Received: 22 October 2020 | Revised: 7 June 2021 | Accepted: 10 August 2021

DOI: 10.1111/biom.13553

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Simultaneous feature selection and outlier detection with optimality guarantees

Luca Insolia[1,2] | **Ana Kenney**[3] | **Francesca Chiaromonte**[2,3] | **Giovanni Felici**[4]

[1] Faculty of Sciences, Scuola Normale Superiore, Pisa, Italy

[2] Institute of Economics & EMbeDS, Sant'Anna School of Advanced Studies, Pisa, Italy

[3] Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania, USA

[4] Istituto di Analisi dei Sistemi ed Informatica, Consiglio Nazionale delle Ricerche, Rome, Italy

Correspondence

Luca Insolia, Faculty of Sciences, Scuola Normale Superiore, Piazza dei Cavalieri, 7, Pisa, 56126, Italy.
Email: luca.insolia@sns.it

Funding information

National Institutes of Health, Grant/Award Number: 5T32LM012415-03; Huck Institutes of the Life Sciences of Penn State University

[Correction added on May 16, after first online publication: CRUI-CARE funding statement has been added.]

Abstract

Biomedical research is increasingly data rich, with studies comprising ever growing numbers of features. The larger a study, the higher the likelihood that a substantial portion of the features may be redundant and/or contain contamination (outlying values). This poses serious challenges, which are exacerbated in cases where the sample sizes are relatively small. Effective and efficient approaches to perform sparse estimation in the presence of outliers are critical for these studies, and have received considerable attention in the last decade. We contribute to this area considering high-dimensional regressions contaminated by multiple *mean-shift outliers* affecting both the response and the design matrix. We develop a general framework and use *mixed-integer programming* to simultaneously perform feature selection and outlier detection with provably optimal guarantees. We prove theoretical properties for our approach, that is, a necessary and sufficient condition for the *robustly strong oracle property*, where the number of features can increase exponentially with the sample size; the optimal estimation of parameters; and the breakdown point of the resulting estimates. Moreover, we provide computationally efficient procedures to tune integer constraints and warm-start the algorithm. We show the superior performance of our proposal compared to existing heuristic methods through simulations and use it to study the relationships between childhood obesity and the human microbiome.

KEYWORDS

breakdown point, mixed-integer programming, regression analysis, robust regression, sparse estimation, strong oracle property

1 | INTRODUCTION

High-dimensional regression problems have become ubiquitous in most application domains, and this is especially true in biomedical research where studies are consistently increasing in size and complexity. In these problems the number of features recorded on each observation (or case) is very large—possibly larger than the sample size, and often growing with the sample size itself. The availability of ever larger numbers of potential predictors increases both the chances that some substantial portion of them are irrelevant, and the chances of contamination in the data (i.e., of some cases following a different model). In principle, these risks may be mitigated in very controlled studies targeting specific populations, but these studies often have smaller sample sizes. In this article, we consider one such study investigating the relationship between childhood obesity and microbiome composition. We use data from Craig *et al.* (2018)—who studied weight gain in very young children as part of the *intervention nurses start infants growing on healthy trajectories* (INSIGHT) project (Paul *et al.*, 2014). Although previous work (Haffajee and Socransky, 2009; Zeigler *et al.*, 2012) focused on the relationship between adult and/or adolescent obesity and microbiome composition, Craig *et al.* (2018) connected infant weight gain (which is known to be predictive of obesity later in life, Taveras *et al.* 2009) to microbiota of the child, as well as the mother. As INSIGHT followed children with repeated visits and extensive data collection from birth to around 3 years of age, its sample size was fairly limited (in the hundreds). In such a setting, eliminating redundant features while accounting for potential contamination with estimation approaches that address both *sparsity* and *statistical robustness* is critical.

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

The notion that one can develop methods for *simultaneous feature selection and outlier detection* (SFSOD) stems from the fact that an MSOM can be equivalently parametrized with the inclusion of binary variables, transforming outlier detection into a feature selection problem (Morgenthaler *et al.*, 2004). This is exactly the avenue we pursue in this article. We propose a discrete and provably optimal approach to perform SFSOD based on the use of L_0 constraints—highlighting its connections with other methods and overcoming the heuristic nature of previous approaches. L_0 constraints have been used separately for feature selection (Bertsimas *et al.*, 2016) and robust estimation (Zioutas *et al.*, 2009)—both of which can be formulated as a *mixed-integer program* (MIP) and solved with optimality guarantees. We combine the two into a novel formulation and take advantage of existing heuristics to produce effective big-$\mathcal{M}$ bounds and warm-starts to reduce the computational burden of MIP.

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

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

The reminder of the article is organized as follows: Section 2 provides the relevant background. Section 3 details our proposal—including a general framework for SFSOD, the MIP formulation and its theoretical properties. Section 4 presents a simulation study comparing our proposal with state-of-the-art methods. Section 5 presents our application investigating the relationships between childhood obesity and microbiome composition. Final remarks are included in Section 6 and additional details are provided in the Supporting Information.

2 | BACKGROUND

Consider a regression model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ the error vector with a $N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ distribution ($\boldsymbol{I}_n$ is the identity matrix of size n), $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ the design matrix, and $\boldsymbol{\beta} \in \mathbb{R}^p$ the vector of regression coefficients. In the following, we briefly review methods for outlier detection, and present the equivalent formulation as a feature selection problem. We then discuss approaches for model selection, focusing on the use of an L_0 constraint for best subset selection.

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

If a regression comprises a single outlier, its position corresponds to the unit with largest absolute Studentized residual, which is a monotone transformation of the deletion residual

$t_i = (y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{(i)})/(\widehat{\sigma}_{(i)}(1+\boldsymbol{x}_i^T(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i)^{1/2}$, where the subscript (i) indicates the removal of the ith unit. Under the null model, a generic t_i follows a Student's t with $n-p-1$ degrees of freedom, which can be computed from an MLE fit based on all units and used as a test for outlying-ness of single data points (Cook and Weisberg, 1982). This can be easily generalized to regressions with multiple outliers. Operationally though, it was considered ineffective—due to the high likelihood of masking (undetected outlying cases) and swamping (nonoutlying cases flagged as outliers) effects—and computationally intractable (Bernholt, 2006). The presence of multiple MSOM outliers motivates the use of high-breakdown point estimators such as the least trimmed squares (LTS), S, and MM (Maronna *et al.* 2006, see also Section 3.3); outlier detection and high-breakdown point estimation are historically distinct but closely related areas of statistical research.

