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

Luca Insolia*
Faculty of Sciences, Scuola Normale Superiore
Institute of Economics & EMbeDS, Sant’Anna School of Advanced Studies

Francesca Chiaromonte
Department of Statistics, Penn State University
Institute of Economics & EMbeDS, Sant’Anna School of Advanced Studies

Runze Li
Department of Statistics, Penn State University

Marco Riani
Department of Economics and Management, University of Parma

Abstract
We propose a general approach to handle data contaminations that might disrupt the performance of feature selection and estimation procedures for high-dimensional linear models. Specifically, we consider the co-occurrence of mean-shift and variance-inflation outliers, which can be modeled as additional fixed and random components, respectively, and evaluated independently. Our proposal performs feature selection while detecting and down-weighting variance-inflation outliers, detecting and excluding mean-shift outliers, and retaining non-outlying cases with full weights. Feature selection and mean-shift outlier detection are performed through a robust class of nonconcave penalization methods. Variance-inflation outlier detection is based on the penalization of the restricted posterior mode. The resulting approach satisfies a robust oracle property for feature selection in the presence of data contamination – which allows the number of features to exponentially increase with the sample size – and detects truly outlying cases of each type with asymptotic probability one. This provides an optimal trade-off between a high breakdown point and efficiency. Computationally efficient heuristic procedures are also presented. We illustrate the finite-sample performance of our proposal through an extensive simulation study and a real-world application.

Keywords: Mean-shift outliers; Nonconvex penalties; Robust estimation; Variable selection; Variance-inflation outliers.

*This work was partially funded by the Huck Institutes of the Life Sciences of Penn State.
1 Introduction

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

Two main contamination mechanisms have been investigated in the literature on linear models (Beckman and Cook 1983), namely: the mean-shift outlier model (MSOM) and the variance-inflation outlier model (VIOM). The MSOM assumes that outlying cases have a shift in mean; maximum likelihood estimation (MLE) leads to their removal from the fit – i.e., to the assignment of 0 weights to the cases identified as outliers. While the MSOM was traditionally studied in low-dimensional scenarios (Cook and Weisberg 1982), it has been recently extended to high-dimensional linear models, where the use of regularization techniques is fundamental (She and Owen 2011; Alfons et al. 2013; Kurnaz et al. 2017; Insolia et al. 2020). The VIOM, which is historically considered as an alternative to the MSOM, assumes that contaminated errors have an inflated variance; outliers are retained but down-weighted in the fit. The VIOM was initially investigated by Cook et al. (1982) and Thompson (1985) in the presence of a single outlier, using MLE and restricted MLE (REMLE), respectively. More recently, Gumedze (2019) developed hypothesis testing procedures for linear models, considering also the presence of multiple outliers. However, when multiple outliers are present, this approach requires the evaluation of a combinatorial number of outlying-ness tests to avoid masking (undetected outlying cases) and swamping (non-outlying cases flagged as outliers). Insolia et al. (2021) proposed the use of robust estimation and REMLE to detect and down-weight multiple VIOM outliers, possibly co-occurring with MSOM outliers, in (low-dimensional) linear models.

High-dimensional settings with VIOM outliers, to the best of our knowledge, have not been explored yet. Here we aim to fill this gap and, like in Insolia et al. (2021), we further consider the co-occurrence of multiple MSOM and VIOM outliers. These are modeled as additional fixed and random components, respectively, which can be estimated independently based on REMLE principles. Specifically, we propose a doubly robust class of nonconcave penalization methods, in which feature selection and MSOM detection rely on a trimmed penalized loss, whereas VIOM detection is based on the penalization of the restricted posterior mode. The resulting procedure: (i) satisfies a robust oracle property for feature selection in the presence of data
contamination, which allows the number of features to exponentially increase with the sample size; (ii) detects MSOM and VIOM outliers with asymptotic probability one; (iii) achieves an optimal trade-off between high breakdown point and efficiency, and thus provides optimal units’ weights. Effective and computationally efficient heuristic procedures are also presented.

Importantly, our approach comprises “hard” trimming sparse estimators as a special case. However, since we rely on nonconcave penalization methods, our proposal satisfies oracle properties under weaker assumptions compared to existing robust estimators based on convex penalties (Kurnaz et al. 2017; Alfons et al. 2013). This provides an important bridge between the latter and $L_0$-constrained formulations with optimality guarantees (Insolia et al. 2020). Moreover, unlike “soft” trimming estimators which produce a general down-weighting for all points (Loh 2017; Smucler and Yohai 2017; Chang et al. 2018; Freue et al. 2019; Amato et al. 2021), our proposal is effective in estimating full weights for non-outlying observations.

The reminder of the paper is organized as follows. Section 2 reviews relevant background literature. Section 3 details our proposal, which is a 3-step procedure, as well as its heuristic counterpart. Section 4 contains numerical studies comparing the empirical properties of different methods both in low- and high-dimensional settings, and Section 5 contains a real-world application. Final remarks are given in Section 6. Further details, extensions and proofs, as well as the source code to replicate our simulation and application studies, are provided in the Supplementary Material.

# 2 Background

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

## 2.1 Outlier Detection

Consider a classical linear regression model of the form $y = X\beta + \varepsilon$, where $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ contains observable responses, $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times p}$ is the design matrix, $\beta \in \mathbb{R}^p$ contains unknown fixed effects (possibly sparse), and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ contains unobservable random errors. Classical assumptions specify that such errors are uncorrelated, homoscedastic and Gaussian, so that $\varepsilon \sim N(0, \sigma^2 I_n)$ for $0 < \sigma^2 < \infty$.

The MSOM postulates that for outlying cases $i \in S_\phi$ (the rationale for this symbol will become clear in Equation 2), $\varepsilon_i \sim N(\mu_{\varepsilon_i}, \sigma^2)$ with $\mu_{\varepsilon_i} \neq 0$. Under the
assumption that $S_φ$ is known and $\text{rank}(X) = p \leq n - |S_φ|$ (where $|·|$ denotes the cardinality of a set), the MLE leads to the exclusion of the units in $S_φ$ from the fit (Cook and Weisberg 1982). If there is a single MSOM outlier, this represents the unit with largest absolute Studentized residual, which is a monotone transformation of the deletion residual $t_i = (y_i - x_i^T\hat{β}(i))/(\hat{σ}(i)(1 + x_i^T X_{(i)}^{-1} x_i)^{1/2})$, where the parenthetical subscript indicates the exclusion of unit $i$ from the fit. Importantly, $t_i$ can be computed very cheaply and, for a generic $i$, follows a Student’s $t$ with $n - p - 1$ degrees of freedom under the null – thus, it can be used to test the outlying-ness of each observation. Although this can be easily generalized to the presence of multiple MSOM outliers, it requires the evaluation of a combinatorial number of fits (i.e., excluding all possible subsets of points of a given size from the fit), which results in a computationally intractable problem. Relatedly, high-breakdown estimators (see Section 3.1) aim at limiting the influence of extreme residuals on the fit (Maronna et al. 2006). Although these are traditionally computed using heuristic approaches, the use of MIP techniques has been recently considered to effectively solve the underlying combinatorial problem with optimality guarantees (Zioutas and Avramidis 2005; Bertsimas and Mazumder 2014). Importantly, high-breakdown point estimators have also been extended to sparse high-dimensional linear models in combination with penalization methods (Alfons et al. 2013; Smucler and Yohai 2017; Kurnaz et al. 2017; Freue et al. 2019). Here $L_0$-constraints, which can be solved through MIP algorithms, provide optimality guarantees and desirable statistical properties for simultaneous feature selection and MSOM detection, with $p$ allowed to increase exponentially with $n$ (Insolia et al. 2020).

The VIOM postulates that for outlying cases $i \in S_γ$ (also this symbol will become clear in Equation 2), $ε_i \sim N(0, σ^2 v_i)$ with $v_i = (1 + ω_i) ≥ 1$. Cook et al. (1982) studied the presence of a single variance-inflated outlier; the MLE estimate of $β$ depends on its $v_i$ and results in a weighted least squares (WLS) fit $\hat{β}(v_i) = (X^T W X)^{-1} X^T W y = \hat{β} - (X^T X)^{-1} X^T ε_i [(1 - w_i)/(1 - (1 - w_i) H_{x,ii})]$, where $W$ is a diagonal matrix containing all ones but $w_i = v_i^{-1}$. The tilde indicates quantities computed from the ordinary least squares (OLS) fit, and $H_{x,ii}$ is the $i$-th diagonal element of $H_x = X(X^T X)^{-1} X^T$. This highlights the fact that the VIOM is asymptotically equivalent to the MSOM as $v_i \rightarrow \infty$. Importantly, in the presence of a single VIOM outlier, the MLE provides a closed-form estimate for $v_i$, which can be used to estimate $β$ and $σ^2$. Similarly, Thompson (1985) used REMLE in place of MLE to estimate the variance components $v_i$ and $σ^2$. REMLE relies on $n - p$ linearly independent error contrasts $A^T ε$, where $A \in \mathbb{R}^{n \times (n-p)}$ is defined such that $A^T A = I_n$ and $AA^T = P_x$, with $P_x = I_n - H_x$ (Patterson and Thompson.
Also REMLE provides a closed-form estimate for the single variance-inflation parameter \( v_i \). Notably, the single VIOM outlier position estimated by MLE and REMLE might differ. A sufficient condition for their agreement is that the unit with maximum absolute OLS residual max(\(|\tilde{e}_i|\)) also has the largest absolute Studentized residual max(\(|t_i|\)) – the latter estimates the outlier position using REMLE, which is equivalent to the outlier position estimated by MLE under an MSOM (Thompson 1985). However, differently from the case of a single VIOM outlier (and of multiple MSOM outliers), multiple variance-inflation parameters \( v \) cannot be estimated in closed-form even if the outliers are known – thus, iterative procedures are required (Gumedze 2019). In order to detect multiple VIOM outliers, possibly concurrent with MSOM outliers, Insolia et al. (2021) proposed the use of robust estimation for outlier detection and of REMLE to estimate optimal units’ weights. Nevertheless, to the best of our knowledge, high-dimensional linear models affected by VIOM contamination have not been explored yet.

