaussian regression'. *Annals of Statistics* 35.4, pp. 1674–1697.

Candes, E. and Tao, T. (2007). 'The Dantzig selector: Statistical estimation when p is much larger than n'. *Annals of Statistics* 35.6, pp. 2313–2351.

Cappart, Q., Chételat, D., Khalil, E. B., Lodi, A., Morris, C., and Veličković, P. (2021). 'Combinatorial optimization and reasoning with graph neural networks'. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 4348–4355.

Chang, L., Roberts, S., and Welsh, A. (2018). 'Robust lasso regression using Tukey's biweight criterion'. *Technometrics* 60.1, pp. 36–47.

Chen, Y., Caramanis, C., and Mannor, S. (2013). 'Robust sparse regression under adversarial corruption'. *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28. 3, pp. 774–782.

Chouldechova, A. and Hastie, T. (2015). 'Generalized additive model selection'. arXiv: 1506.03850.

Christensen, G. and Miguel, E. (2018). 'Transparency, reproducibility, and the credibility of economics research'. *Journal of Economic Literature* 56.3, pp. 920–980.

Christidis, A.-A., Lakshmanan, L., Smucler, E., and Zamar, R. (2020). 'Split regularized regression'. *Technometrics* 62.3, pp. 330–338.

Cohen Freue, G. V., Kepplinger, D., Salibián-Barrera, M., and Smucler, E. (2019). 'Robust elastic net estimators for variable selection and identification of proteomic biomarkers'. *Annals of Applied Statistics* 13.4, pp. 2065–2090.

De Mol, C., Giannone, D., and Reichlin, L. (2008). 'Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?' *Journal of Econometrics* 146.2, pp. 318–328.

Dedieu, A., Hazimeh, H., and Mazumder, R. (2021). 'Learning sparse classifiers: Continuous and mixed integer optimization perspectives'. *Journal of Machine Learning Research* 22, pp. 41–47.

Ding, J., Tarokh, V., and Yang, Y. (2018). 'Model selection techniques: An overview'. *IEEE Signal Processing Magazine* 35.6, pp. 16–34.

Donoho, D. L. and Huber, P. J. (1983). 'The notion of breakdown point'. *A Festschrift for Erich L. Lehmann: In Honor of His Sixty-Fifth Birthday*. Ed. by Bickel, P. J., Doksum, K. A., and Hodges, J. L. Wadsworth Statistics/Probability Series. Belmont, CA, USA: Wadsworth, pp. 157–184.

Efron, B. (1979). 'Bootstrap methods: Another look at the jackknife'. *Annals of Statistics* 7.1, pp. 1–26.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). 'Least angle regression'. *Annals of Statistics* 32.2, pp. 407–499.

Falk, M. (1999). 'A simple approach to the generation of uniformly distributed random variables with prescribed correlations'. *Communications in Statistics - Simulation and Computation* 28.3, pp. 785–791.

Fan, J. and Li, R. (2001). 'Variable selection via nonconcave penalized likelihood and its oracle properties'. *Journal of the American Statistical Association* 96.456, pp. 1348–1360.

Feng, J. and Simon, N. (2019). 'Sparse-input neural networks for high-dimensional nonparametric regression and classification'. arXiv: 1711.07592.

Ferguson, T. S. (1973). 'A Bayesian analysis of some nonparametric problems'. *Annals of Statistics* 1.2, pp. 209–230.

Filippi, S. and Holmes, C. C. (2017). 'A Bayesian nonparametric approach to testing for dependence between random variables'. *Bayesian Analysis* 12.4, pp. 919–938.

Fisher, R. A. (1925). *Statistical methods for research workers*. London, UK: Oliver and Boyd.

Fong, D. Y. T., Kwan, C. W., Lam, K. F., and Lam, K. S. L. (2003). 'Use of the sign test for the median in the presence of ties'. *American Statistician* 57.4, pp. 237–240.

Fong, E., Lyddon, S., and Holmes, C. (2019). 'Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap'. *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97, pp. 1952–1962.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). 'Pathwise coordinate optimization'. *Annals of Applied Statistics* 1.2, pp. 302–332.

Furnival, G. M. and Wilson, R. W. (1974). 'Regressions by leaps and bounds'. *Technometrics* 16.4, pp. 499–511.

Garside, M. J. (1965). 'The best sub-set in multiple regression analysis'. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 14.2/3, pp. 196–200.

Gasparini, M. (1995). 'Exact multivariate Bayesian bootstrap distributions of moments'. *Annals of Statistics* 23.3, pp. 762–768.

Gatu, C. and Kontoghiorghes, E. J. (2006). 'Branch-and-bound algorithms for computing the best-subset regression models'. *Journal of Computational and Graphical Statistics* 15.1, pp. 139–156.

Guo, Y., Berman, M., and Gao, J. (2014). 'Group subset selection for linear regression'. *Computational Statistics and Data Analysis* 75, pp. 39–52.

Guo, Y., Zhu, Z., and Fan, J. (2021). 'Best subset selection is robust against design dependence'. arXiv: 2007.01478.

Gurobi Optimization (2020). *Gurobi 9.0 performance benchmarks*. Tech. rep. URL: https://www.gurobi.com/wp-content/uploads/2020/02/Performance-Gurobi-9.0-1.pdf.

Gutiérrez, L., Barrientos, A. F., González, J., and Taylor-Rodríguez, D. (2019). 'A Bayesian nonparametric multiple testing procedure for comparing several treatments against a control'. *Bayesian Analysis* 14.2, pp. 649–675.

Hall, P. and Hart, J. D. (1990). 'Bootstrap test for difference between means in nonparametric regression'. *Journal of the American Statistical Association* 85.412, pp. 1039–1049.

Hampel, F. R. (1971). 'A general qualitative definition of robustness'. *Annals of Mathematical Statistics* 42.6, pp. 1887–1896.

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, pp. 579–592.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity. The lasso and generalizations*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. Boca Raton, FL, USA: CRC Press.

Hazimeh, H. and Mazumder, R. (2020). 'Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms'. *Operations Research* 68.5, pp. 1517–1537.

Hazimeh, H., Mazumder, R., and Radchenko, P. (2023). 'Grouped variable selection with discrete optimization: Computational and statistical perspectives'. *Annals of Statistics* 51.1, pp. 1–32.

Hocking, R. R. and Leslie, R. N. (1967). 'Selection of the best subset in regression analysis'. *Technometrics* 9.4, pp. 531–540.

Hofmann, M., Gatu, C., and Kontoghiorghes, E. J. (2007). 'Efficient algorithms for computing the best subset regression models for large-scale problems'. *Computational Statistics and Data Analysis* 52.1, pp. 16–29.

— (2010). 'An exact least trimmed squares algorithm for a range of coverage values'. *Journal of Computational and Graphical Statistics* 19.1, pp. 191–204.

Holmes, C. C., Caron, F., Griffin, J. E., and Stephens, D. A. (2015). 'Two-sample Bayesian nonparametric hypothesis testing'. *Bayesian Analysis* 10.2, pp. 297–320.

Huang, L. and Ghosh, M. (2014). 'Two-sample hypothesis testing under Lehmann alternatives and Polya tree priors'. *Statistica Sinica* 24.4, pp. 1717–1733.

Huber, P. J. (1964). 'Robust estimation of a location parameter'. *Annals of Mathematical Statistics* 35.1, pp. 73–101.

— (1981). *Robust statistics*. Wiley Series in Probability and Mathematical Statistics. New York, NY, USA: John Wiley & Sons.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons.

Insolia, L., Kenney, A., Chiaromonte, F., and Felici, G. (2021). 'Simultaneous feature selection and outlier detection with optimality guarantees'. *Biometrics* 78.4, pp. 1592–1603.

Ioannidis, J. P. A. (2005). 'Why most published research findings are false'. *PLoS Medicine* 2.8, pp. 696–701.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). 'Group lasso with overlap and graph lasso'. *Proceedings of the 26th International Conference on Machine Learning*, pp. 433–440.

Janssens, K. H., Deraedt, I., Schalm, O., and Veeckman, J. (1998). 'Composition of 15–17th century archaeological glass vessels excavated in Antwerp, Belgium'. *Modern Developments and Applications in Microbeam Analysis*. Ed. by Love, G., Nicholson, W. A. P., and Armigliato, A. Vol. 15. Mikrochimica Acta Supplement. Vienna, Austria: Springer, Vienna, pp. 253–267.

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, pp. 566–577.

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

Kohavi, R. (1995). 'A study of cross-validation and bootstrap for accuracy estimation and model selection'. *Proceedings of the 14th International Joint Conference on Artifical Intelligence*. Vol. 2, pp. 1137–1143.

Kreber, D. (2019). 'A mixed-integer optimization approach to an exhaustive cross-validated model selection for regression'. URL: http://www.optimization-online.org/DB_HTML/2019/05/7188.html.

Kruschke, J. K. (2013). 'Bayesian estimation supersedes the t test'. *Journal of Experimental Psychology: General* 142.2, pp. 573–603.

Kudo, K., Takano, Y., and Nomura, R. (2020). 'Stochastic discrete first-order algorithm for feature subset selection'. *IEICE Transactions on Information and Systems* E103-D.7, pp. 1693–1702.

Lambert-Lacroix, S. and Zwald, L. (2011). 'Robust regression through the Huber's criterion and adaptive lasso penalty'. *Electronic Journal of Statistics* 5, pp. 1015–1053.

Laurent, B. and Massart, P. (2000). 'Adaptive estimation of a quadratic functional by model selection'. *Annals of Statistics* 28.5, pp. 1302–1338.

Lee, J. and MacEachern, S. N. (2014). 'Inference functions in high dimensional Bayesian inference'. *Statistics and Its Interface* 7.4, pp. 477–486.

Lemberge, P., De Raedt, I., Janssens, K. H., Wei, F., and Van Espen, P. J. (2000). 'Quantitative analysis of 16–17th century archaeological glass vessels using PLS regression of EPXMA and μ-XRF data'. *Journal of Chemometrics* 14.5/6, pp. 751–763.

Lemhadri, I., Ruan, F., Abraham, L., and Tibshirani, R. (2021). 'LassoNet: A neural network with feature sparsity'. *Journal of Machine Learning Research* 22, pp. 1–29.

Li, D., Cao, J., and Zhang, S. (2020). 'Power analysis for cluster randomized trials with multiple binary co-primary endpoints'. *Biometrics* 76.4, pp. 1064–1074.

Li, J. and Chen, W. (2014). 'Forecasting macroeconomic time series: Lasso-based approaches and their forecast combinations with dynamic factor models'. *International Journal of Forecasting* 30.4, pp. 996–1015.

Lim, M. and Hastie, T. (2015). 'Learning interactions via hierarchical group-lasso regularization'. *Journal of Computational and Graphical Statistics* 24.3, pp. 627–654.

Liu, L., Shen, Y., Li, T., and Caramanis, C. (2020). 'High dimensional robust sparse regression'. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*. Vol. 108, pp. 411–421.

Lo, A. Y. (1987). 'A large sample study of the Bayesian bootstrap'. *Annals of Statistics* 15.1, pp. 360–375.