Assuming without loss of generality that outliers occupy the first n_0 positions in the data, the MSOM can be equivalently parametrized as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}_{n_0}\boldsymbol{\phi} + \boldsymbol{\varepsilon}$, where the original design matrix $\boldsymbol{X}$ is augmented with a binary matrix $\boldsymbol{D}_{n_0} = [\boldsymbol{I}_{n_0}, \boldsymbol{0}]^T$ of size $n \times n_0$ indexing the n_0 outliers (Morgenthaler *et al.*, 2004). If $p \leq n - n_0$, the MLE for $\boldsymbol{\phi} \in \mathbb{R}^{n_0}$ provides prediction residuals for the n_0 units excluded from the fit; that is, their residuals under a model that excludes them from the estimation process. This is given by $\widehat{\boldsymbol{\phi}} = [\boldsymbol{I}_{n_0} - \boldsymbol{H}_{MM}]^{-1}(\boldsymbol{y}_M - \boldsymbol{X}_M^T \widehat{\boldsymbol{\beta}}_{(M)})$, where $\boldsymbol{y}_M$ and $\boldsymbol{X}_M$ comprise values for the true set of outliers, $\boldsymbol{H}_{MM} = \boldsymbol{X}_M(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_M^T$, and the associated t-statistics $\boldsymbol{t}_M$ provide (multiple) deletion residuals. However, masking and swamping effects can again arise if $\boldsymbol{D}_{n_0}$ does not index all possible outliers.

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

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

Best subset selection, a traditional hard penalization method, solves feature selection combinatorially, comparing all possible models of size p_0 (Miller, 2002). It can be formulated as an MIP through an L_0 constraint on $\boldsymbol{\beta}$, where the L_0 pseudo-norm is defined as $\|\boldsymbol{\beta}\|_0 = \sum_j I(\beta_j \neq 0)$ $(I(\cdot)$ is the indicator function). The MIP formulation of best subset selection is computationally intractable (Natarajan, 1995) and was previously considered impossible to solve with optimality guarantees for regression problems of realistic size. Nevertheless, improvements in optimization solvers and hardware components, which experienced a 450 billion factor speed-up between 1991 and 2015, now allow one to efficiently solve problems of realistic size with provable optimality (Bertsimas *et al.*, 2016). Modern MIP solvers rely on implicit enumeration methods along with constraints such as *cutting planes* that tighten the relaxed problem (*branch & bound* and *branch & cut*, Schrijver 1986). Optimality is certified monitoring the gap between the best feasible solution and the problem relaxation. Notably, MIP methods can recover the subset of true active features (i.e., they satisfy oracle properties, see Section 3.3) under weaker assumptions compared to soft penalization methods. Here the MIP formulation is not equivalent to the L_0-penalty due to nonconvexity (Shen *et al.*, 2013).

3 | PROPOSED METHODOLOGY

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

$$[\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}] = \underset{\boldsymbol{\beta},\boldsymbol{\phi}}{\arg\min} \sum_{i=1}^{n} \rho(y_i, f(\boldsymbol{x}_i; \boldsymbol{\beta}) + \phi_i) \qquad (1)$$

$$\text{s.t. } R_\omega(\boldsymbol{\beta}) \leq c_\beta, \quad R_\gamma(\boldsymbol{\phi}) \leq c_\phi,$$

where $\rho(\cdot)$ is a loss function, $f(\cdot)$ defines the relation between predictors and response vector, and $R_\omega(\boldsymbol{\beta})$ and $R_\gamma(\boldsymbol{\phi})$ are penalties subject to sparsity-inducing constraints, which may depend on tuning constants ω and γ. Nonzero coefficients in $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\phi}}$ identify active features and outlying units, respectively. Although in this article we focus on linear regression the framework in (1) is very general; it comprises generalized linear models, several classification techniques and nonparametric methods.

Many approaches have been recently developed to solve (1) using ordinary least squares (OLS) as the loss function $\rho(\cdot)$. Both penalties $R_\omega(\boldsymbol{\beta})$ and $R_\gamma(\boldsymbol{\phi})$ are generally convex (Morgenthaler *et al.*, 2004; Menjoge and Welsch, 2010; Lee *et al.*, 2012; Kong *et al.*, 2018) although some nonconvex procedures have been considered (She and Owen, 2011). Robust soft penalization methods also can be cast into (1), abandoning the explicit use of $\boldsymbol{\phi}$ and adopting a robust loss $\rho(\cdot)$ in place of the OLS. These include MM-estimators for ridge regression (Maronna, 2011), sparse-LTS (Alfons *et al.*, 2013), bridge MM-estimators (Smucler and Yohai, 2017), enet-LTS (Kurnaz *et al.*, 2017), penalized elastic net S-estimators (Freue *et al.*, 2019), and penalized M-estimators (Loh, 2017; Chang *et al.*, 2018; Amato *et al.*, 2021), as well as their re-weighted counterparts. Indeed, through specific penalties, M-estimators can be equivalently formulated as feature selection problems (She and Owen, 2011).

Although (1) highlights an important parallel between SFSOD and robust soft penalization, existing heuristic methods suffer several drawbacks. Some rely on restrictive assumptions or their finite-sample and asymptotic performance in terms of feature selection and outlier detection is not well-established. Others rely heavily on an initial subset of nonoutlying cases. Yet others provide a down-weighting of all units, which complicates interpretation and the objective identification of outliers, or have an asymptotic breakdown point of 0%, so they in fact do not tolerate outliers in the first place. Finally, some methods require tuning of other parameters in addition to ω and γ, which can severely increase computational burden.