### 2.2 Feature Selection for Mixed-Effects Linear Models

Mixed-effects linear models are often used to model data with a natural group structure, such as repeated measurements, measurements in time, and measurements in space (Laird and Ware 1982). They extend the classical linear model through the inclusion of a random design matrix characterizing the experiment; namely, \( y = X \beta + Z b + \varepsilon \), where \( Z = [Z_1, \ldots, Z_t] \in \mathbb{R}^{n \times q} \), and \( Z_j \in \mathbb{R}^{n \times q_j} \) indicates the design matrix for the \( j \)-th random effect \( b_j \in \mathbb{R}^{q_j} \), such that \( b = (b_1^T, \ldots, b_t^T)^T \in \mathbb{R}^q \), and \( \sum_j q_j = q \). It is often assumed that \( b \sim N(0, B) \), where \( B = [B_1, \ldots, B_t] \) is a block-diagonal matrix modeling the covariance of each random effect \( b_j \sim N(0, B_j) \), with \( \text{cov}(b_k, b_l) = 0 \) for any \( k \neq l \). Moreover, \( b \) and \( \varepsilon \) are assumed to follow independent Gaussian distributions.

Several methods have been developed to simultaneously estimate fixed and random effects. Henderson’s mixed-model equations lead to the best linear unbiased estimator (BLUE) for the fixed effects \( \beta \) and the best linear unbiased predictor (BLUP) for the random effects \( b \) – which is also known as the empirical Bayes estimator as it maximizes the posterior distribution \( f(b | y) \). However, this approach is unviable to perform feature selection in high-dimensional scenarios (Fan and Li 2012). For this purpose, hypothesis testing procedures have been developed to select relevant random effects (Lin 1997). Different sub-models can be compared through extensions of information criteria, such as the conditional Akaike information criterion (CAIC) (Liang et al. 2008) and its generalizations. Leveraging penalization methods, other approaches perform sparse estimation of the fixed effects \( \beta \). In these, while the di-
mension $p$ of $\beta$ is allowed to increase with the sample size $n$, the random component $b$ is often assumed to contain only truly relevant random effects (Scheldorfer et al. 2011). Yet other approaches use penalization methods to select a given number of fixed and random effects (Bondell et al. 2010; Ibrahim et al. 2011; Peng and Lu 2012). See Müller et al. (2013) and Buscemi and Plaia (2020) for a literature review.

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

\section{Our Proposal}

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

$$
\varepsilon_i \sim \begin{cases} 
N(0, \sigma_{\varepsilon i}^2 v_i) & \forall i \in S_{\gamma} \\
N(\mu_{\varepsilon i}, \sigma^2) & \forall i \in S_{\phi} \\
N(0, \sigma^2) & \text{otherwise} 
\end{cases}
$$

(1)

where $v_i > 1$ and $\mu_{\varepsilon i} \neq 0$. We exclude overlaps between the two types of contamination because such over-parametrization is equivalent to a MSOM assumption (Cook et al. 1982). Moreover, as customary in the robust statistics literature, we let MSOM outliers also affect the design matrix $X$ (with shifts $\mu_{x i}$) creating leverage points (Maronna et al. 2006).

Notably, the outliers in (1) can be equivalently represented adding fixed and random effects to the linear model (Insolia et al. 2021). In symbols

$$
y = X\beta + D_{S_{\gamma}} \gamma + D_{S_{\phi}} \phi + \varepsilon,
$$

(2)

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

\[
y = X\beta + I_n\gamma + I_n\phi + \epsilon
\]

and rely on nonconcave penalization methods to select relevant fixed effects \( \beta \) – but we also enforce sparsity in \( \gamma \in \mathbb{R}^n \), which detects and down-weights VIOM outliers, and \( \phi \in \mathbb{R}^n \), which detects and excludes MSOM outliers from the fit. Specifically, we propose a 3-step procedure based on REMLE principles, that extends and combines the approaches in Fan and Li (2012) and Insolia et al. (2020, 2021). Operationally, the three steps can be solved iteratively (see Section 4), and we first focus on fixed effects estimation, as MSOM outliers can have stronger influence on model estimates.

3.1 Step 1: Feature Selection and MSOM Detection

Suppose that \( S_\gamma \) is known. Then, plugging the MLE estimates for \( \gamma | \beta \) in the joint density distribution \( f(y, \gamma) \) leads to the profile log-likelihood:

\[
l_n(\beta, \hat{\gamma}) \propto \frac{1}{2\sigma^2}(y - X\beta - \phi)^TP_R(y - X\beta - \phi),
\]

which produces a WLS estimator as

\[
P_R = (I_n - B_\gamma)(I_n - B_\gamma)^T + B_\gamma^TD_{S_\gamma}\Gamma^{-1}D_{S_\gamma}^TB_\gamma
= (I_n + D_{S_\gamma}\Gamma D_{S_\gamma}^T)^{-1}
= W,
\]

where \( B_\gamma = (I_n + D_{S_\gamma}\Gamma D_{S_\gamma}^T)^{-1} \). We simultaneously select and estimate fixed effects \( \beta \), while detecting and discarding MSOM outliers from the fit, using a feasible and robustly penalized version of (4), where an integer constraint and a nonconcave penalty are used for MSOM outlier detection and feature selection, respectively. In
\[
[\hat{\beta}, \hat{\phi}] = \arg \min_{\beta, \phi} \frac{1}{2} (y - X\beta - \phi)^T M_R(y - X\beta - \phi) + (n - k_n) \sum_{j=1}^{p} R_\lambda(|\beta_j|) \quad (6)
\]
\[
\text{s.t.} \quad \|\phi\|_0 = \sum_{i=1}^{n} I(\phi_i \neq 0) \leq k_n, \quad (6a)
\]

where \(I(\cdot)\) is the indicator function, and the matrix \(M_R\) is a proxy for the unknown \(P_R/\sigma^2\) (see the Supplementary Material for details). Note that if \(M_R\) is a multiple of the identity matrix, then (6) neglects VIOM outliers – i.e., all points receive binary weights.

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

Conditions List 1 (Penalty function). For any \(\lambda > 0\), the penalty \(R_\lambda(t), t \in [0, \infty)\) is: (i) non-decreasing and concave with \(R_\lambda(0) = 0\), (ii) twice continuously differentiable with first derivative \(R_\lambda'(0^+) > 0\), and (iii) such that \(\sup_{t>0} R_\lambda''(t) \to 0\) for \(\lambda \to 0\).

These conditions are fairly common for concave penalization methods (see for instance Fan and Lv 2011), and are used to develop estimators with three desirable properties: unbiasedness, sparsity and continuity (Fan and Li 2001). We specifically focus on the SCAD penalty \(R_\lambda(\cdot)\) in (6), but others might be considered as well. The SCAD penalty satisfies \(R_\lambda(0) = 0\) and, for \(t \in (0, \infty)\), has \(R_\lambda'(t) = \lambda I(t \leq \lambda) + [(a\lambda - t)/(a - 1)]I(t > \lambda)\), where the constant \(a > 2\) controls nonconcavity and is often set to \(a = 3.7\). This folded-concave penalty is continuously differentiable on \((-\infty, 0) \cup (0, \infty)\) and singular at 0. Since its derivative is zero outside \([-a\lambda, a\lambda]\), it does not shrink and thus bias large coefficient estimates. Obtaining a global minimum with folded-concave penalties such as SCAD is non-trivial. In the following we focus on the local linear approximation (LLA) method (Zou and Li 2008) to obtain a local solution which guarantees oracle properties. However, in principle one can achieve the global minimum using MIP techniques (Liu et al. 2016).
The \( L_0 \)-constraint in (6a) is used for MSOM outlier detection. It depends on an integer tuning parameter \( k_n \geq 0 \) controlling the trimming level – i.e., the number of points which are identified as MSOMs and excluded from the fit. This guarantees the achievability of high-breakdown estimates (see below). Modern MIP solvers can be used to solve the formulation in (6) with optimality guarantees (Bertsimas et al. 2016; Insolia et al. 2020; Kenney et al. 2021). However, in order to reduce the computational burden, one can also use well-established heuristic algorithms (Alfons et al. 2013; Kurnaz et al. 2017).