Lou, Y., Bien, J., Caruana, R., and Gehrke, J. (2016). 'Sparse partially linear additive models'. *Journal of Computational and Graphical Statistics* 25.4, pp. 1026–1040.

Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. B. (2011). 'Oracle inequalities and optimal inference under group sparsity'. *Annals of Statistics* 39.4, pp. 2164–2204.
Lozano, A. C., Meinshausen, N., and Yang, E. (2016). 'Minimum distance lasso for robust high-dimensional regression'. *Electronic Journal of Statistics* 10.1, pp. 1296–1340.
Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019). 'General Bayesian updating and the loss-likelihood bootstrap'. *Biometrika* 106.2, pp. 465–478.
Ma, L. and Wong, W. H. (2011). 'Coupling optional Pólya trees and the two sample problem'. *Journal of the American Statistical Association* 106.496, pp. 1553–1565.
MacEachern, S. (1993). 'An evaluation of Bayes posterior probability regions for a survival curve'. *Journal of Nonparametric Statistics* 3.2, pp. 175–186.
MacEachern, S. N. (2016). 'Nonparametric Bayesian methods: A gentle introduction and overview'. *Communications for Statistical Applications and Methods* 23.6, pp. 445–466.
Mann, H. B. and Whitney, D. R. (1947). 'On a test of whether one of two random variables is stochastically larger than the other'. *Annals of Mathematical Statistics* 18.1, pp. 50–60.
Maronna, R. A. (2011). 'Robust ridge regression for high-dimensional data'. *Technometrics* 53.1, pp. 44–53.
Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust statistics: Theory and methods (with R)*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons.
Mazumder, R. and Radchenko, P. (2017). 'The discrete Dantzig selector: Estimating sparse linear models via mixed integer linear optimization'. *IEEE Transactions on Information Theory* 63.5, pp. 3053–3075.
Mazumder, R., Radchenko, P., and Dedieu, A. (2023). 'Subset selection with shrinkage: Sparse linear modeling when the SNR is low'. *Operations Research* 71.1, pp. 129–147.
McCann, L. and Welsch, R. E. (2007). 'Robust variable selection using least angle regression and elemental set sampling'. *Computational Statistics and Data Analysis* 52.1, pp. 249–257.
McCracken, M. W. and Ng, S. (2016). 'FRED-MD: A monthly database for macroeconomic research'. *Journal of Business and Economic Statistics* 34.4, pp. 574–589.
Meier, L., van de Geer, S., and Bühlmann, P. (2008). 'The group lasso for logistic regression'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1, pp. 53–71.
Menjoge, R. S. and Welsch, R. E. (2010). 'A diagnostic method for simultaneous feature selection and outlier identification in linear regression'. *Computational Statistics and Data Analysis* 54.12, pp. 3181–3193.
Mood, A. M. (1950). *Introduction to the theory of statistics*. McGraw-Hill Series in Probability and Statistics. New York, NY, USA: McGraw-Hill.
Nair, V. et al. (2021). 'Solving mixed integer programs using neural networks'. arXiv: 2012.13349.
Natarajan, B. K. (1995). 'Sparse approximate solutions to linear systems'. *SIAM Journal on Computing* 24.2, pp. 227–234.
Neumeyer, N. and Dette, H. (2003). 'Nonparametric comparison of regression curves: An empirical process approach'. *Annals of Statistics* 31.3, pp. 880–920.
Nguyen, N. H. and Tran, T. D. (2013). 'Robust lasso with missing and grossly corrupted observations'. *IEEE Transactions on Information Theory* 59.4, pp. 2036–2058.
Obozinski, G., Jacob, L., and Vert, J.-P. (2011). 'Group lasso with overlaps: The latent group lasso approach'. arXiv: 1110.0413.

Obozinski, G., Taskar, B., and Jordan, M. (2006). *Multi-task feature selection*. Tech. rep. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.951&rep=rep1&type=pdf.

Open Science Collaboration (2015). 'Estimating the reproducibility of psychological science'. *Science* 349.6251, aac4716.

Percival, D. (2012). 'Theoretical properties of the overlapping groups lasso'. *Electronic Journal of Statistics* 6, pp. 269–288.

Pereira, L. A., Taylor-Rodríguez, D., and Gutiérrez, L. (2020). 'A Bayesian nonparametric testing procedure for paired samples'. *Biometrics* 76.4, pp. 1133–1146.

Polson, N. G., Scott, J. G., and Willard, B. T. (2015). 'Proximal algorithms in statistics and machine learning'. *Statistical Science* 30.4, pp. 559–581.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). 'Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls'. *IEEE Transactions on Information Theory* 57.10, pp. 6976–6994.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). 'Sparse additive models'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.5, pp. 1009–1030.

Rigollet, P. (2015). *18.S997: High dimensional statistics. Lecture notes.*

Rosenbaum, D., Mama, Y., and Algom, D. (2017). 'Stand by your Stroop: Standing up enhances selective attention and cognitive control'. *Psychological Science* 28.12, pp. 1864–1867.

Rosset, S. and Zhu, J. (2007). 'Piecewise linear regularized solution paths'. *Annals of Statistics* 35.3, pp. 1012–1030.

Rousseeuw, P. J. (1984). 'Least median of squares regression'. *Journal of the American Statistical Association* 79.388, pp. 871–880.

Rousseeuw, P. J. and Van Driessen, K. (2006). 'Computing LTS regression for large data sets'. *Data Mining and Knowledge Discovery* 12.1, pp. 29–45.

Rousselet, G. A. and Wilcox, R. R. (2020). 'Reaction times and other skewed distributions: Problems with the mean and the median'. *Meta-Psychology* 4.

Roy, S. N. (1953). 'On a heuristic method of test construction and its use in multivariate analysis'. *Annals of Mathematical Statistics* 24.2, pp. 220–238.

Rubin, D. B. (1981). 'The Bayesian bootstrap'. *Annals of Statistics* 9.1, pp. 130–134.

Savage, V. M. and West, G. B. (2007). 'A quantitative, theoretical framework for understanding mammalian sleep'. *Proceedings of the National Academy of Sciences of the United States of America* 104.3, pp. 1051–1056.

Schwarz, G. (1978). 'Estimating the dimension of a model'. *Annals of Statistics* 6.2, pp. 461–464.

Serneels, S., Croux, C., Filzmoser, P., and Van Espen, P. J. (2005). 'Partial robust M-regression'. *Chemometrics and Intelligent Laboratory Systems* 79.1/2, pp. 55–64.

She, Y. and Owen, A. B. (2011). 'Outlier detection using nonconvex penalized regression'. *Journal of the American Statistical Association* 106.494, pp. 626–639.

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, pp. 807–832.

Smucler, E. and Yohai, V. J. (2017). 'Robust and sparse estimators for linear regression models'. *Computational Statistics and Data Analysis* 111, pp. 116–130.

Srna, S., Schrift, R. Y., and Zauberman, G. (2018). 'The illusion of multitasking and its positive effect on performance'. *Psychological Science* 29.12, pp. 1942–1955.

Stone, M. (1977). 'An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion'. *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 44–47.

Strohmaier, E., Dongarra, J., Simon, H., and Meuer, M. (2022). *Performance Development / TOP500*. URL: https://www.top500.org/statistics/perfdevel/ (visited on 02/02/2022).

Student (1908). 'The probable error of a mean'. *Biometrika* 6.1, pp. 1–25.

Suggala, A. S., Bhatia, K., Ravikumar, P., and Jain, P. (2019). 'Adaptive hard thresholding for near-optimal consistent robust regression'. arXiv: 1903.08192.

Sun, Y. and Genton, M. G. (2011). 'Functional boxplots'. *Journal of Computational and Graphical Statistics* 20.2, pp. 316–334.

Takano, Y. and Miyashiro, R. (2020). 'Best subset selection via cross-validation criterion'. *TOP* 28, pp. 475–488.

Thompson, R. (2022). 'Robust subset selection'. *Computational Statistics and Data Analysis* 169, p. 107415.

Thompson, R., Forbes, C. S., MacEachern, S. N., and Peruggia, M. (2022). 'Familial inference'. arXiv: 2202.12540.

Thompson, R. and Vahid, F. (2022). 'Group selection and shrinkage: Structured sparsity for semiparametric models'. arXiv: 2105.12081.

Tibshirani, R. (1996). 'Regression shrinkage and selection via the lasso'. *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tseng, P. (2001). 'Convergence of a block coordinate descent method for nondifferentiable minimization'. *Journal of Optimization Theory and Applications* 109.3, pp. 475–494.

van de Geer, S. A. and Bühlmann, P. (2009). 'On the conditions used to prove oracle results for the lasso'. *Electronic Journal of Statistics* 3, pp. 1360–1392.

Wang, H. (2009). 'Forward regression for ultra-high dimensional variable screening'. *Journal of the American Statistical Association* 104.488, pp. 1512–1524.

Wang, H., Li, G., and Jiang, G. (2007). 'Robust regression shrinkage and consistent variable selection through the LAD-lasso'. *Journal of Business and Economic Statistics* 25.3, pp. 347–355.

Wang, X., Jiang, Y., Huang, M., and Zhang, H. (2013). 'Robust variable selection with exponential squared loss'. *Journal of the American Statistical Association* 108.502, pp. 632–643.

Welch, B. L. (1947). 'The generalization of 'Student's' problem when several different population variances are involved'. *Biometrika* 34.1/2, pp. 28–35.

Weng, C.-S. (1989). 'On a second-order asymptotic property of the Bayesian bootstrap mean'. *Annals of Statistics* 17.2, pp. 705–710.

Wilcoxon, F. (1945). 'Individual comparisons by ranking methods'. *Biometrics Bulletin* 1.6, pp. 80–83.

Yang, E., Lozano, A. C., and Aravkin, A. (2018). 'A general family of trimmed estimators for robust high-dimensional data analysis'. *Electronic Journal of Statistics* 12.2, pp. 3519–3553.

Yi, C. and Huang, J. (2017). 'Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression'. *Journal of Computational and Graphical Statistics* 26.3, pp. 547–557.

Yin, J., Mutiso, F., and Tian, L. (2021). 'Joint hypothesis testing of the area under the receiver operating characteristic curve and the Youden index'. *Pharmaceutical Statistics* 20.3, pp. 657–674.

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

Zhang, C.-H. (2010). 'Nearly unbiased variable selection under minimax concave penalty'. *Annals of Statistics* 38.2, pp. 894–942.

Zhang, Y., Zhu, J., Zhu, J., and Wang, X. (2023). 'A splicing approach to best subset of groups selection'. *INFORMS Journal on Computing* 35.1, pp. 104–119.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. (2014). 'Lower bounds on the performance of polynomial-time algorithms for sparse linear regression'. *Proceedings of the 27th Conference on Learning Theory*. Vol. 35, pp. 921–948.

Zhao, P. and Yu, B. (2006). 'On model selection consistency of lasso'. *Journal of Machine Learning Research* 7, pp. 2541–2563.

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, pp. 339–353.