3.1 | MIP formulation

Our proposal solves (1) with optimality guarantees, from both optimization and theoretical perspectives. This preserves the intrinsic discreteness of the problem, facilitating implementation, interpretation, and generalizations. We impose two separate integer constraints on $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ in (1), combining in a single framework the use of L_0 constraints for feature selection (Bertsimas *et al.* 2016; Bertsimas and Van Parys 2020; Kenney *et al.* 2021) and outlier detection (Zioutas *et al.* 2009; Bertsimas and Mazumder 2014). In particular, we consider the MIP formulation in (2) where $\mathcal{M}^\beta$ and $\mathcal{M}^\phi$ in constraints (2a) and (2b) are the so-called big-$\mathcal{M}$ bounds (Schrijver, 1986). In our proposal these are vectors of lengths p and n, respectively, which can be tailored for each β_j and ϕ_i. In the L_0-norm constraints (2c) and (2d), k_p and k_n are positive integers modulating sparsity for feature selection and outlier detection, respectively—for the latter, we can think of sparsity as a level of trimming (i.e., outlier removal). In the L_2-norm ridge-like constraint (2e), $\lambda > 0$ can be used to counteract strong collinearities among the features (Hoerl and Kennard, 1970). It also modulates a trade-off between continuity and unbiasedness in the estimation of $\boldsymbol{\beta}$, and allows one to calibrate the intrinsic discreteness of the problem—making its solutions more stable with respect to data perturbations (Breiman, 1995) and weak signal-to-noise ratio regimes (Hastie *et al.*, 2017).

$$[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}] = \underset{\boldsymbol{\beta}, \boldsymbol{z}^\beta, \boldsymbol{\phi}, \boldsymbol{z}^\phi}{\arg\min}\ \frac{1}{n}\rho(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\phi}) \tag{2}$$

$$\text{s.t.} \quad -\mathcal{M}_j^\beta z_j^\beta \le \beta_j \le \mathcal{M}_j^\beta z_j^\beta \tag{2a}$$

$$-\mathcal{M}_i^\phi z_i^\phi \le \phi_i \le \mathcal{M}_i^\phi z_i^\phi \tag{2b}$$

$$\sum_{j=1}^{p} z_j^\beta \le k_p, \quad z_j^\beta \in \{0,1\}, \quad \beta_j \in \mathbb{R}, \quad j = 1, \dots, p \tag{2c}$$

$$\sum_{i=1}^{n} z_i^\phi \le k_n, \quad z_i^\phi \in \{0,1\}, \quad \phi_i \in \mathbb{R}, \quad i = 1, \dots, n \tag{2d}$$

$$\sum_{j=1}^{p} \beta_j^2 \le \lambda. \tag{2e}$$

Although solving (2) plainly with state-of-the-art software may be computationally intractable for large dimensions, with the appropriate implementation it can be used to tackle many real-world applications optimally and efficiently. In general, what it means for a statistical problem to be small or large depends on its structure—for instance, the "signal-to-noise" ratio plays a fundamental role (see Sections 3.3 and 4). From an operational standpoint, in this setting a problem can be considered large if the signal-to-noise ratio is small or moderate (e.g., smaller than 2), and the sample size and number of features are in the thousands or more. Another important advantage of our proposal from an application standpoint is that it allows one to easily incorporate additional constraints to leverage structure in the data—such as groups, ranked features, hierarchical interactions, and compositional information. We note that (2) could undergo a transformation through perspective cut model (Frangioni and Gentile, 2006) that may be of interest to improve the performance of the solution algorithm, here omitted for brevity.

3.2 | Some implementation details

Setting the big-$\mathcal{M}$ bounds for (2) is made even more complicated due to the "double" nature of SFSOD. A robust estimator of the regression coefficients, say $\tilde{\boldsymbol{\beta}}$, can be used to set $\mathcal{M}^\beta = \tilde{\boldsymbol{\beta}}c$ and $\mathcal{M}^\phi = (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})c = \tilde{\boldsymbol{e}}c$, where $c \ge 1$ is a suitable multiplicative constant. We generalize this approach using an *ensemble* $\tilde{\boldsymbol{\beta}}_t$ (for $t = 1, \dots, T$) of preliminary estimators and setting $\mathcal{M}_j^\beta = \max_t(|\tilde{\beta}_{t_j}|)c$ and $\mathcal{M}_i^\phi = \max_t(|\tilde{e}_{t_j}|)c$. The ensemble guarantees that, if at least one of the $\widetilde{\boldsymbol{\beta}}_t$'s is reasonably close to the optimal solution, the MIP will easily recover such solution. Importantly, having also non-robust or nonsparse estimators in the ensemble does not negatively affect solution quality but only convergence speed.

The MIP formulation in (2) critically depends on the big-$\mathcal{M}$ bounds; they should be large enough to retain the optimal solution, yet small enough to avoid unnecessary computations and numerical instability. If identifying suitable bounds is not possible, we use an alternative strategy based on *specially ordered sets of type 1* (SOS-1; Bertsimas *et al.* 2016). These allow only one variable in the set to be nonzero, for example, $(1-z_j^{\beta}, \beta_j) = 0 \iff (1-z_j^{\beta}, \beta_j)$: SOS-1, which can be solved via modern MIP solvers such as `Gurobi` or `CPLEX`. SOS-1 constraints in (2) guarantee that the global optimum can be reached, and generally outperform big-$\mathcal{M}$ bounds when these are difficult to reasonably set.