Intuitively, the breakdown point (BdP) measures the largest fraction of contamination that an estimator can tolerate before it becomes arbitrarily biased (Donoho and Huber 1983). The finite-sample replacement BdP is defined as \( \varepsilon^*(\hat{\beta}, Z) = \min(m/n : \sup_Z \|\hat{\beta}(Z)\|_2 = \infty) \), where \( Z \) denotes the original dataset \( Z = (X, y) \) after the replacement of \( m \) out of \( n \) points with arbitrary values. The following result shows that our proposal achieves the highest possible BdP.

**Proposition 1** (High breakdown-point). For any \( \lambda > 0 \) and \( a > 2 \) the estimator \( \hat{\beta} \) produced by (6) achieves a breakdown point of \( \varepsilon^* = (k_n + 1)/n \).

Thus, in the presence of MSOM contamination, our proposal breaks down only if \( k_n < m_M \). Moreover, this result does not require that the points \( (x_i^T, y_i) \) are in general position. This is necessary for low-dimensional estimators to achieve equivariance (Maronna et al. 2006) – something that cannot be achieved by our proposal (Maronna 2011).

Note that lasso estimation can be considered as the first iteration in computing the SCAD penalty based on the LLA method (Zou and Li 2008). Thus, while SCAD provides stronger theoretical results for feature selection, one can perform MSOM outlier detection with existing robust algorithms based on lasso, e.g., the \textit{sparseLTS} (Alfons et al. 2013) which solves a trimmed loss problem with an \( L_1 \)-penalty using heuristic algorithms. Then, SCAD can be computed on the set of non-outlying cases detected by a robust lasso on the first iteration of LLA; this is the approach followed in our implementation described below.

We remark that the notion of breakdown can be misleading for non-equivariant estimators, such as those produced through penalties (Maronna 2011; Smucler and Yohai 2017; Insolia et al. 2020). Hence, we provide additional guarantees in terms of simultaneous MSOM outlier detection and feature selection. Let \( \theta_0 = (\beta_0^T, \phi_0^T)^T \in \mathbb{R}^{p+n} \) be the true parameter vector, and decompose it as \( \theta_0 = (\theta_S^T, \theta_S^c)^T = (\beta_S^T, \phi_S^T), (\beta_S^c, \phi_S^c)^T \) where \( \theta_S \) contains the \( p_0 \) non-zero coefficients belonging to \( S_\beta \), and the \( m_M \) outlying cases belonging to \( S_\phi \) (\( \cdot^c \) indicates the complement of a set). \( \hat{\theta}_0 \) represents a fixed-effects robust oracle estimator, behaving as if the true sets of active
features and outliers were both known in advance. Let \( \| \cdot \|_\infty \) indicate the matrix infinity norm, and \( \Lambda_{\min}(\cdot) \) and \( \Lambda_{\max}(\cdot) \) the minimum and maximum eigenvalue of a matrix, respectively. We rely on the following conditions to recover \( \hat{\theta}_0 \).

**Conditions List 2** (Fixed-effects robust oracle reconstruction).

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

B. **Design and proxy matrices**: for some constants \( \eta \in (2\tau, 1) \) and \( c_0 > 0 \), the matrices \( (n - m_M)^{-1} (X_{S_\gamma}^T S_\beta X_{S_\gamma} S_{\beta}) \) and \( (n - m_M)^n (X_{S_\gamma}^T S_\beta X_{S_\gamma} S_{\beta})^{-1} \) have minimum and maximum eigenvalues bounded from below and above by \( c_0 \) and \( c_0^{-1} \), respectively. Moreover

\[
\left\| \left( \frac{1}{n - m_M} X_{S_\gamma}^T S_\beta M(R) X_{S_\gamma} S_{\beta} \right)^{-1} \right\|_\infty \leq \frac{\{ \log(n - m_M) \}^{3/4}}{(n - m_M)^\tau R'_\lambda(s_1/2)},
\]

\[
\left\| X_{S_\gamma}^T S_\beta M(R) X_{S_\gamma} S_{\beta} \left( \frac{1}{n - m_M} X_{S_\gamma}^T S_\beta M(R) X_{S_\gamma} S_{\beta} \right)^{-1} \right\|_\infty < \frac{R'_\lambda(0+)}{R'_\lambda(s_1/2)}.
\]

C. **Proxy matrix**: \( \Lambda_{\min}(c_1 M_{\gamma}^{S_\gamma} - \Gamma) \geq 0 \) and \( \Lambda_{\min}(c_1 \log(n - m_M) \Gamma - M_{\gamma}^{S_\gamma}) \geq 0 \) for some constant \( c_1 > 0 \), and \( M_{\gamma}^{S_\gamma} = I_{n - m_M} \). Here \( M_{\gamma}^{S_\gamma} \) and \( M_{\gamma}^{S_\gamma} \) index rows and columns of the proxy matrix \( M_{\gamma} \) corresponding to non-VIOMs and VIOMs, respectively.

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

\[
\Delta_\phi = \min_{\hat{\phi}_{S_\phi}} \frac{\| X_{S_\phi} \hat{\beta}_{S_\phi} + I_{n S_\phi} \hat{\phi}_{S_\phi} - X_{S_\phi} \beta_{S_\phi} - I_{n S_\phi} \phi_{S_\phi} \|^2}{n \max(|S_\phi \setminus S_\beta|, |S_\beta \setminus S_\phi|, 1)}
\]

where \( \hat{\phi}_{S_\phi} \) is any estimate such that \( \hat{S}_\phi \neq S_\phi, |\hat{S}_\phi| \leq m_M \) and \( \hat{\beta}_{S_\phi} \) satisfies \( |\hat{S}_\phi| \leq p_0 \).

Conditions 2(A)-(C) are quite common for nonconcave penalization methods such as SCAD (Fan and Li 2012), and they are based only on the set of non-outlying cases indexed by \( S_\beta \). Condition 2(D) is specifically required to detect MSOM outliers based on \( L_0 \)-constraints (Insolia et al. 2020). It bounds the difficulty of MSOMs detection based on a *minimal degree of separation* between the true and a least favorable model. Intuitively, it requires that MSOM outliers have larger residuals for models of comparable sizes. This relates to the signal-to-noise-ratio and it improves the
The heuristic argument $n > 5p$ which is often advocated for robust estimation methods (Rousseeuw and Van Zomeren 1990). The following result ensures that our proposal provides simultaneous feature selection and MSOM outlier detection consistency.

**Theorem 1** (Robust weak oracle property). Under all conditions in lists 1 and 2, and that $\log p = o((n-m_M)^2)$ and $\sqrt{n-m_M} \lambda \to \infty$ as $(n-m_M) \to \infty$. Then, there exist $k_n$ and a strict local minimizer of (6) such that the resulting robust estimates achieve:

1. Sparsity: $P\left(\hat{\beta}_{S_{\hat{\phi}}} = 0\right) \to 1$;

2. Bounded $L_\infty$-norm: $P\left(\|\hat{\beta}_{S_{\hat{\phi}}} - \beta_{S_{\hat{\phi}}}\|_\infty < (n-m_M)^\tau \log(n-m_M)\right) \to 1$;

3. MSOM detection: $P\left(\hat{S}_{\phi} = S_{\phi}\right) \geq P\left(\hat{\phi} = \phi_0\right) \to 1$.

Here the number of features in $\beta$ is allowed to exponentially increase with the (uncontaminated) sample size $n - m_M$. This is a robust version of the weak oracle property in the sense of Lv and Fan (2009) and Fan and Li (2012).

We remark that existing robust model selection procedures, which explicitly consider only MSOM outliers, can be cast into (3). However, differently from (6), they do not take into account the random structure of the problem, such as VIOM outliers. Relatedly, our approach can be naturally extended to high-dimensional mixed-effects linear models; however, this is left for future work. Moreover, regardless the presence of VIOMs, the use of nonconcave penalties in (6) provides an important bridge between existing trimming estimators, which promote sparsity in the feature space based on convex penalties (Kurnaz et al. 2017; Alfons et al. 2013), and the optimal approach based on $L_0$-constraints (Insolia et al. 2020). Unlike the former, our proposal achieves oracle properties under weaker assumptions, which can be particularly useful for the latter; e.g., to provide better warm-starts and big-$\mathcal{M}$ bounds, and accelerate convergence for MIP techniques.

### 3.2 Step 2: VIOM Detection

VIOM outlier detection, based on sparse estimation of $\gamma$ in (3), differs from sparse estimation of fixed effects ($\beta$ and $\phi$) due to their intrinsic randomness. Indeed, while underfitting $\gamma$, which results in undetected VIOMs, introduces bias in the estimated variance for the fixed effects in $\beta$, the inclusion of irrelevant $\gamma$ components, i.e., wrongly detected VIOMs, decreases the estimator efficiency.
In this section, based on the results from Section 3.1, we consider the augmented design matrix \( \mathbf{X} = [\mathbf{X}_{\mathcal{S}_g}, \mathbf{D}_{\mathcal{S}_g}] \), where \( \mathbf{X}_{\mathcal{S}_g} \) and \( \mathbf{D}_{\mathcal{S}_g} \) index the estimated \( k_p \) active features and \( k_n \) MSOM outliers, respectively. We further assume that \( n - k_n \geq k_p \), and that \( \mathbf{X}^T \mathbf{X} \) is an invertible matrix of size \( (k_p + k_n) \). The corresponding matrix of error contrasts is denoted as \( \mathbf{A} \), and \( \mathbf{P}_x \) is the counterpart of \( \mathbf{P}_x \) using \( \mathbf{X} \) in place of \( \mathbf{X} \).