Appendix A

Robust subset selection

A.1 Computational methods

A.1.1 Improved problem formulations

It is possible to reformulate the robust subsets program (2.9) to yield improved computational performance. Consider the following mixed-integer program:

$$
\begin{aligned}
\min_{\xi,\beta,\eta,s,z} \quad & \frac{1}{2}\sum_{i=1}^{n}(y_i-\xi_i)^2 \\
\text{s.t.} \quad & \xi = X\beta+\eta \\
& -\mathcal{M}_\xi \le \xi_i \le \mathcal{M}_\xi, \qquad i\in[n] \\
& s_j \in \{0,1\}, \qquad j\in[p] \\
& -\mathcal{M}_\beta \le \beta_j \le \mathcal{M}_\beta, \qquad j\in[p] \\
& -s_j\mathcal{M}_\beta \le \beta_j \le s_j\mathcal{M}_\beta, \qquad j\in[p] \\
& \sum_{j=1}^{p} s_j \le k \\
& z_i \in \{0,1\}, \qquad i\in[n] \\
& -\mathcal{M}_\eta \le \eta_i \le \mathcal{M}_\eta, \qquad i\in[n] \\
& -z_i\mathcal{M}_\eta \le \eta_i \le z_i\mathcal{M}_\eta, \qquad i\in[n] \\
& \sum_{i=1}^{n} z_i \le n-h,
\end{aligned}
\tag{A.1}
$$

where $\xi \in \mathbb{R}^n$ is an additional auxiliary variable. Observe that the objective functions in (2.9) and (A.1) differ only in the number of variables involved. Specifically, the objective in (A.1) is a function of $p+n$ variables, whereas the objective in (A.1) is a function of n variables only. In terms of computational performance, it is our experience that mixed-integer solvers respond better to (A.1) because it has fewer quadratic terms. The above formulation also adds additional structure to the mixed-integer program by bounding the ℓ_∞-norms of β and η by $\mathcal{M}_\beta$ and $\mathcal{M}_\eta$ via the bound constraints $-\mathcal{M}_\beta \le \beta_j \le \mathcal{M}_\beta$ and $-\mathcal{M}_\eta \le \eta_i \le \mathcal{M}_\eta$. The Big-M constraints imply these bounds. There are a number of other implied bounds that can also be added to the program. The reader is referred to Bertsimas et al. (2016) for further details.

A.1.2 Proof of Proposition 1

The proof proceeds along the lines of that for Proposition 6 and Theorem 3.1 in Bertsimas et al. (2016).

Proof. We begin by proving the first part of Proposition 1. Let $\hat{\beta}$ denote an update to any k-sparse β:

$$\hat{\beta} \in \mathrm{H}\left(\beta - \frac{1}{\bar{L}_\beta}\nabla_\beta f(\beta, \eta); k\right),$$

and take $\bar{L}_\beta \geq L_\beta$, an upper bound to the partial Lipschitz constant. Then, from Lemma 1, we have the following series of inequalities:

$$\begin{aligned}
f(\beta, \eta) &= Q(\beta, \beta) \\
&\geq \inf_{\|\tilde{\beta}\|_0 \leq k} Q(\tilde{\beta}, \beta) \\
&= \inf_{\|\tilde{\beta}\|_0 \leq k} \left(f(\beta, \eta) + \nabla_\beta f(\beta, \eta)^T(\tilde{\beta} - \beta) + \frac{1}{2}\bar{L}_\beta \|\tilde{\beta} - \beta\|_2^2 \right) \\
&= \inf_{\|\tilde{\beta}\|_0 \leq k} \left(f(\beta, \eta) - \frac{1}{2\bar{L}_\beta}\|\nabla_\beta f(\beta, \eta)\|_2^2 + \frac{1}{2}\bar{L}_\beta \left\| \tilde{\beta} - \left(\beta - \frac{1}{\bar{L}_\beta}\nabla_\beta f(\beta, \eta)\right)\right\|_2^2 \right) \\
&= f(\beta, \eta) - \frac{1}{2\bar{L}_\beta}\|\nabla_\beta f(\beta, \eta)\|_2^2 + \frac{1}{2}\bar{L}_\beta \left\| \hat{\beta} - \left(\beta - \frac{1}{\bar{L}_\beta}\nabla_\beta f(\beta, \eta)\right)\right\|_2^2 \\
&= f(\beta, \eta) + \nabla_\beta f(\beta, \eta)^T(\hat{\beta} - \beta) + \frac{1}{2}\bar{L}_\beta \|\hat{\beta} - \beta\|_2^2 \\
&= f(\beta, \eta) + \nabla_\beta f(\beta, \eta)^T(\hat{\beta} - \beta) + \frac{1}{2}L_\beta \|\hat{\beta} - \beta\|_2^2 + \frac{1}{2}(\bar{L}_\beta - L_\beta)\|\hat{\beta} - \beta\|_2^2 \\
&\geq f(\hat{\beta}, \eta) + \frac{1}{2}(\bar{L}_\beta - L_\beta)\|\hat{\beta} - \beta\|_2^2.
\end{aligned}$$

Taking $\beta = \beta^{(m)}$, $\hat{\beta} = \beta^{(m+1)}$, and $\eta = \eta^{(m)}$, it follows that

$$f(\beta^{(m)}, \eta^{(m)}) - f(\beta^{(m+1)}, \eta^{(m)}) \geq \frac{1}{2}(\bar{L}_\beta - L_\beta)\|\beta^{(m+1)} - \beta^{(m)}\|_2^2. \tag{A.2}$$

Similarly, letting $\hat{\eta}$ denote an update to any $(n-h)$-sparse η:

$$\hat{\eta} \in \mathrm{H}\left(\eta - \frac{1}{\bar{L}_\eta}\nabla_\eta f(\beta, \eta); n - h\right),$$

and applying Lemma 1 with $\eta = \eta^{(m)}$, $\hat{\eta} = \eta^{(m+1)}$, and $\beta = \beta^{(m+1)}$, we obtain

$$f(\beta^{(m+1)}, \eta^{(m)}) - f(\beta^{(m+1)}, \eta^{(m+1)}) \geq \frac{1}{2}(\bar{L}_\eta - L_\eta)\|\eta^{(m+1)} - \eta^{(m)}\|_2^2. \tag{A.3}$$

Adding together (A.2) and (A.3) yields

$$\begin{aligned}
&f(\beta^{(m)}, \eta^{(m)}) - f(\beta^{(m+1)}, \eta^{(m+1)}) \\
&\qquad \geq \frac{1}{2}(\bar{L}_\beta - L_\beta)\|\beta^{(m+1)} - \beta^{(m)}\|_2^2 + \frac{1}{2}(\bar{L}_\eta - L_\eta)\|\eta^{(m+1)} - \eta^{(m)}\|_2^2.
\end{aligned} \tag{A.4}$$

Hence, the sequence $\{f(\beta^{(m)}, \eta^{(m)})\}$ is decreasing, and because $f(\beta, \eta)$ is bounded below by zero, it follows from the monotone convergence theorem that the sequence converges.

For the second part of Proposition 1, we take the sum of (A.4) over $1 \leq m \leq M$ to obtain

$$\begin{aligned}&\sum_{m=1}^{M} \left(f(\beta^{(m)}, \eta^{(m)}) - f(\beta^{(m+1)}, \eta^{(m+1)}) \right) \\ &\quad \geq \frac{1}{2} \sum_{m=1}^{M} \left((\bar{L}_\beta - L_\beta) \|\beta^{(m+1)} - \beta^{(m)}\|_2^2 + (\bar{L}_\eta - L_\eta) \|\eta^{(m+1)} - \eta^{(m)}\|_2^2 \right).\end{aligned} \tag{A.5}$$

The inequality (A.5) implies that

$$\begin{aligned}&f(\beta^{(1)}, \eta^{(1)}) - f(\beta^{(M+1)}, \eta^{(M+1)}) \\ &\quad \geq \frac{M}{2} \min_{1 \leq m \leq M} \left((\bar{L}_\beta - L_\beta) \|\beta^{(m+1)} - \beta^{(m)}\|_2^2 + (\bar{L}_\eta - L_\eta) \|\eta^{(m+1)} - \eta^{(m)}\|_2^2 \right) \\ &\quad \geq \frac{M}{2} \min(\bar{L}_\beta - L_\beta, \bar{L}_\eta - L_\eta) \min_{1 \leq m \leq M} \left(\|\beta^{(m+1)} - \beta^{(m)}\|_2^2 + \|\eta^{(m+1)} - \eta^{(m)}\|_2^2 \right).\end{aligned}$$

Because $\{f(\beta^{(m)}, \eta^{(m)})\}$ is decreasing and converges to $f(\beta^\star, \eta^\star)$, it follows that

$$\begin{aligned}\min_{1 \leq m \leq M} \left(\|\beta^{(m+1)} - \beta^{(m)}\|_2^2 + \|\eta^{(m+1)} - \eta^{(m)}\|_2^2 \right) &\leq 2 \frac{f(\beta^{(1)}, \eta^{(1)}) - f(\beta^{(M+1)}, \eta^{(M+1)})}{M \min(\bar{L}_\beta - L_\beta, \bar{L}_\eta - L_\eta)} \\ &\leq 2 \frac{f(\beta^{(1)}, \eta^{(1)}) - f(\beta^\star, \eta^\star)}{M \min(\bar{L}_\beta - L_\beta, \bar{L}_\eta - L_\eta)},\end{aligned}$$

with the final inequality that which we set out to obtain. □

A.2 Breakdown point

A.2.1 Proof of Theorem 1

The proof below follows steps similar to those used in the proof of the breakdown point in Bertsimas and Mazumder (2014) for the objective value of the least quantile of squares estimator. We use the following standard result in the proof.

Lemma 5. *Let $\Theta(X, Y)$ be the optimal objective value to the robust subset selection problem (2.2). Then $\Theta(X, Y)$ satisfies the equality*

$$\Theta(X, Y) = \min_{I \in \mathcal{I}} \min_{\beta \in \mathrm{B}} \frac{1}{2} \sum_{i \in I} (y_i - x_i^T \beta)^2,$$

where

$$\mathcal{I} = \{I \subseteq [n] : |I| \geq h\} \quad \textit{and} \quad \mathrm{B} = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq k\}.$$

We are now ready to prove Theorem 1.