The formulation in (2) also, and critically, requires the tuning of k_p, k_n and, if a ridge-like constraint is included in the model, λ. Performing this simultaneously along an extensive grid of values can be computationally unviable for MIP. We therefore proceed as follows: **(i)** fix λ (possibly, in turn, to a few values in a meaningful range); **(ii)** fix k_n to a starting value larger than a reasonable expectation on the amount of contamination in the problem (n_0); **(iii)** holding fixed the k_n starting value from (ii), tune k_p through cross-validation or an information criterion; **(iv)** holding fixed the k_p value selected in (iii), refine downward the value of k_n. See also She and Owen (2011) for a discussion on parameter tuning for outlier detection. To the best of our knowledge, this is still an open research area, especially in high-dimensional settings. In our numerical studies, we found that there was little difference in choosing one tuning approach over the other. Thus, in this work we mainly focus on a robust counterpart of the BIC. We provide further details concerning feature standardization, cross-validation, selecting a trimming level, and using information criteria in Web Appendix B.

3.3 | Theoretical results

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

Without loss of generality, we assume that (2) has a unique global minimum, and that the loss function is such that $\rho(\boldsymbol{x}) \geq 0$ with $\rho(\mathbf{0}) = 0$ (this is the case for OLS and many other instances, such as estimation in quantile regression and robust estimators). Our first result connects our proposal to a large class of penalized methods based on trimming.

Theorem 1 (Sparse trimming). *For any λ, n, p, k_n and k_p, the $\widehat{\boldsymbol{\beta}}$ estimator produced solving (2) is the same as the one produced solving*

$$\arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n-k_n}\{\rho(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\}_{i:n} = \frac{1}{n}\sum_{i=1}^{n-k_n}\{\rho(e_i)\}_{i:n} \tag{3}$$

$$\text{s.t.} \quad (2a), (2c), (2e),$$

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

Theorem 1 demonstrates the equivalence of our formulation to a trimmed loss problem, where the level of trimming is directly controlled by the L_0 constraint on $\boldsymbol{\phi}$. This extends a well-known result for unpenalized OLS and motivates the formulation in (1) as a general framework for SFSOD. In particular, (2) includes some trimmed likelihood estimators as special cases (Hadi and Luceño, 1997). Thus, our proposal inherits their desirable properties, such as equivariance if the points are in general position (Maronna *et al.*, 2006).

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

Theorem 2 (MIP breakdown point). *For any λ, n, p, k_n and k_p, where $(y_i, \boldsymbol{x}_i^T)$ are not necessarily in general position, the BdP of the $\widehat{\boldsymbol{\beta}}$ estimator produced solving (2) is $\epsilon^* = (k_n + 1)/n$.*

Thus, $k_n \geq n_0$ is the only requirement to achieve the largest possible BdP. Similar results were obtained for the least quantile estimator (Bertsimas and Mazumder, 2014), the LTS estimator with a lasso penalty (Alfons *et al.*, 2013), and MM-estimators with a ridge or elastic net penalty

(Maronna 2011; Kurnaz *et al.* 2017). However, there are two caveats: the BdP can be misleading for nonequivariant estimators (Smucler and Yohai, 2017), and it only guarantees against the worst-case scenario—infinite maximum bias—as it does not account for large but finite biases in $\widehat{\boldsymbol{\beta}}$. This motivates the development of additional theoretical guarantees.

Next, we exclude the ridge-like penalty and take $\rho(\cdot) = \|\cdot\|_2^2$, making (2) a *mixed-integer quadratic program* (MIQP). In this setting, we prove that under certain conditions our approach satisfies the *robustly strong oracle property* (see Definition 1, based on Fan *et al.* 2014). In the following, we use the L_0 sparsity assumption on $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ as in Zhang and Zhang (2012). Recall that MSOM leads to outlier removal (see Section 2), and we showed in Theorem 1 that the L_0 constraint on $\boldsymbol{\phi}$ controls the level of trimming from the fit, thus this sparsity assumption on $\boldsymbol{\phi}$ is equivalent to the presence of MSOM outliers. In our SFSOD problem, as customary in feature selection literature, let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\phi}_0^T)^T \in \mathbb{R}^{p+n}$ be the true parameter vector, and decompose it as $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{\mathcal{S}}^T, \boldsymbol{\theta}_{\mathcal{S}^c}^T)^T = \{(\boldsymbol{\beta}_{\mathcal{S}_\beta}^T, \boldsymbol{\phi}_{\mathcal{S}_\phi}^T), (\boldsymbol{\beta}_{\mathcal{S}_\beta^c}^T, \boldsymbol{\phi}_{\mathcal{S}_\phi^c}^T)\}^T$ where $\boldsymbol{\theta}_{\mathcal{S}}$ contains only the true nonzero regression coefficients. Let the *robust oracle estimator* be $\widehat{\boldsymbol{\theta}}_0 = (\boldsymbol{A}_{\mathcal{S}}^T \boldsymbol{A}_{\mathcal{S}})^{-1} \boldsymbol{A}_{\mathcal{S}}^T \boldsymbol{y}$, where $\boldsymbol{A}_{\mathcal{S}} = (\boldsymbol{X}_{\mathcal{S}_\beta}, \boldsymbol{I}_{\mathcal{S}_\phi})$ is the $n \times (p_0 + n_0)$ matrix restricted to the active features belonging to $\mathcal{S}_\beta$ and the outlying cases belonging to $\mathcal{S}_\phi$. The robust oracle estimator is akin to the oracle estimator in feature selection—where the oracle is simply the OLS solution across the active set, when the features belonging to it are known. Our robust oracle estimator extends this concept taking also outliers into account. Specifically, $\widehat{\boldsymbol{\theta}}_0$ behaves as if the sets of active features and outliers were both known in advance. Indeed, if we know which points are outliers, we can include dummies for them, effectively removing their influence on the fit and making the OLS the optimal estimator.