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

\[
f \left( \gamma_{\mathcal{S}_y} | \mathbf{A}^T \mathbf{y} \right) = f \left( \mathbf{A}^T \mathbf{y} | \gamma_{\mathcal{S}_y} \right) f \left( \gamma_{\mathcal{S}_y} \right)
\]

\[
= (\mathbf{y} - \mathbf{D}_{\mathcal{S}_y} \gamma_{\mathcal{S}_y})^T \mathbf{P}_x (\mathbf{y} - \mathbf{D}_{\mathcal{S}_y} \gamma_{\mathcal{S}_y}) + \gamma_{\mathcal{S}_y}^T \boldsymbol{\Gamma}^{-1} \gamma_{\mathcal{S}_y}.
\]

However, (7) cannot be used to estimate \( \gamma \) as it relies on the unknown set of VIOM outliers \( \mathcal{S}_y \), as well as their covariance matrix \( \boldsymbol{\Gamma} \). We replace (7) with the following objective function

\[
\hat{\gamma} = \arg \min_{\gamma} (\mathbf{y} - \gamma)^T \mathbf{P}_x (\mathbf{y} - \gamma) + \gamma^T \mathbf{M}_{\gamma}^{-1} \gamma + (n - k_n) \sum_{i \in \mathcal{S}_\phi} R_\lambda(\gamma_i)
\]

(8)

where \( \mathbf{M}_{\gamma} \) is a proxy for \( \boldsymbol{\Gamma} \) (see the Supplementary Material for details). In principle the penalty function \( R_\lambda(\cdot) \) might differ from the one in (6), but for simplicity we consider nonconcave penalties such as SCAD also here.

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

**Conditions List 3** (VIOM reconstruction).

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

\[
\left\| (\mathbf{P}_x^{\mathcal{S}_y} + \boldsymbol{\Gamma}^{-1})^{-1} \right\|_{\infty} \leq \frac{(n - m_M)^{-1-\delta/2}}{R_\lambda'(\sqrt{n - m_M b_0}/2)}.
\]

\[
\max_{i \in \mathcal{S}_y \cap \mathcal{S}_\phi} \left\| \mathbf{P}_x^{\mathcal{S}_y} \mathbf{D}_{\mathcal{S}_y} (\mathbf{P}_x^{\mathcal{S}_y} + \boldsymbol{\Gamma}^{-1})^{-1} \right\|_2 \leq \frac{R_\lambda'(0+)}{R_\lambda' (\sqrt{n - m_M b_0}/2)}.
\]

B. **VIOM strength:** \( \sup \{ \sqrt{n - m_M b_0} \} R_\lambda''(t) = o ((n - m_M)^{-1}) \).
C. Proxy matrix: \( \Lambda_{\text{min}}(\mathcal{M}^S_{\Phi}) \geq 0 \) and \( \Lambda_{\text{min}}(\mathcal{M}^S_{\gamma} - \Gamma) \geq 0 \).

Similar conditions can be found in Fan and Li (2012) to perform feature selection on random effects using nonconcave penalties. The following result shows that our proposal detects VIOM outliers with asymptotic probability one, and effectively down-weights them.

**Theorem 2** (VIOM treatment). Under all conditions in lists 1-3, and that \( b_0(n - m_M)^{\delta - 1/2} \to \infty \) as \( (n - m_M) \to \infty \), there exists \( \lambda \) such that a strict local minimizer of (8) satisfies:

1. VIOM detection: \( P(\hat{S}_\gamma = S_\gamma) \to 1 \);
2. VIOM down-weighting: \( \max_{i \in S_\gamma} \| \hat{\gamma}_i - \gamma_i \| \leq (n - m_M)^{-\delta} \) for \( \delta \in (0, \frac{1}{2}) \).

**3.3 Step 3: Weights Estimation**

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

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

The following definition extends the robustly strong oracle property in the sense of Insolia et al. (2020) to the concurrent presence of MSOM and VIOM outliers.

**Definition 1** (Doubly robust strong oracle property). Let \( \mathcal{S} = \{S_\beta, S_\phi, S_\gamma\} \), and define the doubly robust strong oracle estimator \( \hat{\beta}_{\hat{S}} = \hat{\beta} | \hat{S} \) as the solution for \( \beta \) in (2). An estimator \( \hat{\beta}_{\hat{S}} \) satisfies the doubly robust strong oracle property if there exist tuning parameters which ensure \( P(\hat{S} = \mathcal{S}) \geq P(\hat{\beta}_{\hat{S}} = \hat{\beta}_S) \to 1 \) in the presence of MSOM and VIOM outliers.
The following result refines Theorems 1 and 2, and ensures that our proposal achieves the doubly robust strong oracle property – allowing us to rely on large sample inference.

**Theorem 3** (Doubly robust strong oracle property). Under all conditions in lists 1-3, as \((n - m_M) \rightarrow \infty\) there exist tuning parameters \(k_n\) and \(\lambda_s\)’s in (6) and (8) such that the resulting estimator plugging \(\hat{S}\) in (2) achieves:

1. **Asymptotic unbiasedness:**
   \[
   \|E\hat{\beta} - \beta_0\|_2^2 \leq 2P(\hat{S} \neq S) \left\{ \|\beta_0\|_2^2 + \lambda_M \left( \|\hat{W}^{1/2}X\beta_0\|_2^2 + \sigma^2 \text{tr}(\hat{W}) \right) \right\} \rightarrow 0
   \]
   where \(\text{tr}(\cdot)\) is the matrix trace, \(\lambda_M = \Lambda_{\max}\{(X^T_S\hat{W}XS_\beta)^+\} > 0\) and \(\{\hat{S}_\beta : \hat{S}_\beta \neq S_\beta\}\).

2. **Optimal MSE:**
   \[
   E\|\hat{\beta} - \beta_0\|_2^2 \leq \sigma^2 \text{tr}(\Sigma_X^{-1})/\text{tr}(\hat{W}) \\
   + 2P(\hat{S} \neq S) \left\{ (\lambda_M + \lambda_{M_s}) \left( \|\hat{W}^{1/2}X\beta_0\|_2^2 + \sigma^2 \text{tr}(\hat{W}) \right) \right\}
   \]
   where \(\lambda_{M_s} = \Lambda_{\max}\{(X^T_S\hat{W}XS_\beta)^-\} \) and \(\Sigma_X = (X^T_S\hat{W}XS_\beta)\).

3. **Asymptotic normality:**
   \[
   \sqrt{n}(\hat{\beta} - \beta_0) \rightarrow^d N(0, \sigma^2(\Sigma_X/n)^{-1}).
   \]

Importantly, this result provides also some intuition on the estimator’s behavior when it does not retrieve the doubly robust oracle solution, as well as in finite-sample settings. Indeed, points 1 and 2 in Theorem 3 depend on the probability of not recovering the true model, in terms of active features and/or outlying cases – which increases estimation biases and MSE. Finally, weights estimates obtained in Step 3 can be used to update the proxy matrices used in Sections 3.1 and 3.2, suggesting an iterative strategy whereby the process in Steps 1-3 is repeated improving model selection and estimation results (see Section 4). A similar approach was proposed in Fan and Li (2012) to select and estimate fixed and random effects; here our iteration includes an additional third step to update the weights.

### 3.4 A Heuristic Procedure

Here we present a computationally lean heuristic procedure similar to two-stage regression for mixed-models, which is inspired by our main proposal; namely:
1. Solve (6) using the proxy matrix $\mathcal{M}_R = I_n$. Let $y^* = y_{\hat{S}_\phi}$ and $X^* = X_{\hat{S}_\phi, \hat{S}_\beta}$ comprise response and predictor values restricted to the selected relevant features and non-outlying cases.

2. Consider again (6) using $y^*$, $X^*$ and $\gamma_{\hat{S}_\phi}$ in place of $y$, $X$ and $\phi$, respectively. Using $\mathcal{M}_R = I_{n-k_n}$ and leaving the estimation of $\beta$ unpenalized, solve the model relaxing the $L_0$-constraint (e.g., using SCAD or lasso). Let $\hat{\gamma}_{\hat{S}_\beta}$ indicate the resulting sparse estimates.

3. Consider $y^* = X^* \beta + \epsilon$ and, similar to Section 3.3, estimate weights for the units $i \in \hat{S}_\gamma$ using REMLE and use WLS to update the estimation of $\beta$.

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

We remark that Steps 1 and 2 of our heuristic procedure require a careful tuning process, which is critical to estimate the weights in a data-driven fashion and guarantee their “adaptiveness” (i.e., the breakdown point and the efficiency of the corresponding $\beta$ estimates). In the Supplementary Material we describe the robust BIC proposed for this tuning, and discuss connections between our heuristic procedure, ridge and $M$-estimation.