Proof. We proceed by completing the proof of Theorem 1 in two parts, showing that the inequalities $b(\Theta; X, Y) > (n-h)/n$ and $b(\Theta; X, Y) \leq (n-h+1)/n$ both hold. The former inequality is proven first. Suppose that exactly $m = n - h$ observations of the original sample (X, Y) are arbitrarily contaminated, and denote this new contaminated sample

$(\tilde{X}, \tilde{Y})$. Let I^0 contain only the indices of the uncontaminated observations. Because $I^0 \in \mathcal{I}$, it follows from Lemma 5 that

$$\Theta(\tilde{X}, \tilde{Y}) = \min_{I \in \mathcal{I}} \min_{\beta \in \mathrm{B}} \frac{1}{2} \sum_{i \in I} (\tilde{y}_i - \tilde{x}_i^T \beta)^2 \leq \min_{\beta \in \mathrm{B}} \frac{1}{2} \sum_{i \in I^0} (\tilde{y}_i - \tilde{x}_i^T \beta)^2. \tag{A.6}$$

The right-hand side of (A.6) does not depend on any contaminated observations and is finite. Thus, the breakdown point is strictly larger than $(n-h)/n$. Suppose that one additional observation is arbitrarily contaminated such that $m = n - h + 1$. Therefore, every $I \in \mathcal{I}$ includes a contaminated observation, say the observation indexed by c. Let $I^\star$ and $\beta^\star$ denote an optimal solution to the robust subsets problem (2.2). Then the optimal objective value is lower bounded as

$$\Theta(\tilde{X}, \tilde{Y}) = \frac{1}{2} \sum_{i \in I^\star} (\tilde{y}_i - \tilde{x}_i^T \beta^\star)^2 \geq \frac{1}{2} (\tilde{y}_c - \tilde{x}_c^T \beta^\star)^2. \tag{A.7}$$

The right-hand side of (A.7) can be made arbitrarily large because $\tilde{y}_c$ can be made arbitrarily large. Thus, the breakdown point is less than or equal to $(n-h+1)/n$. We conclude that $b(\Theta; X, Y) = (n-h+1)/n$. □

A.3 Experiments

A.3.1 Comparisons of estimators

The results of Section 2.5.1 for $p_0 = 10$ nonzero coefficients are reported in Figures A.1, A.2, and A.3. The findings in the low-dimensional setup are broadly consistent with those for $p_0 = 5$, while the high-dimensional setup proves more formidable. Robust subsets maintains superior support recovery when either Y or X are contaminated and performs within statistical precision when both are contaminated. However, when both are contaminated, none of the estimators offer more than marginal improvement in prediction over the null model, even when SNR $= 9$.

A.3.2 Comparisons of algorithms

The results of Section 2.5.2 for $p = 1,000$ predictors are reported in Table A.1. When Y and X are contaminated, and in the absence of warm start information, the solution from the solver is typically low-quality. The combination of heuristics and mixed-integer optimisation continues to produce the best outcomes across the board. In both contamination settings, setting $\tau = 1$ produces optimality gaps that are far superior to those from setting $\tau = 1.5$. As with $p = 500$, using the smaller value of $\tau = 1$ does not harm the number of true positive selections or the objective gap.

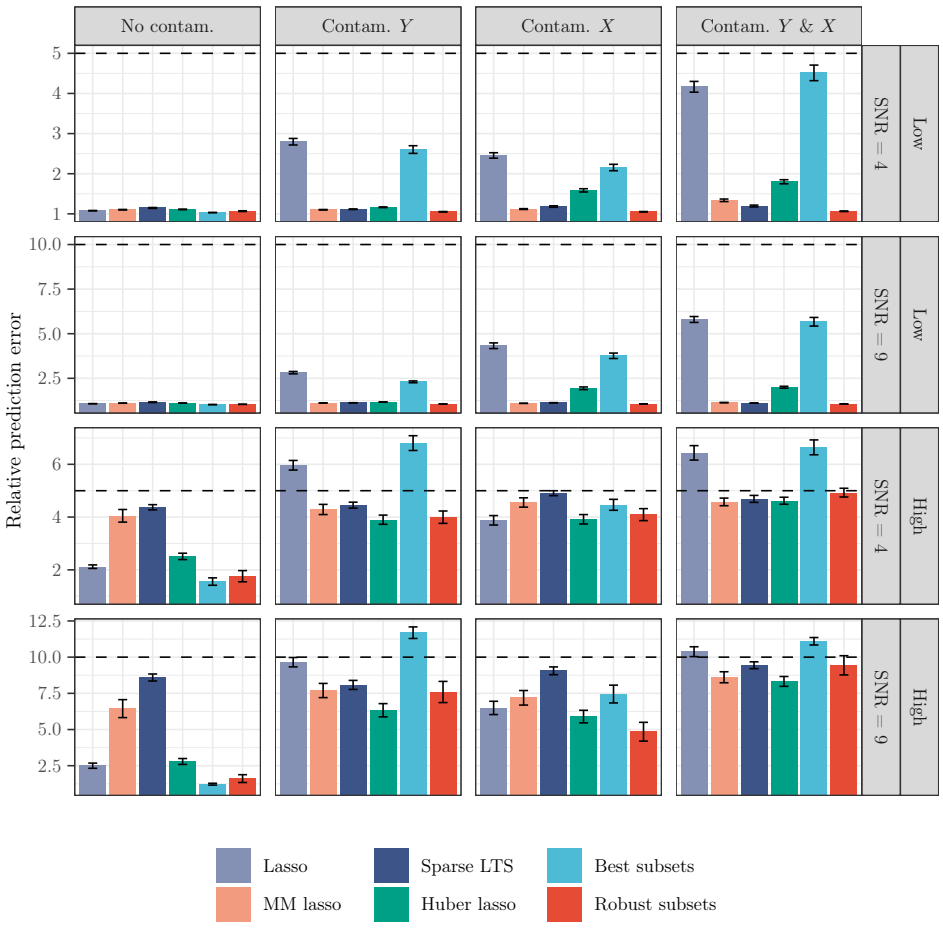

Figure A.1: Relative prediction error estimated over 30 simulations with $p_0 = 10$. The vertical bars represent averages, and the error bars denote (one) standard errors. The dashed horizontal lines indicate the relative prediction error from the null model.

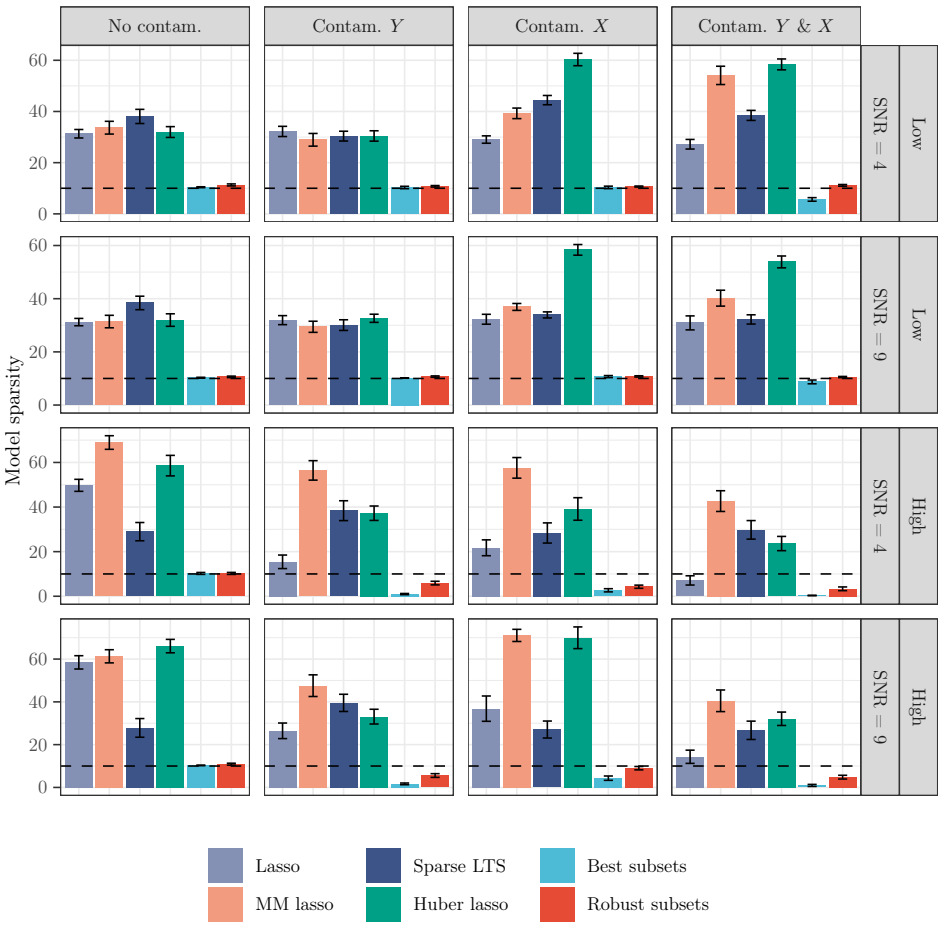

Figure A.2: Model sparsity estimated over 30 simulations with $p_0 = 10$. The vertical bars represent averages, and the error bars denote (one) standard errors. The dashed horizontal lines indicate the true model sparsity.

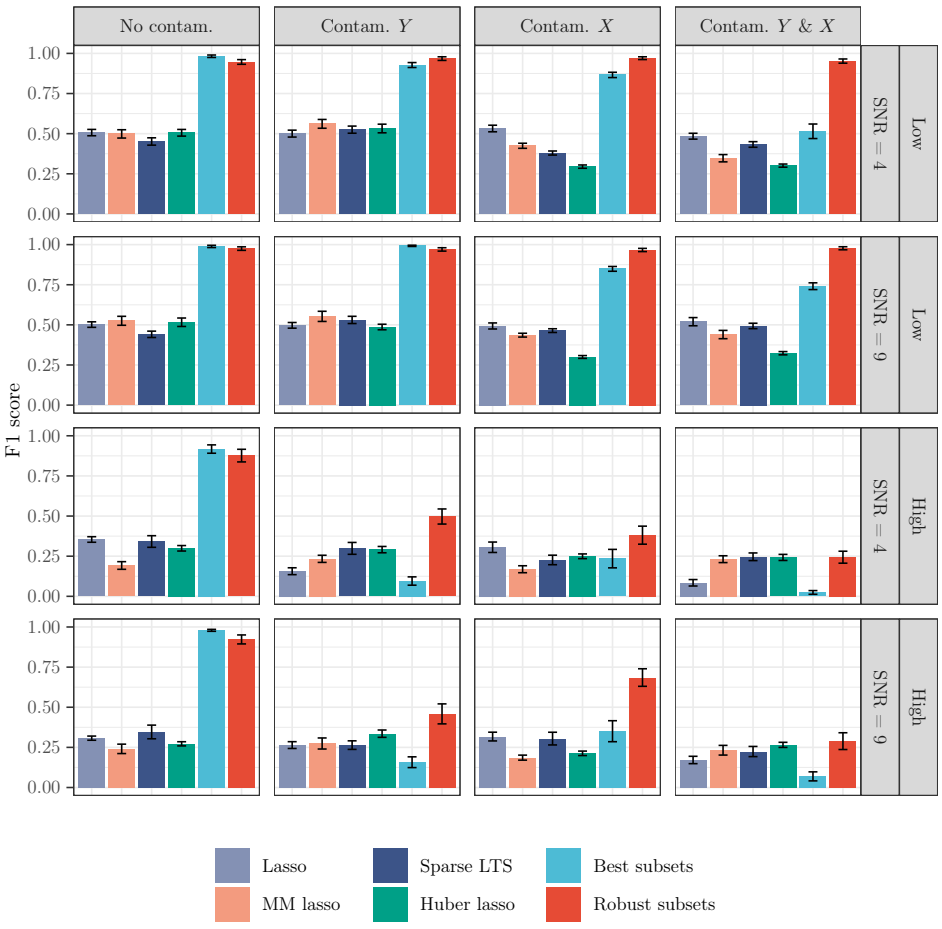

Figure A.3: F1 score estimated over 30 simulations with $p_0 = 10$. The vertical bars represent averages, and the error bars denote (one) standard errors.