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

Such robust version of the oracle property is stronger and more general than the oracle property in the sense of Fan and Li (2001), as it implies both SFSOD consistency and sign consistency (see also Bradic *et al.* 2011). Thus, SFSOD consistency depends on the achievability of the robust oracle estimator that we investigate by extending the theoretical framework in Shen *et al.* (2013) for feature selection. This requires weaker assumptions compared to other penalization methods (Zhang and Zhang, 2012), and we generalize it to the presence of MSOM outliers. Intuitively, if the robust oracle estimator is achievable (i.e., if it has the lowest objective for models of the same size), it is also the solution of our MIQP when the integer constraints are set to $k_p = p_0$ and $k_n = n_0$. Achievability depends on the difficulty of the problem, as measured by the *minimal degree of separation* between the true and a least favorable model—indexed by the supports $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, respectively. This is defined as $\Delta_\theta(\boldsymbol{A}) = \min_{\boldsymbol{\theta}_{\widetilde{\mathcal{S}}}} \|\boldsymbol{A}_{\mathcal{S}} \boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{A}_{\widetilde{\mathcal{S}}} \boldsymbol{\theta}_{\widetilde{\mathcal{S}}}\|_2^2 / \{n \max(|\mathcal{S} \backslash \widetilde{\mathcal{S}}|, 1)\}$ (for $\boldsymbol{\theta}_{\widetilde{\mathcal{S}}} : \widetilde{\mathcal{S}} \neq \mathcal{S}, |\widetilde{\mathcal{S}}_\beta| \leq p_0, |\widetilde{\mathcal{S}}_\phi| \leq n_0$), which relates to the signal-to-noise ratio and can be bounded as $\Delta_\theta \leq \Delta_\beta + \Delta_\phi$ (with Δ_β and Δ_ϕ defined similarly to Δ_θ using $\boldsymbol{X}$ and $\boldsymbol{I}_n$, respectively). We control this level of difficulty in Theorem 3, which provides a *necessary* condition for SFSOD consistency over $B(u, l) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_0 \leq u, \Delta_\theta \geq l\}$, the L_0-band with upper and lower radii u and l, respectively (a subset of the L_0-ball $B(u, 0)$ excluding a neighborhood of the origin).

Theorem 3 (Necessary condition for SFSOD consistency). *For any support estimate* $\widehat{\mathcal{S}}$ *and* $u > l > 0$, $\sup_{\boldsymbol{\theta}_0 \in B(u,l)} P(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$ *implies that*

$$\Delta_\theta \geq l = \frac{\sigma^2}{n} \max\left\{d_\beta \log(p), d_\phi \log(n)\right\}, \tag{4}$$

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

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

Theorem 4 (MIQP oracle reconstruction). *For any* n, p, n_0 *and* p_0*, the* $\widehat{\boldsymbol{\theta}}_{L_0}$ *estimator produced solving (2) with* $k_p = p_0$ *and* $k_n = n_0$ *is such that*

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

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

Theorem 5 (MIQP robustly strong oracle property). *Assume that $u_\theta = u_\phi + u_\beta$, where $u_\phi < n - k_p$ and $u_\beta < \min(n - k_n, p)$, and that there exists a constant $d_\theta > 36$ such that $l_\theta = d_\theta \sigma^2 / n \max\{\log(p), \log(n)\}$. Then, under (3) and for $(n, p) \to \infty$, the estimator $\widehat{\theta}_{L_0}$ produced solving (2) with $k_p = p_0$ and $k_n = n_0$ satisfies*

1. *Robustly strong oracle property: $\sup_{\theta_0 \in B(u_\theta, l_\theta)} P(\widehat{\mathcal{S}}^{L_0} = \mathcal{S}) \geq \sup_{\theta_0 \in B(u_\theta, l_\theta)} P(\widehat{\theta}_{L_0} = \widehat{\theta}_0) \to 1$ uniformly over $B(u_\theta, l_\theta) = \{\theta : \|\theta\|_0 = (p_0 + n_0) \leq u_\theta, \Delta_\theta \geq l_\theta\}$.*
2. *Asymptotic normality: $\sqrt{n}(\widehat{\theta}_{L_0} - \theta_0) \to^d N(\mathbf{0}, \mathbf{\Sigma}_\theta)$, where $\mathbf{\Sigma}_\theta = \sigma^2(\mathbf{A}_S^T \mathbf{A}_S / n)^{-1}$.*

Theorem 5(1) provides a sufficient condition for SFSOD consistency and the robust oracle reconstruction up to a constant term d_θ. Note that the number of features is allowed to exponentially increase with the sample size—so these properties hold also in ultra-high dimensional settings where $p = O(e^{n\alpha})$ with $\alpha = \Delta_\theta / (d_\theta \sigma^2) > 0$. Theorem 5(2) guarantees asymptotic normality and efficiency of MIQP estimates, which achieve the Cramèr–Rao lower bound as if the true sets of features and outliers were known a priori. Thus, although finite-sample inference with our approach can be problematic, as with other robust and/or regularization approaches, "large sample" statistical inference can be performed. Importantly, existing penalized M-estimators provide weaker results under stronger assumptions (Loh, 2017; Smucler and Yohai, 2017; Amato *et al.*, 2021). We conclude with a result showing that our proposal attains optimal parameter estimation with respect to the L_2-norm in the presence of MSOM outliers.

Theorem 6 (MIQP optimal parameter estimation). *Under the same conditions of Theorem 5, the estimator $\widehat{\theta}_{L_0}$ produced solving (2) with $k_p = p_0$ and $k_n = n_0$ provides*

1. *Optimal L_2-norm prediction error: $n^{-1} E\|\mathbf{A}(\widehat{\theta}_{L_0} - \theta_0)\|_2^2 = \sigma^2(p_0 + n_0)/n$.*
2. *Risk-minimax optimality for parameter estimation: $\sup_{\theta_0 \in B(u_\theta, l_\theta)} \; n^{-1} E\|\mathbf{A}(\widehat{\theta}_{L_0} - \theta_0)\|_2^2 = \inf_{\tau_n} \sup_{\theta_0 \in B(u_\theta, l_\theta)} n^{-1} E\|\mathbf{A}(\tau_n - \theta_0)\|_2^2 = \sigma^2 u_\theta / n$.*