4 Simulation Study

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

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

$$\text{MSE}(\hat{\beta}_j) = \frac{1}{t} \sum_{i=1}^{t} (\hat{\beta}_{ij} - \beta_j)^2 = \frac{1}{t} \sum_{i=1}^{t} (\hat{\beta}_{ij} - \overline{\beta}_j)^2 + (\overline{\beta}_j - \beta_j)^2, \quad (9)$$

where $\overline{\beta}_j = \frac{1}{t} \sum_{i=1}^{t} \hat{\beta}_{ij}$, and we average the MSE across coefficients to produce $\text{MSE}(\hat{\beta}) = \frac{1}{p} \sum_{j=1}^{p} \text{MSE}(\hat{\beta}_j)$. (ii) For low-dimensional settings without MSOMs, we also consider the MSE of a weighted estimate of the error variance

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

where the $e_i$’s are the raw residuals and the $\hat{w}_i$’s the estimated weights. This takes into account weight estimates regardless of whether some units are in fact contaminated. The MSE decomposition for $\hat{s}^2$ is computed as in (9), with $\sigma_{\text{SNR}}^2$ and $\hat{s}^2$ replacing $\beta$ and $\overline{\beta}$, respectively. (iii) Let the non-zero entries of $\tau = \phi + \gamma$ indicate MSOMs and/or VIOMs. Outlier detection accuracy is measured in terms of false positive and false negative rates

$$\text{FPR}(\hat{\tau}) = \frac{|\{i \in \{1, \ldots, n\} : \hat{\tau}_i \neq 0 \land \tau_i = 0\}|}{|\{i \in \{1, \ldots, n\} : \tau_i = 0\}|}, \quad (10)$$

$$\text{FNR}(\hat{\tau}) = \frac{|\{i \in \{1, \ldots, n\} : \hat{\tau}_i = 0 \land \tau_i \neq 0\}|}{|\{i \in \{1, \ldots, n\} : \tau_i \neq 0\}|}. \quad (11)$$

These indicate the proportion of uncontaminated units wrongly detected as outliers, and of undetected contaminated units, respectively. (iv) For sparse settings, we also consider feature selection accuracy – which is measured in terms of FPR and FNR as in (10) and (11), using $\beta_j$ and $\hat{\beta}_j$ (for $j = 1, \ldots, p$) in place of $\tau_i$ and $\hat{\tau}_i$, respectively.

### 4.1 Scenario 1: Low-Dimensional VIOMs

Here we set $p = p_0 = 2$, with $\beta = (2, 2)^T$ and $\text{SNR} = 3$. The proportion of VIOM outliers is $m_\varepsilon / n = 0.25$ and $v = 10$. The sample size $n$ increases from 50 to 500 with 10 equispaced values. Data for each setting are replicated $t = 100$ times.
We consider the *oracle benchmark* (Opt), i.e., a WLS fit based on the true population weights $\mathbf{w}$, along with: (a) OLS, the ordinary least squares estimator; (b) LTS, the least trimmed sum of squares estimator with trimming set to the true $m_V/n$ (Maronna et al. 2006); (c) MM85, an MM-estimator using a preliminary LTS and Tukey’s bisquare loss function, with tuning constant set to achieve 85% nominal efficiency (Maronna et al. 2006); (d) MM95, as in (c), with 95% nominal efficiency; (e) FSRws, which utilizes a variant of forward search and single REMLE weights as described in Insolia et al. (2021); (f) Heur, our heuristic procedure (Section 3.4), where in Step 2 $\gamma$ is estimated by adaptive lasso initialized with OLS residuals, and in Step 3 each weight is estimated independently using REMLE as in FSRws; (g) SCADws, our main proposal (Section 3), where in Step 3 weights are estimated by a REMLE fit on the active random components of $\gamma$ detected by SCAD – as in FSRws and Heur, these weights are estimated independently.

Figure 1 shows the MSE for $\hat{\beta}$; SCADws and MM85 generally outperform other methods, Heur and MM95 perform comparably, FSRws improves on LTS and OLS (which perform poorly across sample sizes). Figure 2 shows the MSE for $s^2$. Notably, SCADws generally outperforms other methods, including the oracle estimator – likely because some VIOM outliers which are down-weighted by the latter do not carry sizeable residuals. Nevertheless, SCADws is capable of estimating full weights for these points. Relatedly, non-outlying cases with large residuals by chance are given full weight by the oracle estimator, but not necessarily by SCADws (see circled dots on the right panel of Figure 3). MM85 outperforms MM95, highlighting the drawbacks of $M$-estimators with pre-specified efficiency values. Heur performs comparably, although its estimates have larger biases, and it outperforms LTS and OLS, which provide strongly biased estimates because each point receives a binary or full weight. The performance of FSRws decreases for smaller sample sizes, where outliers are more often undetected.

The two left panels of Figure 3 show FPR and FNR for VIOM detection across methods, respectively. Overall, SCADws outperforms other methods; its decrease in terms of FPR along sample sizes is partially compensated by an increase in FNR. FSRws is close to SCADws for larger sample sizes, but for smaller ones it fails to detect some outliers (low FPR and high FNR). Heur performs similarly to SCADws, and MM-estimators perform poorly in these metrics due to a general down-weighting of all units. These trends demonstrate the ability of SCADws to detect truly outlying cases as the sample size increases. On the other hand, while FSRws tends to be more conservative across sample sizes, LTS has a more aggressive behavior resulting in larger FPR and lower FNR. The right panel of Figure 3 shows a scatterplot summarizing results for a typical simulation ($n = 500$). True VIOM outliers, as well
as the ones detected by SCADws, are highlighted.

4.2 Scenario 2: High-Dimensional VIOMs and MSOMs

Here we mimic Scenario 1, but we use sparse fixed effects in $\beta$ and introduce MSOM outliers. Specifically, we set $p = 30$ with $p_0 = 3$ active features. The proportions of VIOM and MSOM outliers are set to $m_V/n = 0.15$ and $m_M/n = 0.05$. Mean shifts
Figure 3: Scenario 1. Left: comparisons of FPR and FNR for outlier detection across procedures and sample sizes. Right: scatterplot summarizing results for a typical simulation with $n = 500$ – true VIOMs and VIOMs detected by SCADws are highlighted.

are set to $\mu_x = -10$ and $\mu_X = 10$ in order to create bad leverage points. The sample size $n$ ranges from 60 to 150 (with 10 equispaced values). Data for each setting are again replicated $t = 100$ times.

The oracle benchmark (Opt) is computed using population weights and the active feature set. In addition to it, we consider: (a) lasso; (b) sparseLTS (Alfons et al. 2013); (c) TaL, adaptive lasso with Tukey’s bisquare loss, a preliminary sparseLTS fit, and tuning constant fixed to achieve 85% nominal efficiency (Chang et al. 2018); (d) Heur, as in Scenario 1, but with a preliminary fixed-effects selection and MSOM detection using robust SCAD. (f) SCADws, as in Scenario 1, but with a preliminary fixed-effects selection and MSOM detection based on (6); (g) SCAD2s, two iterations of SCADws where weights estimated in the first iteration are used to update the proxy matrices and re-run our 3-step procedure; (h) SCADopt, similar to SCADws, but with proxy matrices built with VIOM population weights; For simplicity, robust methods all use the true trimming level $m_M/n$.

Figure 4 shows the MSE for $\hat{\beta}$. As expected, SCADopt resembles very closely the oracle estimator. SCAD2s, which improves upon SCADws, outperforms other feasible estimation methods. TaL performs comparably but has higher biases, and Heur improves upon sparseLTS. Lasso breaks down due to the presence of MSOM outliers.

The left panels of Figure 5 show FPR and FNR for outlier detection. Unlike the oracle estimator, SCADopt is capable of estimating full weights for VIOMs with negligible residuals (higher FNR), and it is not prone to detecting non-outliers with large residuals by chance (very low FPR). Notably though, although weights need to be estimated, also SCADws and SCAD2s perform well in both these metrics.
SCAD2s reduces FPR and slightly increases FNR compared to SCADws, which results is an overall performance increase for the iterative approach. Heur provides larger FPR and smaller FNR. SparseLTS has FPR equal to 0 and large FNR, as it detects only extreme MSOM outliers. TaL performs poorly due to a general down-weighting of all points.

The right panels of Figure 5 show FPR and FNR for feature selection. SCADopt performs comparably to the oracle estimator. SCAD2s, which improves upon SCADws, generally outperforms other methods. TaL produces higher FPR across sample sizes, and Heur provides denser solutions – but still sparser than sparseLTS. Lasso performs poorly also here, since it breaks down. We note that most robust methods are at times affected by MSOMs for smaller sample size (larger FNR and MSE) where their detection is harder.

5 An application to the Boston Housing Data

The Boston Housing dataset (http://lib.stat.cmu.edu/datasets/boston) contains $n = 506$ housing location and 13 predictors; namely: 1. crim (the per capita crime rate), 2. zn (the proportion of residential land zoned for lots over 25,000 sq.ft), 3. indus (the proportion of non-retail business acres), 4. chas (a “Charles River” dummy), 5. nox (the nitrogen oxides concentration in parts per 10 million), 6. rm (the average number of rooms per dwelling), 7. age (the proportion of owner-occupied units built prior to 1940), 8. dis (a weighted mean distance to five Boston employ-
Figure 5: Scenario 2. Comparisons of FPR and FNR for outlier detection (left) and feature selection (right) across procedures and sample sizes.