	True pos.	Obj. gap (%)	Opt. gap (%)	Term. (%)	Time (min.)
Contamination of Y					
Heuristics	5.0 (0.0)	0.2 (0.2)	-	-	1.3 (0.1)
MIO	5.0 (0.0)	0.0 (0.0)	100.0 (0.0)	0.0 (0.0)	30.0 (0.0)
MIO+heur. (1)	5.0 (0.0)	0.0 (0.0)	0.7 (0.7)	96.7 (3.3)	3.7 (1.0)
MIO+heur. (1.5)	5.0 (0.0)	0.0 (0.0)	71.1 (5.4)	10.0 (5.5)	30.5 (0.6)
Contamination of Y and X					
Heuristics	4.9 (0.1)	1.8 (1.3)	-	-	7.8 (0.4)
MIO	3.8 (0.3)	30.5 (8.7)	100.0 (0.0)	0.0 (0.0)	30.0 (0.0)
MIO+heur. (1)	5.0 (0.0)	0.0 (0.0)	68.8 (3.8)	3.3 (3.3)	37.3 (0.7)
MIO+heur. (1.5)	5.0 (0.0)	0.0 (0.0)	100.0 (0.0)	0.0 (0.0)	37.9 (0.4)

Table A.1: True positive selections, relative objective gap, relative optimality gap, termination rate, and runtime estimated over 30 simulations with $n = 100$, $p = 1,000$, $p_0 = 5$, and SNR $= 4$. Averages or proportions are reported next to (one) standard errors in parentheses.

Appendix B

Group selection and shrinkage: Structured sparsity for semiparametric models

B.1 Computation

B.1.1 Proof of Proposition 2

Proof. The subscript k is dropped from c_k, λ_{0k}, and λ_{1k} to simplify the notation. Since the objective is treated as a function in the kth group of coordinates $\boldsymbol{\nu}_k$ only, we have

$$\begin{aligned}\bar{F}_c(\boldsymbol{\nu};\tilde{\boldsymbol{\nu}}) &\propto \nabla_k L(\tilde{\boldsymbol{\nu}})^\top(\boldsymbol{\nu}_k - \tilde{\boldsymbol{\nu}}_k) + \frac{c}{2}\|\boldsymbol{\nu}_k - \tilde{\boldsymbol{\nu}}_k\|^2 + \lambda_0 1(\|\boldsymbol{\nu}_k\| \neq 0) + \lambda_1\|\boldsymbol{\nu}_k\| \\ &\propto \frac{c}{2}\left\|\boldsymbol{\nu}_k - \left(\tilde{\boldsymbol{\nu}}_k - \frac{1}{c}\nabla_k L(\tilde{\boldsymbol{\nu}})\right)\right\|^2 + \lambda_0 1(\|\boldsymbol{\nu}_k\| \neq 0) + \lambda_1\|\boldsymbol{\nu}_k\| \\ &= \frac{c}{2}\|\boldsymbol{\nu}_k - \hat{\boldsymbol{\nu}}_k\|^2 + \lambda_0 1(\|\boldsymbol{\nu}_k\| \neq 0) + \lambda_1\|\boldsymbol{\nu}_k\|,\end{aligned}$$

where $\hat{\boldsymbol{\nu}}_k = \tilde{\boldsymbol{\nu}}_k - 1/c\nabla_k L(\tilde{\boldsymbol{\nu}})$. When $\lambda_1 = 0$, it is not hard to see a minimiser of $\bar{F}_c(\boldsymbol{\nu};\tilde{\boldsymbol{\nu}})$ is

$$\boldsymbol{\nu}_k^\star = \begin{cases} \hat{\boldsymbol{\nu}}_k & \text{if } \|\hat{\boldsymbol{\nu}}_k\| \geq \sqrt{\dfrac{2\lambda_0}{c}} \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{B.1}$$

When $\lambda_0 = 0$ and $\lambda_1 > 0$, the minimiser is

$$\boldsymbol{\nu}_k^\star = \begin{cases} \left(1 - \dfrac{\lambda_1}{c\|\hat{\boldsymbol{\nu}}_k\|}\right)_+ \hat{\boldsymbol{\nu}}_k & \text{if } \left(1 - \dfrac{\lambda_1}{c\|\hat{\boldsymbol{\nu}}_k\|}\right)_+ \|\hat{\boldsymbol{\nu}}_k\| \geq 0 \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{B.2}$$

This expression follows from the proximal operator for the l_2-norm (Beck 2017). Combining (B.1) with (B.2) leads to the result of the proposition. □

B.1.2 Proof of Lemma 3

Proof. Denote by $\boldsymbol{\nu}^\star$ the result of applying the thresholding function (3.7) to $\tilde{\boldsymbol{\nu}}$. Starting from the inequality (3.4) with $\boldsymbol{\nu} = \boldsymbol{\nu}^\star$, we add $R(\boldsymbol{\nu}^\star)$ to both sides to obtain

$$F(\boldsymbol{\nu}^\star) \leq L(\tilde{\boldsymbol{\nu}}) + \nabla_k L(\tilde{\boldsymbol{\nu}})^\top(\boldsymbol{\nu}_k^\star - \tilde{\boldsymbol{\nu}}_k) + \frac{c_k}{2}\|\boldsymbol{\nu}_k^\star - \tilde{\boldsymbol{\nu}}_k\|^2 + R(\boldsymbol{\nu}^\star),$$

where the left-hand side follows from definition (3.5). Adding $\bar{c}_k/2\|\boldsymbol{\nu}_k^\star - \tilde{\boldsymbol{\nu}}_k\|^2$ to both sides and rearranging terms leads to

$$\begin{aligned} F(\boldsymbol{\nu}^\star) &\leq L(\tilde{\boldsymbol{\nu}}) + \nabla_k L(\tilde{\boldsymbol{\nu}})^\top(\boldsymbol{\nu}_k^\star - \tilde{\boldsymbol{\nu}}_k) + \frac{\bar{c}_k}{2}\|\boldsymbol{\nu}_k^\star - \tilde{\boldsymbol{\nu}}_k\|^2 + R(\boldsymbol{\nu}^\star) + \frac{c_k - \bar{c}_k}{2}\|\boldsymbol{\nu}_k^\star - \tilde{\boldsymbol{\nu}}_k\|^2 \\ &= \bar{F}_{\bar{c}_k}(\boldsymbol{\nu}^\star; \tilde{\boldsymbol{\nu}}) + \frac{c_k - \bar{c}_k}{2}\|\boldsymbol{\nu}_k^\star - \tilde{\boldsymbol{\nu}}_k\|^2. \end{aligned}$$

Using $\bar{F}_{\bar{c}_k}(\boldsymbol{\nu}^\star; \tilde{\boldsymbol{\nu}}) \leq \bar{F}_{\bar{c}_k}(\tilde{\boldsymbol{\nu}}; \tilde{\boldsymbol{\nu}}) = F(\tilde{\boldsymbol{\nu}})$, we reorganise terms to get

$$F(\tilde{\boldsymbol{\nu}}) - F(\boldsymbol{\nu}^\star) \geq \frac{\bar{c}_k - c_k}{2}\|\boldsymbol{\nu}_k^\star - \tilde{\boldsymbol{\nu}}_k\|^2. \tag{B.3}$$

Now, define the vector

$$\boldsymbol{\eta}_k^{(m)} := \begin{cases} (\boldsymbol{\nu}_1^{(m+1)\top}, \ldots, \boldsymbol{\nu}_k^{(m+1)\top}, \boldsymbol{\nu}_{k+1}^{(m)\top}, \ldots, \boldsymbol{\nu}_g^{(m)\top})^\top & \text{if } k > 0 \\ \boldsymbol{\nu}^{(m)} & \text{otherwise.} \end{cases}$$

Take $\tilde{\boldsymbol{\nu}} = \boldsymbol{\eta}_{k-1}^{(m)}$ and $\boldsymbol{\nu}^\star = \boldsymbol{\eta}_k^{(m)}$ and sum both sides of the inequality (B.3) over $1 \leq k \leq g$ to get

$$\sum_{k=1}^{g}[F(\boldsymbol{\eta}_{k-1}^{(m)}) - F(\boldsymbol{\eta}_k^{(m)})] = F(\boldsymbol{\eta}_0^{(m)}) - F(\boldsymbol{\eta}_g^{(m)}) \geq \sum_{k=1}^{g}\frac{\bar{c}_k - c_k}{2}\|\boldsymbol{\nu}_k^{(m+1)} - \boldsymbol{\nu}_k^{(m)}\|^2.$$

By definition $\boldsymbol{\eta}_0^{(m)} = \boldsymbol{\nu}^{(m)}$ and $\boldsymbol{\eta}_g^{(m)} = \boldsymbol{\nu}^{(m+1)}$, establishing $\{F(\boldsymbol{\nu}^{(m)})\}_{m\in\mathbb{N}}$ is decreasing. Since $F(\boldsymbol{\nu})$ is bounded below, $\{F(\boldsymbol{\nu}^{(m)})\}_{m\in\mathbb{N}}$ must converge. □

B.1.3 Proof of Theorem 2

The proof relies on the following lemma that the active set must stabilise in finitely many iterations. The lemma is established by contradiction along the lines of Dedieu et al. (2021, Theorem 1).

Lemma 6. *Let $\bar{c}_k > c_k$ for all $k = 1, \ldots, g$. Then the sequence of iterates $\{\boldsymbol{\nu}^{(m)}\}_{m\in\mathbb{N}}$ stabilises to a fixed support within a finite number of iterations.*

Proof. Suppose the support does not stabilise in finitely many iterations. Choose an m such that $\text{gs}(\boldsymbol{\nu}^{(m+1)}) \neq \text{gs}(\boldsymbol{\nu}^{(m)})$. Then at least one group was added or removed from the support, i.e., there is a k such that either (1) $\boldsymbol{\nu}_k^{(m)} = \mathbf{0}$ and $\boldsymbol{\nu}_k^{(m+1)} \neq \mathbf{0}$ or (2) $\boldsymbol{\nu}_k^{(m)} \neq \mathbf{0}$ and $\boldsymbol{\nu}_k^{(m+1)} = \mathbf{0}$. Consider case (1). It follows from Lemma 3

$$F(\boldsymbol{\nu}^{(m)}) - F(\boldsymbol{\nu}^{(m+1)}) \geq \frac{\bar{c}_k - c_k}{2}\|\boldsymbol{\nu}_k^{(m+1)}\|^2,$$

and, because $\boldsymbol{\nu}_k^{(m+1)}$ is the output of the thresholding function (3.7), it holds $\boldsymbol{\nu}_k^{(m+1)} \geq \sqrt{2\lambda_{0k}/\bar{c}_k}$. These inequalities together imply

$$F(\boldsymbol{\nu}^{(m)}) - F(\boldsymbol{\nu}^{(m+1)}) \geq (\bar{c}_k - c_k)\frac{\lambda_{0k}}{\bar{c}_k}.$$

Similar working yields the same inequality for case (2). For $\bar{c}_k > c_k$, the quantity on the right-hand side is strictly positive. Hence, a change to the support yields a strict decrease in the objective value. However, if the support changes infinitely many times, this contradicts that $F(\boldsymbol{\nu})$ is bounded below. Thus, the support must stabilise in finitely many iterations. □

We are now ready to prove Theorem 2.