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

4 | SIMULATION STUDY

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

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

Replicating each scenario a certain number of (independent) times, say q, and creating (independent) test data, say $(\mathbf{y}^*, \mathbf{X}^*)$, from the same generating scheme but without contamination, we compare methods with a variety of criteria, namely: (i) out-of-sample prediction performance, measured by the *root mean squared prediction error* $\text{RMSPE} = \{n^{-1} \sum_{i=1}^n (y_i^* - \mathbf{x}_i^* \widehat{\boldsymbol{\beta}})^2\}^{1/2}$; (ii) estimation accuracy for $\boldsymbol{\beta}$, measured by the *average mean squared error* $\text{MSE}(\widehat{\boldsymbol{\beta}}) = p^{-1} \sum_{j=1}^p \text{MSE}(\widehat{\beta}_j)$, where for each $\widehat{\beta}_j$ we form $\text{MSE}(\widehat{\beta}_j) = q^{-1} \sum_{i=1}^q (\widehat{\beta}_{ji} - \beta_j)^2 = (\overline{\beta}_j - \beta_j)^2 + q^{-1} \sum_{i=1}^q (\widehat{\beta}_{ji} - \overline{\beta}_j)^2$, decomposed in squared bias and variance (here $\overline{\beta}_j = q^{-1} \sum_{i=1}^q \widehat{\beta}_{ji}$); (iii) feature selection accuracy, measured by the *false positive rate* $\text{FPR}(\widehat{\boldsymbol{\beta}}) = |\{j \in \{1, \ldots, p\} : \widehat{\beta}_j \neq 0 \wedge \beta_j = 0\}| / |\{j \in \{1, \ldots, p\} : \beta_j = 0\}|$ and the *false negative rate* $\text{FNR}(\widehat{\boldsymbol{\beta}}) = |\{j \in \{1, \ldots, p\} : \widehat{\beta}_j = 0 \wedge$

TABLE 1 Mean (SD in parenthesis) of RMSPE, variance and squared bias for $\hat{\boldsymbol{\beta}}$, FPR and FNR for feature selection and outlier detection (as well as the corresponding F_1 scores), and computing time, based on 1000 simulation replications

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

$\beta_j \neq 0\}|/|\{j \in \{1, \dots, p\} : \beta_j \neq 0\}|$, as well as the F_1 score—which is a mixture of the two defined as $F_1(\hat{\boldsymbol{\beta}}) = (1 - \text{FNR})/\{(1 - \text{FNR}) + (\text{FPR} + \text{FNR})/2\}$; (iv) outlier detection accuracy, which is similarly measured by FPR($\hat{\boldsymbol{\phi}}$), FNR($\hat{\boldsymbol{\phi}}$) and $F_1(\hat{\boldsymbol{\phi}})$; (v) computational burden, measured as CPU time in seconds (this is used as a rough evaluation, since software implementations of different methods are not entirely comparable).

Using the robust oracle estimator as a benchmark, we compare the following estimators: (a) sparse-LTS (Alfons *et al.*, 2013), (b) enet-LTS (Kurnaz *et al.*, 2017), and (c) our MIP proposal (see Section 3). All methods trim the true number of outliers ($k_n = n_0$) and only their feature sparsity level is tuned. See Web Appendix B for implementation details.

Table 1 provides means and standard deviations (SD) of simulation results over $q = 1000$ replications. Our proposal substantially outperforms competing methods in most criteria. In particular, for the low-dimensional setting ($p = 50$), its RMSPE converges faster to the oracle solution and the variance of its $\hat{\boldsymbol{\beta}}$ decreases faster as n increases (the bias is essentially nonexistent for all methods). Notably, the FPR($\hat{\boldsymbol{\beta}}$) of sparse-LTS and enet-LTS increases with the sample size, while our approach avoids these type II errors. Even with these sparser solutions, we retain comparable (and at times lower) FNR($\hat{\boldsymbol{\beta}}$). Our method struggles most when $n = 50$, suggesting that additional work for tuning MIP may be beneficial in under-sampled problems. All methods perform very well in terms of FPR($\hat{\boldsymbol{\phi}}$) and FNR($\hat{\boldsymbol{\phi}}$), though enet-LTS is slightly worse for $n = 50$. As expected, the computational burden of our procedure is substantially higher than that of the competing heuristic methods—though we note that averages here are not representative, as there is a marked right skew in the distribution of computing times across replications. For comparison we provide medians and median absolute deviations (MAD) in Web Table 1 and find that results are even stronger. For example, the average computing time with $n = 150$ and $p = 200$ is 832.13 min compared to a median of 518.92 min. Our experience suggests that the growth in computational burden is mainly due to increases in the absolute number of outliers as the sample size increases.

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

5 | CONNECTING CHILDHOOD OBESITY AND MICROBIOME COMPOSITION

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

We focused on the *oral* microbiota of children and their mothers, which were found to contain interesting signals in Craig *et al.* (2018). Based on the preprocessing in Craig *et al.* (2018), we retained 215 child and 215 maternal oral samples. Correspondingly, we considered the abundances of 67 and 62 bacterial groups, respectively—which the original authors obtained aggregating phylogenetically sparse and correlated abundance data (we further filtered based on those with a MAD of 0 and/or exhibiting 0 counts in half or more of the samples). We also log-transformed the abundances of each group to mitigate skews. We focused on one among the phenotypes studied in Craig *et al.* (2018); namely, the *conditional weight gain score* (CWG)—a continuous measure computed from weight gain between birth and six months (a positive CWG indicates an accelerated weight gain) that is commonly used in paediatric research (Savage *et al.*, 2016).

We thus applied our approach along with sparse-LTS, enet-LTS, and classical lasso to two main models; the regressions of children's CWG on log-transformed bacterial groups abundances in oral samples of the children themselves, and of their mothers, respectively. The problem sizes were 215 × 68 and 215 × 63 with the inclusion of an intercept term. In addition to applying our approach on the full data sets, we split the data at random into training ($n^{\text{tr}} \approx 0.8n$) and test ($n^{\text{te}} \approx 0.2n$) sets to assess out-of-sample prediction performance of the various procedures. We also considered a different splitting ratio ($n^{\text{tr}} \approx 0.9n$ and $n^{\text{te}} \approx 0.1n$) and obtained similar results (see Web Table 4). Since the true contamination level of a given test set is unknown, we calculated a trimmed median squared prediction error (TMSPE) at 50% to be conservative. The trimming level when fitting each robust procedure was set to that found on the full data set (20% for both children and maternal regressions, which corresponds to 43 cases). We repeated the analysis on eight different training/test splits for all methods. Table 2 provides, for each of the two regressions, medians and MADs of results over the eight splits—including TMSPE and the number of features selected on the training set ($\hat{p}_0^{\text{tr}}$). The last column contains the total number of features selected on the full data ($\hat{p}_0^{\text{full}}$).