Using all predictors plus an intercept, we applied the LTS estimator with increasing trimming and computed the robust BIC (see Supplementary Material). This helps identify a reasonable trimming level to use across different methods. The left panel of Figure 6 shows that the curve flattens for low levels, with a noticeable drop only for very small amounts of trimming. With a conservative 10% trimming, we used SCAD2s to select the relevant features on the full dataset. These are the predictors number 1, 5, 6, 8, 9, 10, 11, 12, 13 (plus the intercept). The central panel of Figure 6 shows the robust BIC recomputed on these features alone. There is some evidence of both MSOM outliers (the curve achieves a maximum around 5% trimming) and VIOM outliers (the curve flattens starting from 15-10%). Using again 10% trimming, the right panel of Figure 6 shows the residuals obtained by SCAD2s on the full dataset. Cases detected as MSOM and VIOM outliers are highlighted.

Next, we extended the analysis along lines similar to Chang et al. (2018). We considered 20 random splits of the data in training and testing sets (300 and 206 units, respectively). Based on the observations above we used again 10% trimming across robust methods. The left panel of Figure 7 shows box-plots of the sparsity levels, i.e., the number of features retained by different methods, across the 20 random training sets. Some methods do not provide sparse estimates by definition, but also lasso and our heuristic proposal provide very dense solutions. TaL and sparseLTS provide, respectively, sparser and denser solutions compared to SCAD2s and SCADws. SCAD2s appears to induce slightly more sparsity than SCADws. The
Figure 6: Left: robust BIC computed on all points and features. Center: robust BIC computed on all points and only the features selected using SCAD2s. Right: SCAD2s residuals labeled as non-outlying (blue), MSOM (red), and VIOM (green).

right panel of Figure 7 shows the distribution of the selected features across the 20 random training sets. The solution for SCAD2s is in line with prior analyses and, unlike TaL, supports the relevance of predictors number 8 and 9 (dis and rad).

Figure 7: Box-plots of the estimated sparsity levels (left) and distribution of the selected features for sparse methods (right) across 20 random training sets for different methods.

Figure 8 compares the prediction accuracy of different methods across the 20 random training/testing splits based on the mean absolute (MAPE) and trimmed mean squared (TMSPE) prediction errors, with an upper 10% trimming. SCADws and SCAD2s provide a good trade-off between model parsimony and prediction accuracy. They outperform TaL (the only method generating sparser solutions) in terms of prediction, independently of the considered quantile. Our heuristic procedure performs very well – often better than non-sparse robust estimators – in terms
of prediction, but it has very dense solutions.

Figure 8: Box-plots of MAPE (left) and TMSPE (right) across 20 random training/testing splits for different methods.

6 Final Remarks

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

The theoretical results characterizing our proposal include its high breakdown point, a robust oracle property – which allows the number of feature to increase exponentially with the sample size – and the detection of each type of outliers with probability tending to one. Moreover, including a computationally cheap extra step, our proposal achieves a doubly strong oracle property. This provides optimal units’ weights and thus an optimal trade-off between high-breakdown point and efficiency.

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

Moreover, as our theoretical results critically rely on tuning parameters controlling the trade-off between sparsity and efficiency, we are interested in the development of suitable information criteria for sparse models affected by different sources of contamination, extending the robust BIC introduced in this work. We are also developing more effective ways to build proxy matrices used in our procedure, as well as iterative approaches. Finally, we are exploring how to include into our framework cellwise contamination (Alqallaf et al. 2009), which is recently receiving a lot of attention for high-dimensional settings.

SUPPLEMENTARY MATERIAL

Appendix A: Theoretical Results

Proof of Proposition 1. For any trimming level \(k_n\), the objective function in (6) subject to integer constraints in (6a) can be equivalently formulated as

\[
Q(\hat{\beta}) = \frac{1}{2} \sum_{i=1}^{n-k_n} [(y_i^* - \beta^T x_i^*)^2]_{i:n} + (n - k_n) \sum_{j=1}^{p} R_{\lambda}(|\beta_j|) \tag{A.1}
\]

where \((t_1)_{1:n} \leq \ldots \leq (t_n)_{n:n}\) denote the order statistics of \(t_i\), \(y^* = \sqrt{M_R} y\) and \(X^* = \sqrt{M_R} X = (x_1^*, \ldots, x_n^*)^T\). This relies on the fact that a weighted regression of \(y\) on \(X\) is equivalent to an unweighted regression of \(y^*\) on \(X^*\), and we also use Proposition 1 in Insolia et al. (2020) to transform the mean-shift model based on \(\phi\) to a trimmed loss problem without explicit mean shift parameters. Then, denote the contaminated dataset as \(\tilde{Z} = [\tilde{y}, \tilde{X}] = [(y + \Delta_y), (X + \Delta_X)]\). We first show that the BdP \(\varepsilon^* \geq (n - k_n + 1)/n\), and then \(\varepsilon^* \leq (n - k_n + 1)/n\).

For the first part of the proof assume that \(\tilde{Z}\) contains \(m_M \leq k_n\) outliers. Consider \(\hat{\beta} = 0\), so that the associated loss

\[
Q(0) = \sum_{i=1}^{n-k_n} (y_i^2)_{i:n} \leq \sum_{i=1}^{n-k_n} (y_i^2)_{i:n} \leq (n - k_n) M_y^2,
\]

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

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

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

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

$$Q(\hat{\beta}) = \sum_{i=1}^{n-m_M} [(\bar{y}_i^* - \bar{\beta}^T \bar{x}_i^*)^2]_{i:n} + \sum_{h=n-m_M+1}^{n-k_n} [(\bar{y}_h^* - \bar{\beta}^T \bar{x}_h^*)^2]_{h:n} + (n - k_n) \sum_{j=1}^{p} R_\lambda(|\hat{\beta}_j|)$$

$$\geq \left[(\bar{y}_i^* - \bar{\beta}^T \bar{x}_i^*) + (\Delta y_i - \bar{\beta}^T \Delta x_i)\right]_{i=n-m_M+1} + (n - k_n) \sum_{j=1}^{p} R_\lambda(|\hat{\beta}_j|) \quad (A.2)$$

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

**Proof of Theorem 1.** It extends Theorem 1 in Fan and Li (2012) to the presence of MSOM contamination. Specifically, we can use the same argument, but their conditions must hold at least on $n - m_M$ (uncontaminated) points as opposed to $n$. Since $k_n$ largest residuals (say, $k_n = m_M$) are always discarded from our loss in (6), we thus need to ensure that these trimmed points encompass MSOM outliers. Condition 2(D) guarantees this, similarly to Theorem 3 in Insolia et al. (2020), so that MSOM outliers have largest residuals for any model of size $k_p \leq p_0$. See Fan and Li (2012) for details. ■

25
Proof of Theorem 2. This result immediately follows from Theorem 2 in Fan and Li (2012) specifically focusing on VIOM outliers as random effects (i.e., our term $I_n \gamma$ instead of $Z b$). However, in Fan and Li (2012) the dimension of the random effects $b$ can increase exponentially with the sample size $n$, but in our formulation $\gamma$ can only increase linearly with $n - k_n$. Thus, our conditions in list 2 might be relaxed to account only for VIOMs. Nevertheless, these more general conditions allow one to extend our results also to the presence of additional (pure) random effects, whose size can increase exponentially with $n - k_n$. ■

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

Here with a slight abuse of notation, we denote $P(S) = P(\hat{S} = S)$ and $P(\hat{S}) = P(\hat{S} \neq S)$, where $\hat{S} = \{\hat{S}_\beta, \hat{S}_\phi, \hat{S}_\gamma\}$. Furthermore, we indicate as $\hat{\beta}|\hat{S}$ the estimated coefficients conditionally on the selected model, which is abbreviated as $\hat{\beta}_{\hat{S}}$. It is also assumed that, conditioned on any selected model $\hat{S}$, units weights $\hat{W}$ are deterministic.

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

$$
\|E\hat{\beta} - \beta_0\|_2^2 = \|E\hat{\beta}_S P(S) + E\hat{\beta}_{\hat{S}} P(\hat{S}) - \beta_0\|_2^2 \\
\leq 2\|E\hat{\beta}_S P(S) - \beta_0\|_2^2 + 2\|E\hat{\beta}_{\hat{S}} P(\hat{S})\|_2^2 \\
= 2\|E\{(X^T S \hat{W} X_S)^+ X^T S \hat{W} y\} P(S) - \beta_0\|_2^2 + 2P(\hat{S})\|E\hat{\beta}_{\hat{S}}\|_2^2 \\
= 2\|P(S)\beta_0 - \beta_0\|_2^2 + 2P(\hat{S})\|E\hat{\beta}_{\hat{S}}\|_2^2 \\
= 2\|\beta_0 P(S) - 1\|_2^2 + 2P(\hat{S})\|E\hat{\beta}_{\hat{S}}\|_2^2 \\
= 2P(\hat{S})\{\|\beta_0\|_2^2 + \|E\hat{\beta}_{\hat{S}}\|_2^2\}. 
$$

(A.3)