Proof. From Lemma 6, there exists a finite M such that the iterates of the subsequence $\{\boldsymbol{\nu}^{(m)}\}_{m \geq M}$ share the same active set, say $\mathcal{A}$. Hence, for all $m \geq M$ and $k \in \mathcal{A}$ we have

$$\bar{F}_{\bar{c}_k}(\boldsymbol{\nu}; \boldsymbol{\nu}^{(m)}) \propto \frac{\bar{c}_k}{2} \left\| \boldsymbol{\nu}_k - \left(\boldsymbol{\nu}_k^{(m)} - \frac{1}{\bar{c}_k} \nabla_k L(\boldsymbol{\nu}^{(m)}) \right) \right\|^2 + \lambda_{1k} \|\boldsymbol{\nu}_k\|.$$

Thus, the group subset penalty can be treated as fixed. Denote by $\nabla^2_{\mathbf{v}} \bar{F}_{\bar{c}_k}(\boldsymbol{\nu}; \boldsymbol{\nu}^{(m)})$ the second directional derivative of $\bar{F}_{\bar{c}_k}(\boldsymbol{\nu}; \boldsymbol{\nu}^{(m)})$ along the vector $\mathbf{v} \in \mathbb{R}^{\sum_{k=1}^g p_k}$. The infimum of the minimal eigenvalue of $\nabla^2_{\mathbf{v}} \bar{F}_{\bar{c}_k}(\boldsymbol{\nu}; \boldsymbol{\nu}^{(m)})$ over all $\boldsymbol{\nu}_k$ and $\mathbf{v}$ is $\bar{c}_k$. Since $\bar{c}_k > 0$, $\bar{F}_{\bar{c}_k}(\boldsymbol{\nu}; \boldsymbol{\nu}^{(m)})$ is strictly convex in $\boldsymbol{\nu}_k$. Furthermore, under the statement of the theorem, either (a) $\lambda_{1k} > 0$ for all $k = 1, \ldots, g$ (so the objective is coercive) or (b) no elements of $\boldsymbol{\nu}$ tend to $\pm\infty$. It follows then that the level set $\{\boldsymbol{\nu} \in \mathbb{R}^{\sum_{k=1}^g p_k} : F(\boldsymbol{\nu}) \leq F(\boldsymbol{\nu}^{(0)})\}$ is bounded when the initialisation $\boldsymbol{\nu}^{(0)} \in \mathbb{R}^{\sum_{k=1}^g p_k}$. Hence, by the descent property of Lemma 3, the sequence $\{\boldsymbol{\nu}^{(m)}\}_{m \in \mathbb{N}}$ is bounded and therefore has a limit point $\boldsymbol{\nu}^\star$. These conditions are sufficient to invoke Tseng (2001, Proposition 5.1) and establish $\boldsymbol{\nu}^\star$ is a stationary point of $\bar{F}_{\bar{c}_k}(\boldsymbol{\nu}; \boldsymbol{\nu}^\star)$. We conclude by the equality $\bar{F}_{\bar{c}_k}(\boldsymbol{\nu}^\star; \boldsymbol{\nu}^\star) = F(\boldsymbol{\nu}^\star)$ that $\boldsymbol{\nu}^\star$ is also a stationary point of $F(\boldsymbol{\nu})$. □

B.1.4 Proof of Proposition 3

Proof. Under the conditions of Theorem 2, Algorithm 3 is guaranteed to converge to a stationary point $\hat{\boldsymbol{\nu}}^{(t)}$ such that for all $k \notin \mathcal{A}^{(t)}$ it holds

$$\frac{\left(\|\nabla_k L(\hat{\boldsymbol{\nu}}^{(t)})\| - \lambda_{1k}\right)_+}{\bar{c}_k} < \sqrt{\frac{2\lambda_{0k}^{(t)}}{\bar{c}_k}} = \sqrt{\frac{2\lambda_0^{(t)} p_k}{\bar{c}_k}}.$$

Then initialising Algorithm 3 with $\hat{\boldsymbol{\nu}}^{(t)}$ and using $\lambda_0 = \lambda_0^{(t+1)}$ such that

$$\lambda_0^{(t+1)} < \max_{k \notin \mathcal{A}^{(t)}} \left(\frac{\left(\|\nabla_k L(\hat{\boldsymbol{\nu}}^{(t)})\| - \lambda_{1k}\right)_+^2}{2 p_k \bar{c}_k} \right)$$

leads to $\hat{\boldsymbol{\nu}}^{(t+1)} \neq \hat{\boldsymbol{\nu}}^{(t)}$. □

B.1.5 Other components

We briefly outline other algorithmic components of `grpsel`. These components are similar to those used in the coordinate descent literature (Friedman et al. 2007; Breheny and Huang 2011; Hazimeh and Mazumder 2020).

Gradient screening

Rather than cycling through all g groups in each coordinate descent round, it is convenient to restrict the updates to a smaller set of screened groups. The initialisation $\boldsymbol{\nu}^{(0)}$ can be used to compute the coordinate-wise gradients $\{\|\nabla_k L(\boldsymbol{\nu}^{(0)})\| / \sqrt{p_k}\}_{k \notin \mathcal{A}^{(0)}}$ which are already available as a consequence of selecting λ_0 dynamically (see Proposition 3). The inactive groups whose gradients are among the top 500 largest are classed as "strong," in addition to the active set of groups. The remaining groups are classed as "weak." The coordinate descent updates are restricted to the strong groups until convergence

is achieved, at which time a further round over the weak groups is performed. If the solution does not change after this further round, convergence is declared. Otherwise, any weak groups that have become active are shifted to the strong set, and the process is repeated.

Gradient screening is also used in local search. Rather than searching through all inactive groups in the inner loop of Algorithm 4, only the inactive groups whose gradients are among the largest 5% are enumerated.

Gradient ordering

The solutions produced by coordinate descent often benefit from greedily ordering the groups. At the beginning of the algorithm, the groups are sorted according to their gradients. Any groups in $\mathcal{A}^{(0)}$ are placed first. The coordinate descent updates then proceed using this new ordering.

Active set updates

The set of active groups typically stabilises after several rounds of coordinate descent updates. At this time, several additional rounds are required for the nonzero coefficients to converge. Rather than cycling through the full set (or screened set) of groups, the updates are restricted to the active groups only, usually a small subset. Once convergence is achieved on the active set, a further round is performed over the inactive set to confirm overall convergence.

B.2 Error bounds

B.2.1 Proof of Theorem 3

The proof requires the following lemma.

Lemma 7. *Let $\delta \in (0, 1]$. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a fixed matrix and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ be a $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ random vector. Define $\boldsymbol{\theta} := \sum_{k=1}^{g} \bar{\boldsymbol{\nu}}_k$ and the random event*

$$A_{\bar{\boldsymbol{\nu}}} := \left\{ |\boldsymbol{\varepsilon}^\top \mathbf{X}\boldsymbol{\theta}| \geq \|\mathbf{X}\boldsymbol{\theta}\| C_1 \sigma \sqrt{s p_{\max} + s \log\left(\frac{g}{s}\right) + \log(\delta^{-1})} \right\}.$$

Then, for some numerical constant $C_1 > 0$, the probability of the union $\cup_{\bar{\boldsymbol{\nu}} \in \mathcal{V}(2s)} A_{\bar{\boldsymbol{\nu}}}$ is at most δ.

Proof. For $\mathcal{A} \subseteq \{1, \dots, g\}$, a set of active groups, denote by $\mathcal{S}_\mathcal{A} := \cup_{k \in \mathcal{A}} \mathcal{G}_k$ the set of active predictors. Denote the singular value decomposition of $\mathbf{X}_\mathcal{A}$ by $\mathbf{U}_\mathcal{A} \mathbf{D}_\mathcal{A} \mathbf{V}_\mathcal{A}^\top$, where $\mathbf{X}_\mathcal{A}$ are the columns of $\mathbf{X}$ indexed by $\mathcal{S}_\mathcal{A}$. Define the set of unit vectors $\mathcal{B}_2^r := \{\mathbf{u} \in \mathbb{R}^r : \|\mathbf{u}\| \leq 1\}$, and the set of s group-sparse subsets $\mathcal{P}(s) := \{\mathcal{A} \subseteq \{1, \dots, g\} : |\mathcal{A}| = s\}$. For all $\bar{\boldsymbol{\nu}} \in \mathcal{V}(2s)$ such that $\boldsymbol{\theta} \neq \mathbf{0}$, it holds

$$\frac{|\boldsymbol{\varepsilon}^\top \mathbf{X}\boldsymbol{\theta}|}{\|\mathbf{X}\boldsymbol{\theta}\|} = \frac{|\boldsymbol{\varepsilon}^\top \mathbf{U}_\mathcal{A} \mathbf{D}_\mathcal{A} \mathbf{V}_\mathcal{A}^\top \boldsymbol{\theta}_\mathcal{A}|}{\|\mathbf{D}_\mathcal{A} \mathbf{V}_\mathcal{A}^\top \boldsymbol{\theta}_\mathcal{A}\|} \leq \max_{\mathcal{A} \in \mathcal{P}(2s)} \sup_{\mathbf{u} \in \mathcal{B}_2^{|\mathcal{S}_\mathcal{A}|}} |\boldsymbol{\varepsilon}^\top \mathbf{U}_\mathcal{A} \mathbf{u}|.$$

For any $t \in \mathbb{R}$, this inequality implies

$$\mathrm{P}\left(\cup_{\bar{\boldsymbol{\nu}} \in \mathcal{V}(2s)} \left\{ |\boldsymbol{\varepsilon}^\top \mathbf{X}\boldsymbol{\theta}| \geq \|\mathbf{X}\boldsymbol{\theta}\| t \right\} \right) \leq \mathrm{P}\left(\max_{\mathcal{A} \in \mathcal{P}(2s)} \sup_{\mathbf{u} \in \mathcal{B}_2^{|\mathcal{S}_\mathcal{A}|}} |\boldsymbol{\varepsilon}^\top \mathbf{U}_\mathcal{A} \mathbf{u}| \geq t \right).$$

Applying Boole's inequality to the right-hand side yields

$$\mathrm{P}\left(\max_{\mathcal{A}\in\mathcal{P}(2s)}\sup_{\mathbf{u}\in\mathcal{B}_2^{|\mathcal{S}_\mathcal{A}|}}|\boldsymbol{\varepsilon}^\top\mathbf{U}_\mathcal{A}\mathbf{u}|\geq t\right)\leq\sum_{\mathcal{A}\in\mathcal{P}(2s)}\mathrm{P}\left(\sup_{\mathbf{u}\in\mathcal{B}_2^{|\mathcal{S}_\mathcal{A}|}}|\boldsymbol{\varepsilon}^\top\mathbf{U}_\mathcal{A}\mathbf{u}|\geq t\right).$$