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

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

Switching to outlier detection, our procedure detected 43 outliers for both the children and the maternal regression. Note that 17 infants with particularly extreme CWG scores in either direction (see Web Figure 1) were detected as outliers in both regressions, with extreme (positive or negative) standardized residuals (see Web Figure 2). The child with the highest CWG and the largest residual in the children regression was one of the few infants (15/215) whose

TABLE 2 Median (MAD in parenthesis) of TMSPE and the number of features selected on the training set on eight train-test splits. Last column: number of features selected on the full data. Robust methods use 20% trimming

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

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

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

6 | FINAL REMARKS

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

Our approach relies on L_0 constraints—extending prior work where they were used separately for feature selection or outlier detection. Our simultaneous MIP formulation can handle problems of considerable size, and produces solutions that improve on existing heuristic methods. Although our formulation provides provably optimal solutions from the optimization perspective, it is crucial to tune its integer constraints. Thus, we also provide computationally efficient, data-driven approaches to induce sparsity in the coefficients and the estimated amount of contamination. Theoretical properties characterizing our proposal include its high breakdown point, the *robustly strong oracle property*—which holds in ultra-high dimensional settings where the number of predictors grows exponentially with the sample size—and optimality in parameter estimation with respect to the L_2-norm (i.e., optimal prediction error and risk-minimaxity). Our proposal requires weaker assumptions than prior methods in the literature and, unlike such methods, it allows the sparsity level and/or the amount of contamination to grow with the number of predictors and/or the sample size.

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

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

ACKNOWLEDGMENTS

This work was partially funded by the NIH B2D2K training grant and the Huck Institutes of the Life Sciences of Penn State. Computations were performed on the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer.

OpenAccess Funding provided by Scuola Superiore Sant'Anna within the CRUI-CARE Agreement.

OPEN RESEARCH BADGES

This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at https://github.com/LucaIns/SFSOD_MIP.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are openly available in Sequence Read Archive (SRA) at www.ncbi.nlm.nih.gov/bioproject/PRJNA420339 BioProject number `PRJNA42033` (raw microbiota reads), and in the database of Genotypes and Phenotypes (dbGaP) at www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001498.v1.p1, study number `phs001498.v1.p1` (phenotype information).

ORCID

Luca Insolia https://orcid.org/0000-0003-4169-5446
Ana Kenney https://orcid.org/0000-0002-2209-2431
Giovanni Felici https://orcid.org/0000-0003-0544-5407

REFERENCES

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

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

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

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

Beckman, R.J.& Cook, R.D. (1983) Outlier … … …. s. *Technometrics*, 25, 119–149.

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

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

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

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

Bradic, J., Fan, J.& Wang, W. (2011) Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B*, 73, 325–349.

Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics*, 37, 373–384.

Cabrera-Rubio, R., Collado, M.C., Laitinen, K., Salminen, S., Isolauri, E.& Mira, A. (2012) The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery. *The American Journal of Clinical Nutrition*, 96, 544–551.

Chang, L., Roberts, S.& Welsh, A. (2018) Robust lasso regression using Tukey's biweight criterion. *Technometrics*, 60, 36–47.

Cook, R.D. & Weisberg, S. (1982) *Residuals and influence in regression*. New York, NY: Chapman and Hall.

Craig, S.J., Blankenberg, D., Parodi, A. C.L., Paul, I.M., Birch, L.L., Savage, J.S., et al. (2018) Child weight gain trajectories linked to oral microbiota composition. *Scientific Reports*, 8, 1–14.

dbGaP (2017) INSIGHT cohort microbiome study. Available at: www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001498.v1.p1. dbGaP accession number phs001498.v1.p1. [Accessed July 5 2020].

Donoho, D.L. & Huber, P.J. (1983) The notion of breakdown point. In: Bickel, P., Doksum, K.A. & Hodges, J.L. (Eds.) *A festschrift for Erich L. Lehmann*. Belmont, CA: Wadsworth, pp. 157–184.

Fan, J.& Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J., Xue, L.& Zou, H. (2014) Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42, 819–849.

Frangioni, A.& Gentile, C. (2006) Perspective cuts for a class of convex 0–1 mixed integer programs. *Mathematical Programming*, 106, 225–236.

Freue, G. V.C., Kepplinger, D., Salibián-Barrera, M.& Smucler, E. (2019) Robust elastic net estimators for variable selection and identification of proteomic biomarkers. *The Annals of Applied Statistics*, 13, 2065–2090.

Gatu, C., Yanev, P.I.& Kontoghiorghes, E.J. (2007) A graph approach to generate all possible regression submodels. *Computational Statistics & Data Analysis*, 52, 799–815.

Gómez, A. (2021) Outlier detection in time series via mixed-integer conic quadratic optimization. *SIAM Journal on Optimization*, 31, 1897–1925.

Hadi, A.S.& Luceño, A. (1997) Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics & Data Analysis*, 25, 251–272.

Haffajee, A.D.& Socransky, S.S. (2009) Relation of body mass index, periodontitis and tannerella forsythia. *Journal of Clinical Periodontology*, 36, 89–99.

Hastie, T., Tibshirani, R. & Tibshirani, R. (2020) Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35, 579–592.

Hoerl, A.E.& Kennard, R.W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.

Insolia, L., Chiaromonte, F. & Riani, M. (2021) A robust estimation approach for mean-shift and variance-inflation outliers. In Bura, E. & Li, B. (Eds.) *Festschrift in Honor of R. Dennis Cook*. Berlin: Springer, pp. 17–41.