Further, using Jensen’s inequality and the fact that $\|Ab\| \leq \|A\|\|b\|$ provides

$$
\|E\hat{\beta}_{\hat{S}}\|_2 \leq E\|(X^T S \hat{W} X_S)^+ X^T S \hat{W} y\|_2 \\
\leq E\| (X^T S \hat{W} X_S)^+ X^T S \hat{W}^{1/2} \hat{W}^{1/2} y\|_2^2 \\
= \Lambda_{\text{max}} \{(X^T S \hat{W} X_S)^+\} E\|\hat{W}^{1/2} X \beta_0 + \hat{W}^{1/2} \epsilon\|_2^2 \\
= \Lambda_{\text{max}} \{(X^T S \hat{W} X_S)^+\} E(\|\hat{W}^{1/2} X \beta_0\|_2^2 + \epsilon^T \hat{W} \epsilon) \\
= \Lambda_{\text{max}} \{(X^T S \hat{W} X_S)^+\} (\|\hat{W}^{1/2} X \beta_0\|_2^2 + \text{tr}(\hat{W}) \sigma^2) \\
\leq \Lambda_{\text{max}} \{(X^T S \hat{W} X_S)^+\} (\|X \beta_0\|_2^2 + n\sigma^2),
$$

(A.4)
where \( \lambda_{\text{max}}(\cdot) \) denotes the largest eigenvalue, and for a real matrix \( A \), the spectral norm \( \| A \|_2 = \sqrt{\lambda_{\text{max}}(A A^T)} = \sqrt{\lambda_{\text{max}}(A^T A)} \). In our case,

\[
\| (X_S^T \tilde{W} X_S)^+ X_S^T \tilde{W}^{1/2} \|_2^2 = \lambda_{\text{max}} \{ (X_S^T \tilde{W} X_S)^+ X_S^T \tilde{W} X_S (X_S^T \tilde{W} X_S)^+ \}
\]

\[
= \lambda_{\text{max}} \{ (X_S^T \tilde{W} X_S)^+ \},
\]

where the last equality follows from the property of a generalized inverse \( A^+ A A^+ = A^+ \). Combining (A.3) and (A.4) leads to the desired results. \( \blacksquare \)

**Proof of Theorem 3(2).** Introducing the WLS oracle estimator \( \hat{\beta}_0 \) and using the fact that

\[
E \| \hat{\beta}_0 - \beta_0 \|_2 = E \| (X_S^T \tilde{W} X_S)^+ X_S^T \tilde{W} \varepsilon \|_2 = 0
\]

provides

\[
E \| \tilde{\beta} - \beta_0 \|_2^2 = E \| \tilde{\beta} + \hat{\beta}_0 - \hat{\beta}_0 - \beta_0 \|_2^2
\]

\[
= E \| \tilde{\beta} - \hat{\beta}_0 \|_2^2 + E \| \hat{\beta}_0 - \beta_0 \|_2^2
\]

\[
= E \| \tilde{\beta} - \hat{\beta}_0 \|_2^2 + \sigma^2 \text{tr}(\Sigma_X^{-1}) / \text{tr}(\tilde{W})
\]

(A.5)

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

\[
E \| \tilde{\beta} - \hat{\beta}_0 \|_2^2 = E \| \tilde{\beta} - \hat{\beta}_0 \|_2^2 P(S) + E \| \hat{\beta}_0 - \beta_0 \|_2^2 P(\tilde{S})
\]

(A.6)

where the first equality relies on the law of total expectations and the last one uses the fact that \( \tilde{\beta}_\tilde{S} = \hat{\beta}_0 \) conditioned on \( \{ \tilde{S} = S \} \).

Further, note that

\[
E \| \tilde{\beta}_\tilde{S} - \beta_0 \|_2^2 \leq 2 \{ E \| \tilde{\beta}_\tilde{S} \|_2^2 + E \| \hat{\beta}_0 \|_2^2 \}
\]

\[
= 2 \{ E \| (X_S^T \tilde{W} X_S)^+ X_S^T \tilde{W} \varepsilon \|_2^2 + E \| (X_S^T \tilde{W} X_S)^+ X_S^T \tilde{W} \varepsilon \|_2^2 \}
\]

\[
\leq 2 E \| \tilde{W}^{1/2} \varepsilon \|_2^2 \left[ \lambda_{\text{max}} \{ (X_S^T \tilde{W} X_S)^+ \} + \lambda_{\text{max}} \{ (X_S^T \tilde{W} X_S)^+ \} \right],
\]

(A.7)

where the first upper bound follows from \( \| a + b \|_2^2 \leq 2(\| a \|_2^2 + \| b \|_2^2) \), and the second one uses \( \| A b \| \leq \| A \| \| b \| \). Finally, combining

\[
E \| \tilde{W}^{1/2} \varepsilon \|_2^2 \leq E (\| \tilde{W}^{1/2} X \beta_0 \|_2^2 + \varepsilon^T \tilde{W} \varepsilon) = \| \tilde{W}^{1/2} X \beta_0 \|_2^2 + \text{tr}(\tilde{W}) \sigma^2 \leq \| X \beta_0 \|_2^2 + n \sigma^2
\]

with (A.5), (A.6), and (A.7) concludes the proof. \( \blacksquare \)
Proof of Theorem 3(3). Under the conditions in lists 1-3, as \((n - m_M) \to \infty\), it follows that \(P(\hat{S} = S) \to 1\) for some suitable constants. Thus, \(\hat{\beta}\) has an asymptotic normal distribution as it is a linear combination of normal distributions

\[
\hat{\beta} = (X_S^T \hat{W} X_S)^{-1} X_S^T \hat{W} y \\
= (X_S^T \hat{W} X_S)^{-1} X_S^T \hat{W} (X_S \beta_0 + \epsilon) \\
= \beta_0 + (X_S^T \hat{W} X_S)^{-1} X_S^T \hat{W} \epsilon \sim N(\beta_0, \sigma^2(X_S^T \hat{W} X_S)^{-1}),
\]

and \(\hat{W} = V^{-1}\) guarantees that it asymptotically reaches maximum efficiency.

\section*{Appendix B: Technical Details}

\subsection*{B.1 Choice of the Proxy Matrix \(M\)}

For mixed-effects linear models without data contamination as in Section 2.2, \textcite{FanLi2012} propose to replace \(\sigma^{-2} B\) in (5) with a proxy matrix \(M_b\). They show that under mild conditions it is safe to choose \(M_b = \log(n) I_n\), as the eigenvalues of \(Z^T P_Z Z\) and \(Z Z^T\) have magnitude increasing with \(n\), so that they are likely to dominate the eigenvalues of \(M_b\) for a large enough \(n\). While this choice excludes cross-correlations in the random effects, it avoids the estimation of a large number of parameters as in the case of an unstructured covariance matrix.

In our formulation the terms \(M_R\) and \(M_\gamma\) in (6) and (8) are proxies for the unknown \(P_R\) and \(\Gamma\), respectively. Following \textcite{FanLi2012}, in our implementation we use \(M_R = M_\gamma = \log(n) I_n\) on the first iteration. If the 3-step procedure is re-iterated, such as in SCAD2s, we use estimated weights \(\hat{W}\) from the previous iteration for their update.

\subsection*{B.2 Weights Estimation}

The formulation in (8) highlights that if \(\hat{\gamma}_i = 0\) also the corresponding variance inflation \(\hat{\omega}_i = 0\). However, it might be of interest to estimate \(\omega_i\) when the corresponding \(\hat{\gamma}_i \neq 0\). A similar reasoning holds for step 3 of the heuristic method described in Section 3.4. Note that

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

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

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

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

### B.3 Parameter Tuning

For feature selection and MSOM detection is essential to tune the sparsity level and the amount of trimming. We propose to combine the approach in Insolia et al. (2020) and Riani et al. (2021). Specifically, in low-dimensional models affected by MSOM contamination, Riani et al. (2021) introduced the following robust version of BIC to tune the trimming level for hard-trimming estimators:

\[
\text{BICW} = -n \log \left\{ \frac{R(\hat{\beta}_h)}{(\sigma_h^2 h)} \right\} - \{p + k_n\} \log n,
\]

where \( h = n - k_n \) and \( R(\hat{\beta}_h) \) is the residual sum of square based on the \( h \) observations contributing to the loss. The associated variance of the truncated normal distribution containing a central portion \( h/n \) of the full distribution is

\[
\sigma^2(h) = 1 - \frac{2n}{h} \Phi^{-1} \left( \frac{n + h}{2n} \right) \phi \left\{ \Phi^{-1} \left( \frac{n + h}{2n} \right) \right\},
\]
where \( \phi(\cdot) \) and \( \Phi(\cdot) \) denote the probability cumulative density function for the Gaussian distribution, respectively.

In our work, we consider the following extension of BICW for high-dimensional settings:

\[
\text{BICr} = -n \left[ \log \left( \frac{R(\hat{\beta}_h)}{(\sigma_h^2 h)} \right) \right] - \{k_p + k_n\} \log n,
\]

where \( k_p = |\hat{S}_\beta| \) denotes the sparsity level for feature selection. This formulation improves and extends the robust BIC proposed in Insolia et al. (2020). In principle one should consider a range of trimming values \( k_n \) and shrinkage parameter \( \lambda \) (the latter determines \( k_p \)). However, to reduce the computational burden, we often fix one of the two parameters and tune only the other. Moreover, to take into account the co-occurrence of VIOM outliers this might be generalized further, similarly to the CAIC and extended CAIC discussed in Section 2.2.