We bound the supremum over $\mathcal{B}_2^{|\mathcal{S}_\mathcal{A}|}$ using an ϵ-net argument. Let $\mathcal{E}^{|\mathcal{S}_\mathcal{A}|}$ be an ϵ-net of $\mathcal{B}_2^{|\mathcal{S}_\mathcal{A}|}$ with respect to l_2-norm that satisfies $|\mathcal{E}^{|\mathcal{S}_\mathcal{A}|}|\leq(3/\epsilon)^{|\mathcal{S}_\mathcal{A}|}$. Such an $\mathcal{E}^{|\mathcal{S}_\mathcal{A}|}$ is guaranteed to exist for $\epsilon\in(0,1)$ (Rigollet 2015, Lemma 1.18). Setting $\epsilon=1/2$, it holds for any $\mathcal{A}\in\mathcal{P}(2s)$ and any $\mathbf{z}\in\mathcal{E}^{|\mathcal{S}_\mathcal{A}|}$

$$\sup_{\mathbf{u}\in\mathcal{B}_2^{|\mathcal{S}_\mathcal{A}|}}|\boldsymbol{\varepsilon}^\top\mathbf{U}_\mathcal{A}\mathbf{u}|\leq 2\sup_{\mathbf{z}\in\mathcal{E}^{|\mathcal{S}_\mathcal{A}|}}|\boldsymbol{\varepsilon}^\top\mathbf{U}_\mathcal{A}\mathbf{z}|.$$

Applying Boole's inequality to this bound yields

$$\sum_{\mathcal{A}\in\mathcal{P}(2s)}\mathrm{P}\left(2\sup_{\mathbf{z}\in\mathcal{E}^{|\mathcal{S}_\mathcal{A}|}}|\boldsymbol{\varepsilon}^\top\mathbf{U}_\mathcal{A}\mathbf{z}|\geq t\right)\leq\sum_{\mathcal{A}\in\mathcal{P}(2s)}\sum_{\mathbf{z}\in\mathcal{E}^{|\mathcal{S}_\mathcal{A}|}}\mathrm{P}\left(2|\boldsymbol{\varepsilon}^\top\mathbf{U}_\mathcal{A}\mathbf{z}|\geq t\right).$$

The cardinality of $\mathcal{P}(2s)$ satisfies $|\mathcal{P}(2s)|=\binom{g}{2s}\leq(eg/(2s))^{2s}$. For any $\mathcal{A}\in\mathcal{P}(2s)$, the cardinality of $\mathcal{E}^{|\mathcal{S}_\mathcal{A}|}$ satisfies $|\mathcal{E}^{|\mathcal{S}_\mathcal{A}|}|\leq 6^{|\mathcal{S}_\mathcal{A}|}\leq 6^{2sp_{\max}}$. Since $\mathbf{U}_\mathcal{A}$ is orthonormal and $\mathbf{z}$ has unit length, the random variable $\boldsymbol{\varepsilon}^\top\mathbf{U}_\mathcal{A}\mathbf{z}\sim\mathcal{N}(0,\sigma^2)$. Using a standard Gaussian tail bound (Rigollet 2015, Lemma 1.4), we have

$$\mathrm{P}\left(2|\boldsymbol{\varepsilon}^\top\mathbf{U}_\mathcal{A}\mathbf{z}|\geq t\right)\leq 2\exp\left(-\frac{t^2}{8\sigma^2}\right).$$

It follows from the chain of inequalities above

$$\mathrm{P}\left(\cup_{\bar{\boldsymbol{\nu}}\in\mathcal{V}(2s)}\left\{|\boldsymbol{\varepsilon}^\top\mathbf{X}\boldsymbol{\theta}|\geq\|\mathbf{X}\boldsymbol{\theta}\|t\right\}\right)\leq 2\exp\left(-\frac{t^2}{8\sigma^2}+2sp_{\max}\log(6)+2s\log\left(\frac{eg}{2s}\right)\right).$$

Setting $t\geq\sqrt{8\sigma^2[\log(2)+2sp_{\max}\log(6)+2s\log(eg/(2s))+\log(\delta^{-1})]}$ concludes the proof. □

We are now ready to prove Theorem 3.

Proof. Take any $\bar{\boldsymbol{\nu}}\in\mathcal{V}(s)$ and any $\boldsymbol{\beta}\in\mathbb{R}^p$ such that $\boldsymbol{\beta}=\sum_{k=1}^g\bar{\boldsymbol{\nu}}_k$. Optimality of $\hat{\bar{\boldsymbol{\nu}}}$ and $\hat{\boldsymbol{\beta}}=\sum_{k=1}^g\hat{\bar{\boldsymbol{\nu}}}_k$ implies

$$\frac{1}{n}\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2\leq\frac{1}{n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2,$$

which, after some algebra, leads to

$$\frac{1}{n}\|\mathbf{f}^0-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2\leq\frac{1}{n}\|\mathbf{f}^0-\mathbf{X}\boldsymbol{\beta}\|^2+\frac{2}{n}|\boldsymbol{\varepsilon}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})|.$$

Observe $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\sum_{k=1}^g(\hat{\bar{\boldsymbol{\nu}}}_k-\bar{\boldsymbol{\nu}}_k)$, with at most $2s$ components of $(\hat{\bar{\boldsymbol{\nu}}}_1-\bar{\boldsymbol{\nu}}_1,\ldots,\hat{\bar{\boldsymbol{\nu}}}_g-\bar{\boldsymbol{\nu}}_g)$ not equal to $\mathbf{0}$. An application of Lemma 7 thus yields the high-probability upper bound

$$\begin{aligned}\frac{2}{n}|\boldsymbol{\varepsilon}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})|&\leq\frac{2}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|C_1\sigma\sqrt{sp_{\max}+s\log\left(\frac{g}{s}\right)+\log(\delta^{-1})}\\&\leq\frac{2}{n}\left(\|\mathbf{f}^0-\mathbf{X}\hat{\boldsymbol{\beta}}\|+\|\mathbf{f}^0-\mathbf{X}\boldsymbol{\beta}\|\right)C_1\sigma\sqrt{sp_{\max}+s\log\left(\frac{g}{s}\right)+\log(\delta^{-1})},\end{aligned}$$

where the last line follows from Minkowski's inequality for l_p-norms. Using Young's inequality ($2ab \leq \alpha a^2 + \alpha^{-1} b^2$ for $\alpha > 0$), the first term on the right-hand side is bounded as

$$\frac{2}{n}\|\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}\| C_1 \sigma \sqrt{s p_{\max} + s \log\left(\frac{g}{s}\right) + \log(\delta^{-1})} \\ \leq \frac{\alpha}{n}\|\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \frac{C_1^2 \sigma^2}{\alpha n}\left[s p_{\max} + s \log\left(\frac{g}{s}\right) + \log(\delta^{-1})\right].$$

A bound for the second term on the right-hand side follows similarly. Putting the results together and rearranging terms, we arrive at

$$\frac{1}{n}\|\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \leq \frac{1+\alpha}{(1-\alpha)n}\|\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{2C_1^2\sigma^2}{\alpha(1-\alpha)n}\left[s p_{\max} + s \log\left(\frac{g}{s}\right) + \log(\delta^{-1})\right],$$

holding with probability at least $1 - \delta$ for $\alpha \in (0, 1)$. Taking $C \geq 2C_1^2$ completes the proof. □

B.2.2 Proof of Theorem 4

The proof requires the following lemma.

Lemma 8. *Let $\delta \in (0, 1]$. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a fixed matrix and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ be a $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ random vector. Let γ_k be the maximal eigenvalue of $\mathbf{X}_k^\top \mathbf{X}_k / n$, where $\mathbf{X}_k$ is the submatrix of $\mathbf{X}$ corresponding to group k. Define the random event*

$$A_k = \left\{\|\mathbf{X}_k^\top \boldsymbol{\varepsilon}\| \geq \sqrt{n\gamma_k}\sigma\sqrt{p_k + 2\sqrt{p_k \log(g) + p_k \log(\delta^{-1})} + 2\log(g) + 2\log(\delta^{-1})}\right\}.$$

Then the probability of the union $\cup_k A_k$ is at most δ.

Proof. Denote the singular value decomposition of $\mathbf{X}_k$ by $\mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^\top$. Using the properties of the operator norm, it holds

$$\|\mathbf{X}_k^\top \boldsymbol{\varepsilon}\| = \|\mathbf{V}_k \mathbf{D}_k \mathbf{U}_k^\top \boldsymbol{\varepsilon}\| \leq \sqrt{n\gamma_k}\|\mathbf{U}_k^\top \boldsymbol{\varepsilon}\|.$$

For any $t \in \mathbb{R}$, this inequality implies

$$\mathrm{P}\left(\cup_k \left\{\|\mathbf{X}_k^\top \boldsymbol{\varepsilon}\| \geq \sqrt{n\gamma_k} t\right\}\right) \leq \mathrm{P}\left(\cup_k \left\{\|\mathbf{U}_k^\top \boldsymbol{\varepsilon}\| \geq t\right\}\right).$$

Applying Boole's inequality to the right-hand side yields

$$\mathrm{P}\left(\cup_k \left\{\|\mathbf{U}_k^\top \boldsymbol{\varepsilon}\| \geq t\right\}\right) \leq \sum_{k=1}^{g} \mathrm{P}\left(\|\mathbf{U}_k^\top \boldsymbol{\varepsilon}\| \geq t\right).$$

Since $\mathbf{U}_k$ is orthonormal, the random variable $\|\mathbf{U}_k^\top \boldsymbol{\varepsilon}\|^2 / \sigma^2 \sim \chi^2(p_k)$. Using a standard chi-squared tail bound (Laurent and Massart 2000, Lemma 1), we have for $t = p_k + \sqrt{2p_k x} + 2x$ and $x > 0$

$$\mathrm{P}\left(\|\mathbf{U}_k^\top \boldsymbol{\varepsilon}\| \geq \sigma\sqrt{t}\right) \leq \exp(-x).$$

It follows from the chain of inequalities above

$$\mathrm{P}\left(\cup_k \left\{\|\mathbf{X}_k^\top \boldsymbol{\varepsilon}\| \geq \sqrt{n\gamma_k}\sigma\sqrt{p_k + \sqrt{2p_k x} + 2x}\right\}\right) \leq \exp(-x + \log(g)).$$

Setting $x \geq \log(g) + \log(\delta^{-1})$ concludes the proof. □

We are now ready to prove Theorem 4.