Kenney, A., Chiaromonte, F. & Felici, G. (2021) MIP-BOOST: efficient and effective L_0 feature selection for linear regression. *Journal of Computational and Graphical Statistics*, https://doi.org/10.1080/10618600.2020.1845184.

Kong, D., Bondell, H.D.& Wu, Y. (2018) Fully efficient robust estimation, outlier detection and variable selection via penalized regression. *Statistica Sinica*, 28, 1031–1052.

Kurnaz, F.S., Hoffmann, I.& Filzmoser, P. (2017) Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems*, 172, 211–222.

Lee, Y., MacEachern, S.N.& Jung, Y. (2012) Regularization of case-specific parameters for robustness and efficiency. *Statistical Science*, 27, 350–372.

Liu, J., Cosman, P.C.& Rao, B.D. (2017) Robust linear regression via ℓ_0 regularization. *IEEE Transactions on Signal Processing*, 66, 698–713.

Liu, T.& Jiang, H. (2019) Minimizing sum of truncated convex functions and its applications. *Journal of Computational and Graphical Statistics*, 28, 1–10.

Liu, H., Wang, J., He, T., Becker, S., Zhang, G., Li, D. & Ma, X. (2018) Butyrate: a double-edged sword for health? *Advances in Nutrition*, 9, 21–29.

Loh, P. (2017) Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45, 866–896.

Maronna, R.A. (2011) Robust ridge regression for high-dimensional data. *Technometrics*, 53, 44–53.

Maronna, R.A., Martin, R.D. & Yohai, V.J. (2006) *Robust statistics: theory and methods*. New York, NY: John Wiley & Sons.

McCann, L. (2006) *Robust model selection and outlier detection in linear regressions*. PhD thesis, Massachusetts Institute of Technology.

Menjoge, R.S.& Welsch, R.E. (2010) A diagnostic method for simultaneous feature selection and outlier identification in linear regression. *Computational Statistics & Data Analysis*, 54, 3181–3193.

Miller, A.J. (2002) *Subset selection in regression*, 2nd edition. Boca Raton, FL: Chapman and Hall/CRC.

Morgenthaler, S., Welsch, R.E. & Zenide, A. (2004) Algorithms for robust model selection in linear regression. In: Hubert, M., Pison, G., Struyf, A. & Van Aelst, S. (Eds.) *Theory and applications of recent robust methods*. Basel: Springer, pp. 195–206.

Müller, S.& Welsh, A.H. (2005) Outlier robust model selection in linear regression. *Journal of the American Statistical Association*, 100, 1297–1310.

Natarajan, B.K. (1995) Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24, 227–234.

Paul, I.M., Williams, J.S., Anzman-Frasca, S., Beiler, J.S., Makova, K.D., Marini, M.E., et al. (2014) The intervention nurses start infants growing on healthy trajectories (insight) study. *BMC Pediatrics*, 14, 1–15.

Savage, J.S., Birch, L.L., Marini, M., Anzman-Frasca, S.& Paul, I.M. (2016) Effect of the insight responsive parenting intervention on rapid infant weight gain and overweight status at age 1 year: a randomized clinical trial. *BMC Pediatrics*, 170, 742–749.

Schrijver, A. (1986) *Theory of linear and integer programming*. New York, NY: John Wiley & Sons.

She, Y.& Owen, A.B. (2011) Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106, 626–639.

Shen, X., Pan, W.& Zhu, Y. (2012) Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107, 223–232.

Shen, X., Pan, W., Zhu, Y.& Zhou, H. (2013) On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65, 807–832.

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

SRA (2017) INSIGHT oral and gut microbiome. Available at: www.ncbi.nlm.nih.gov/bioproject/PRJNA420339. NCBI BioProject number PRJNA420339. [Accessed 5 July 2020].

Taveras, E.M., Rifas-Shiman, S.L., Belfort, M.B., Kleinman, K.P., Oken, E.& Gillman, M.W. (2009) Weight status in the first 6 months of life and obesity at 3 years of age. *Pediatrics*, 123, 1177–1183.

Taylan, P., Yerlikaya-Özkurt, F., Bilgiç Uçak, B. & Weber, G. (2020) A new outlier detection method based on convex optimization: application to diagnosis of Parkinson's disease. *Journal of Applied Statistics*, https://doi.org/10.1080/02664763.2020.1864815.

Taylan, P., Yerlikaya-Özkurt, F.& Weber, G. (2014) An approach to the mean shift outlier model by Tikhonov regularization and conic programming. *Intelligent Data Analysis*, 18, 79–94.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58, 267–288.

Vital, M., Howe, A.C.& Tiedje, J.M. (2014) Revealing the bacterial butyrate synthesis pathways by analyzing (meta) genomic data. *mBio*, 5, 1–11.

Yerlikaya-Özkurt, F.& Taylan, P. (2020) New computational methods for classification problems in the existence of outliers based on conic quadratic optimization. *Communications in Statistics-Simulation and Computation*, 49, 753–770.

Zeigler, C.C., Persson, G.R., Wondimu, B., Marcus, C., Sobko, T.& Modéer, T. (2012) Microbiota in the oral subgingival biofilm is associated with obesity in adolescence. *Obesity*, 20, 157–164.

Zhang, C.-H.& Zhang, T. (2012) A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27, 576–593.

Zioutas, G., Pitsoulis, L.& Avramidis, A. (2009) Quadratic mixed integer programming and support vectors for deleting outliers in robust regression. *Annals of Operations Research*, 166, 339–353.

SUPPORTING INFORMATION

Web Appendices, Tables and Figures referenced in Sections 3–5, as well as the code to replicate both our simulation study and the microbiome application, are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Insolia, L., Kenney, A., Chiaromonte, F., Felici, G. Simultaneous feature selection and outlier detection with optimality guarantees. *Biometrics*. 2022;781592–1603.
https://doi.org/10.1111/biom.13553