### B.4 Parallel Between our Heuristic Approach and M-estimation

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

Step 1 is equivalent to an adaptive hard-trimming, sparse estimator (i.e., it selects features and assigns binary weights) and guarantees an high-breakdown point. This step aims to exclude MSOMs and select only the relevant features (see for instance Alfons et al. 2013; Kurnaz et al. 2017; Insolia et al. 2020). Step 2 corresponds to an adaptive “truncated” M-estimator, where only the most extreme cases are down-weighted. In full generality, this estimator takes the form \( \hat{\beta} = \arg \min_{\beta} \sum_{i=1}^{n} \rho(e_i/\sigma). \) Here the idea is that the \( n - m_M - m_V \) uncontaminated points receive full weights as in OLS, but only VIOMs are down-weighted according to the \( \rho(\cdot) \) function in use, and MSOMs (if present) are excluded from the fit. For instance, this has a parallel with the hyperbolic tangent \( \rho(\cdot) \) function, which can be considered as refinement of Hampel’s piecewise linear redescending function and is related to the change of variance curve (Hampel et al. 1981). Tanh-estimators are more easily defined in terms of their derivatives, and the corresponding \( \psi(\cdot) \) function is

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

for suitable constants \( k, A, B, c_1, \) and \( c_2, \) where \( 0 < c_1 < c_2 \) satisfies

\[
c_1 = \{A(k - 1)\}^{1/2} \tanh \frac{1}{2} \left\{ (k - 1)B^2/A \right\}^{1/2} (c_2 - c_1).
\]
These constants are traditionally computed iteratively, based on the Newton-Raphson algorithm and numerical integration. Figure B.1 shows the corresponding $\rho$, $\psi$, and weight functions for $c_2 = 4$ and $k = 4.5$.

Unlike tanh-estimators, our heuristic proposal does not pre-specify a trade-off between breakdown point and efficiency, but this is adaptively tuned as follows. The rejection point $c_2$ approximately corresponds to the smallest standardized residual for the MSOMs detected at step 1. Similarly, the constant $c_1$ is set to the value of the largest standardized residual for points which are not affected by MSOM or VIOM. Specifically, for our heuristic proposal, $c_1$ and $c_2$ can be computed based on order statistics from the scaled residuals obtained at step 1. Ideally, assuming without loss of generality that all outliers have sizeable residuals, these corresponds to the $(n - m_V - m_M)$-th and $(n - m_M)$-th order statistics of the absolute standardized residuals, respectively.

**Appendix C: Code**

Our code is available upon request.

**References**

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

Alqallaf, F. A., Van Aelst, S., Yohai, V. J. and Zamar, R. H. (2009), ‘Propagation of outliers in multivariate data’, *The Annals of Statistics* 37(1), 311–331.
Amato, U., Antoniadis, A., De Feis, I. and Gijbels, I. (2021), ‘Penalised robust estimators for sparse and high-dimensional linear models’, *Statistical Methods & Applications* **30**(1), 1–48.

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

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

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

Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010), ‘Joint variable selection for fixed and random effects in linear mixed-effects models’, *Biometrics* **66**(4), 1069–1077.

Buscemi, S. and Plaia, A. (2020), ‘Model selection in linear mixed-effect models’, *AStA Advances in Statistical Analysis* **104**(4), 529–575.

Chang, L., Roberts, S. and Welsh, A. (2018), ‘Robust lasso regression using Tukey’s biweight criterion’, *Technometrics* **60**(1), 36–47.

Cook, R. D., Holschuh, N. and Weisberg, S. (1982), ‘A note on an alternative outlier model’, *Journal of the Royal Statistical Society: Series B (Methodological)* **44**(3), 370–376.

Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall, New York.

Donoho, D. L. and Huber, P. J. (1983), The notion of breakdown point, in P. Bickel, K. A. Doksum and J. L. Hodges, eds, ‘A festschrift for Erich L. Lehmann’, Wadsworth, pp. 157–184.

Fan, J. and Li, R. (2001), ‘Variable selection via nonconcave penalized likelihood and its oracle properties’, *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fan, J. and Lv, J. (2011), ‘Nonconcave penalized likelihood with NP-dimensionality’, *IEEE Transactions on Information Theory* **57**(8), 5467–5484.

Fan, Y. and Li, R. (2012), ‘Variable selection in linear mixed effects models’, *The Annals of Statistics* **40**(4), 2043–2068.

Freue, G. V. C., Kepplinger, D., Salibián-Barrera, M. and Smucler, E. (2019), ‘Robust elastic net estimators for variable selection and identification of proteomic biomarkers’, *The Annals of Applied Statistics* **13**(4), 2065–2090.
Grandvalet, Y. (1998), Least absolute shrinkage is equivalent to quadratic penalization, in ‘International Conference on Artificial Neural Networks’, Springer, pp. 201–206.

Gumedze, F. N. (2019), ‘Use of likelihood ratio tests to detect outliers under the variance shift outlier model’, *Journal of Applied Statistics* 46(4), 598–620.

Gumedze, F. N., Welham, S. J., Gogel, B. J. and Thompson, R. (2010), ‘A variance shift model for detection of outliers in the linear mixed model’, *Computational Statistics & Data Analysis* 54(9), 2128–2144.

Hampel, F. R., Rousseeuw, P. J. and Ronchetti, E. (1981), ‘The change-of-variance curve and optimal redescending M-estimators’, *Journal of the American Statistical Association* 76(375), 643–648.

Hoerl, A. E. and Kennard, R. W. (1970), ‘Ridge regression: applications to nonorthogonal problems’, *Technometrics* 12(1), 69–82.

Ibrahim, J. G., Zhu, H., Garcia, R. I. and Guo, R. (2011), ‘Fixed and random effects selection in mixed effects models’, *Biometrics* 67(2), 495–503.

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

Insolia, L., Kenney, A., Chiaromonte, F. and Felici, G. (2020), ‘Simultaneous feature selection and outlier detection with optimality guarantees’, *arXiv preprint arXiv:2007.06114*.

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* pp. 1–12.

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

Laird, N. M. and Ware, J. H. (1982), ‘Random-effects models for longitudinal data’, *Biometrics* 38(4), 963–974.

Liang, H., Wu, H. and Zou, G. (2008), ‘A note on conditional AIC for linear mixed-effects models’, *Biometrika* 95(3), 773–778.

Lin, X. (1997), ‘Variance component testing in generalised linear models with random
effects’, *Biometrika* **84**(2), 309–326.

Liu, H., Yao, T. and Li, R. (2016), ‘Global solutions to folded concave penalized nonconvex learning’, *The Annals of Statistics* **44**(2), 629–659.

Liu, H. and Yu, B. (2013), ‘Asymptotic properties of lasso+mLS and lasso+ridge in sparse high-dimensional linear regression’, *Electronic Journal of Statistics* **7**, 3124–3169.

Loh, P. (2017), ‘Statistical consistency and asymptotic normality for high-dimensional robust $M$-estimators’, *The Annals of Statistics* **45**(2), 866–896.

Lv, J. and Fan, Y. (2009), ‘A unified approach to model selection and sparse recovery using regularized least squares’, *The Annals of Statistics* **37**(6A), 3498–3528.

Maronna, R. A. (2011), ‘Robust ridge regression for high-dimensional data’, *Technometrics* **53**(1), 44–53.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, John Wiley & Sons, Ltd.

Müller, S., Scealy, J. L. and Welsh, A. H. (2013), ‘Model selection in linear mixed models’, *Statistical Science* **28**(2), 135–167.

Patterson, H. D. and Thompson, R. (1971), ‘Recovery of inter-block information when block sizes are unequal’, *Biometrika* **58**(3), 545–554.

Peng, H. and Lu, Y. (2012), ‘Model selection in linear mixed effect models’, *Journal of Multivariate Analysis* **109**, 109–129.

Riani, M., Atkinson, A. C., Corbellini, A. and Fabrizio, L. (2021), ‘Information criteria for outlier detection avoiding arbitrary significance levels’, *Submitted*.

Rousseeuw, P. J. and Van Zomeren, B. C. (1990), ‘Unmasking multivariate outliers and leverage points’, *Journal of the American Statistical Association* **85**(411), 633–639.

Schelldorfer, J., Bühlmann, P. and van De Geer, S. (2011), ‘Estimation for high-dimensional linear mixed-effects models using $\ell_1$-penalization’, *Scandinavian Journal of Statistics* **38**(2), 197–214.

She, Y. and Owen, A. B. (2011), ‘Outlier detection using nonconvex penalized regression’, *Journal of the American Statistical Association* **106**(494), 626–639.

Smucler, E. and Yohai, V. J. (2017), ‘Robust and sparse estimators for linear regression models’, *Computational Statistics & Data Analysis* **111**, 116–130.
Thompson, R. (1985), ‘A note on restricted maximum likelihood estimation with an alternative outlier model’, *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1), 53–55.

Tibshirani, R. (1996), ‘Regression shrinkage and selection via the lasso’, *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Zhang, C. (2010), ‘Nearly unbiased variable selection under minimax concave penalty’, *The Annals of Statistics*, 38(2), 894–942.

Zioutas, G. and Avramidis, A. (2005), ‘Deleting outliers in robust regression with mixed integer programming’, *Acta Mathematicae Applicatae Sinica*, 21(2), 323–334.

Zou, H. (2006), ‘The adaptive lasso and its oracle properties’, *Journal of the American Statistical Association*, 101(476), 1418–1429.

Zou, H. and Li, R. (2008), ‘One-step sparse estimates in nonconcave penalized likelihood models’, *The Annals of Statistics*, 36(4), 1509–1533.