Proof. For any $\bar{\boldsymbol{\nu}} \in \mathcal{V}(s)$ and any $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $\boldsymbol{\beta} = \sum_{k=1}^g \bar{\boldsymbol{\nu}}_k$, we have

$$\frac{1}{n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + 2\sum_{k=1}^g \lambda_k \|\hat{\boldsymbol{\nu}}_k\| \leq \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\sum_{k=1}^g \lambda_k \|\bar{\boldsymbol{\nu}}_k\|,$$

which leads to

$$\frac{1}{n}\|\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \leq \frac{1}{n}\|\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{2}{n}|\boldsymbol{\varepsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})| + 2\sum_{k=1}^g \lambda_k \left(\|\bar{\boldsymbol{\nu}}_k\| - \|\hat{\bar{\boldsymbol{\nu}}}_k\|\right). \tag{B.4}$$

The Cauchy-Schwarz inequality and Minkowski's inequality are applied in turn to get

$$\frac{2}{n}|\boldsymbol{\varepsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})| \leq \frac{2}{n}\sum_{k=1}^g \|\mathbf{X}_k^\top \boldsymbol{\varepsilon}\|\|\hat{\bar{\boldsymbol{\nu}}}_k - \bar{\boldsymbol{\nu}}_k\| \leq \frac{2}{n}\sum_{k=1}^g \|\mathbf{X}_k^\top \boldsymbol{\varepsilon}\|(\|\hat{\bar{\boldsymbol{\nu}}}_k\| + \|\bar{\boldsymbol{\nu}}_k\|).$$

Applying Lemma 8, and using the assumed lower bound on λ_k, yields

$$\frac{2}{n}\sum_{k=1}^g \|\mathbf{X}_k^\top \boldsymbol{\varepsilon}\|(\|\hat{\bar{\boldsymbol{\nu}}}_k\| + \|\bar{\boldsymbol{\nu}}_k\|) \leq 2\sum_{k=1}^g \lambda_k(\|\hat{\bar{\boldsymbol{\nu}}}_k\| + \|\bar{\boldsymbol{\nu}}_k\|)$$

with high-probability. Plugging this bound into (B.4), we arrive at

$$\frac{1}{n}\|\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \leq \frac{1}{n}\|\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}\|^2 + 4\sum_{k=1}^g \lambda_k \|\bar{\boldsymbol{\nu}}_k\|,$$

holding with probability at least $1 - \delta$. □

B.2.3 Proof of Theorem 5

Proof. Begin with inequality (B.4). First, we bound the term $2/n|\boldsymbol{\varepsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|$. Lemma 7 gives the high-probability upper bound

$$\begin{aligned}\frac{2}{n}|\boldsymbol{\varepsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})| &\leq \frac{2}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\| C_1 \sigma \sqrt{s p_{\max} + s\log\left(\frac{g}{s}\right) + \log(\delta^{-1})} \\ &\leq \frac{2}{n}\left(\|\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}\| + \|\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}\|\right) C_1 \sigma \sqrt{s p_{\max} + s\log\left(\frac{g}{s}\right) + \log(\delta^{-1})}.\end{aligned}$$

Using Young's inequality ($2ab \leq \alpha/2a^2 + 2/\alpha b^2$), the first term on the right-hand side is bounded as

$$\begin{aligned}&\frac{2}{n}\|\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}\| C_1 \sigma \sqrt{s p_{\max} + s\log\left(\frac{g}{s}\right) + \log(\delta^{-1})} \\ &\qquad \leq \frac{\alpha}{2n}\|\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \frac{2C_1^2\sigma^2}{\alpha n}\left[s p_{\max} + s\log\left(\frac{g}{s}\right) + \log(\delta^{-1})\right].\end{aligned}$$

The second term on the right-hand side is bounded similarly. Now, we bound the remaining term $2\sum_{k=1}^g \lambda_k(\|\bar{\boldsymbol{\nu}}_k\| - \|\hat{\bar{\boldsymbol{\nu}}}_k\|)$ in (B.4). Minkowski's inequality and Assumption 1 give

$$2\sum_{k=1}^g \lambda_k(\|\bar{\boldsymbol{\nu}}_k\| - \|\hat{\bar{\boldsymbol{\nu}}}_k\|) \leq 2\lambda_{\max}\sum_{k=1}^g \|\bar{\boldsymbol{\nu}}_k - \hat{\bar{\boldsymbol{\nu}}}_k\| \leq \frac{2\lambda_{\max}}{\sqrt{n}\phi(2s)}\|\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|.$$

Using Minkowski's inequality again, we have

$$\frac{2\lambda_{\max}}{\sqrt{n}\phi(2s)}\|\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\| \leq \frac{2\lambda_{\max}}{\sqrt{n}\phi(2s)}\left(\|\mathbf{f}^0-\mathbf{X}\hat{\boldsymbol{\beta}}\|+\|\mathbf{f}^0-\mathbf{X}\boldsymbol{\beta}\|\right).$$

Two applications of Young's inequality yields

$$\frac{2\lambda_{\max}}{\sqrt{n}\phi(2s)}\left(\|\mathbf{f}^0-\mathbf{X}\hat{\boldsymbol{\beta}}\|+\|\mathbf{f}^0-\mathbf{X}\boldsymbol{\beta}\|\right) \leq \frac{4\lambda_{\max}^2}{\alpha\phi(2s)^2}+\frac{\alpha}{2n}\|\mathbf{f}^0-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2+\frac{\alpha}{2n}\|\mathbf{f}^0-\mathbf{X}\boldsymbol{\beta}\|^2.$$

Finally, putting the bounds together and simplifying the resulting expression, we have

$$\frac{1}{n}\|\mathbf{f}^0-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \leq \frac{1+\alpha}{(1-\alpha)n}\|\mathbf{f}^0-\mathbf{X}\boldsymbol{\beta}\|^2+\frac{4C_1^2\sigma^2}{\alpha(1-\alpha)n}\left[sp_{\max}+s\log\left(\frac{g}{s}\right)+\log(\delta^{-1})\right] + \frac{4\lambda_{\max}^2}{\alpha(1-\alpha)\phi(2s)^2},$$

holding with probability at least $1-\delta$ for $\alpha \in (0,1)$. □

B.3 Data analyses

B.3.1 Macroeconomic data preprocessing

All series are made stationary using standard transformations given in McCracken and Ng (2016). Some series contain missing observations and outliers, which are also treated as missing. These missing values are imputed using the `na_kalman` function of the `R` package `imputeTS`. An observation x_i is treated as an outlier if $|x_i - Q_2|/(Q_3 - Q_1) > 4.5$ where Q_1, Q_2, and Q_3 are the respective quartiles of the data.

Appendix C

Familial inference

C.1 Huber family

C.1.1 Proof of Lemma 4

Proof. A sufficient condition for the result of the lemma is for $\lambda - |x_0 - \mu(\lambda)|$ to be increasing as a function of λ. To establish the function is increasing, consider its gradient:

$$\frac{\partial}{\partial \lambda}\left(\lambda - |x_0 - \mu(\lambda)|\right) = 1 + \operatorname{sign}(x_0 - \mu(\lambda))\frac{\partial \mu(\lambda)}{\partial \lambda}. \tag{C.1}$$

A sufficient condition for the gradient to be positive, and hence for $\lambda - |x_0 - \mu(\lambda)|$ to be increasing, is that $|\partial\mu(\lambda)/\partial\lambda| < 1$. The first-order condition for optimality of $\mu(\lambda)$ is

$$\mathcal{L}'(\mu(\lambda)) = -\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i(x_i - \mu(\lambda)) - \sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i\lambda \operatorname{sign}(x_i - \mu(\lambda)) = 0,$$

which gives

$$-\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i\lambda \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i(x_i - \mu(\lambda))} = 1. \tag{C.2}$$

Using the bound $|x_i - \mu(\lambda)| < \lambda$ in the denominator of (C.2) yields

$$\begin{aligned}
-\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i\lambda \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i(x_i - \mu(\lambda))} &> \left|-\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i\lambda \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i\lambda}\right| \\
&= \left|-\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i}\right| \\
&= \left|\frac{\partial \mu(\lambda)}{\partial \lambda}\right|.
\end{aligned}$$

Together with (C.2), the above bound shows $|\partial\mu(\lambda)/\partial\lambda| < 1$, and hence the gradient (C.1) is positive. We conclude $\lambda - |x_0 - \mu(\lambda)|$ is increasing, thereby establishing the result of the lemma. □

C.1.2 Proof of Proposition 4

The proof of Proposition 4 requires the following lemma.

Lemma 9. *Let $s_i := \operatorname{sign}(x_i - \tilde{\mu})$, where $\tilde{\mu}$ is the weighted median. Suppose sample point x_0 satisfies $|x_0 - \mu(\lambda^\star)| \geq \lambda^\star$ for some $\lambda^\star > 0$. Then $\operatorname{sign}(x_0 - \mu(\lambda^\star)) = s_0$.*

Proof. We proceed using proof by contradiction and suppose $\operatorname{sign}(x_0 - \mu(\lambda^\star)) \neq s_0$. This event can only occur if there exists a $0 < \lambda < \lambda^\star$ such that $|x_0 - \mu(\lambda)| < \lambda$, since for the sign of $x_0 - \mu(\lambda)$ to change the residual must cross through zero. But the existence of such a λ contradicts Lemma 4 since $|x_0 - \mu(\lambda^\star)| \geq \lambda^\star$. Hence, it must be the case that $\operatorname{sign}(x_0 - \mu(\lambda^\star)) = \operatorname{sign}(x_0 - \mu(\lambda))$ for all $0 < \lambda < \lambda^\star$. The result of the lemma immediately follows from the fact that $\lim_{\lambda \to 0} \mu(\lambda) = \tilde{\mu}$. □

We are now ready to prove Proposition 4.

Proof. By equation (4.5), $\gamma = \lambda - \lambda^+$. Since (λ^+, μ^+) is a knot point, one or more sample points cross from the square piece of the Huber function to the absolute piece and satisfy $|x_i - \mu^+| = \lambda^+$. Among all sample points eligible to cross (i.e., all i satisfying $|x_i - \mu| < \lambda$), those with with maximal absolute deviation from μ^+ cross:

$$\lambda^+ = \max_{i:|x_i - \mu| < \lambda} \left(|x_i - \mu^+| \right).$$

Together, the above expressions for γ and λ^+ give

$$\gamma = \lambda - \max_{i:|x_i - \mu| < \lambda} \left(|x_i - \mu^+| \right) = \min_{i:|x_i - \mu| < \lambda} \left(\lambda - |x_i - \mu^+| \right).$$

Since $|x_i - \mu^+| = \lambda^+$ for i satisfying the above equalities, we can invoke Lemma 9 to get

$$\gamma = \min_{i:|x_i - \mu| < \lambda} \left(\lambda - s_i(x_i - \mu^+) \right).$$

Now, making the substitution $\mu^+ = \mu + \gamma \partial \mu(\lambda)/\partial \lambda$ per equation (4.6) and rearranging terms leads to

$$0 = \min_{i:|x_i - \mu| < \lambda} \left(\lambda - s_i(x_i - \mu) - \gamma \left(1 - s_i \frac{\partial \mu(\lambda)}{\partial \lambda} \right) \right). \tag{C.3}$$

We have $1 - s_i \partial \mu(\lambda)/\partial \lambda > 0$ since $|\partial \mu(\lambda)/\partial \lambda| < 1$ (as established in the proof of Lemma 4). Hence, equality (C.3) remains valid after division by $1 - s_i \partial \mu(\lambda)/\partial \lambda$ inside the minimisation. Performing the division and isolating γ yields

$$\gamma = \min_{i:|x_i - \mu| < \lambda} \left(\frac{\lambda - s_i(x_i - \mu)}{1 - s_i \partial \mu(\lambda)/\partial \lambda} \right),$$

as per the result of the proposition. □