TEST (2022) 31:563–594
https://doi.org/10.1007/s11749-021-00792-w

ORIGINAL PAPER

Penalized robust estimators in sparse logistic regression

Ana M. Bianco[1] · Graciela Boente[2] · Gonzalo Chebi[2]

Received: 17 October 2020 / Accepted: 20 October 2021 / Published online: 12 November 2021

Abstract

Sparse covariates are frequent in classification and regression problems where the task of variable selection is usually of interest. As it is well known, sparse statistical models correspond to situations where there are only a small number of nonzero parameters, and for that reason, they are much easier to interpret than dense ones. In this paper, we focus on the logistic regression model and our aim is to address robust and penalized estimation for the regression parameter. We introduce a family of penalized weighted M-type estimators for the logistic regression parameter that are stable against atypical data. We explore different penalization functions including the so-called Sign penalty. We provide a careful analysis of the estimators convergence rates as well as their variable selection capability and asymptotic distribution for fixed and random penalties. A robust cross-validation criterion is also proposed. Through a numerical study, we compare the finite sample performance of the classical and robust penalized estimators, under different contamination scenarios. The analysis of real datasets enables to investigate the stability of the penalized estimators in the presence of outliers.

Keywords Logistic regression · Outliers · Penalty functions · Robust estimation · Sparse models

Mathematics Subject Classification 62F35 · 62F12

✉ Graciela Boente
gboente@dm.uba.ar

Ana M. Bianco
abianco@dm.uba.ar

Gonzalo Chebi
gonzalo.chebi@gmail.com

[1] Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Ciudad Universitaria, Pabellón 2, Buenos Aires 1428, Argentina

[2] Departamento de Matemáticas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Ciudad Universitaria, Pabellón 1, Buenos Aires 1428, Argentina

1 Introduction

Sparse regression models assume that the number of actually relevant predictors, k, is lower than the number of measured covariates. Hastie et al. (2015) describe that *a sparse statistical model is one in which only a relatively small number of parameters (or predictors) play an important role*, leading to models that are much easier to interpret than dense ones. This type of models has raised a paradigm shift in statistics, since the traditional approach to classical issues such as regression or classification assumes that no restrictions are imposed when estimating the parameters. In these circumstances, penalized regression estimators are a useful tool when the practitioner is interested in automatic variable selection. We refer to (Efron and Hastie 2016) for an overview of adapted inference methods. For instance, the ℓ_1 regularization, which is related to the LASSO estimators introduced in Tibshirani (1996), bets on the sparsity principle and is effective for variable selection, but tends to choose too many features. Zou and Hastie (2005) considered an alternative regularization, namely the Elastic Net penalty, which combines both ℓ_1 and ℓ_2 norms. Elastic Net preserves the sparsity of LASSO and maintains some of the desirable predictive properties of Ridge regression. Fan and Li (2001) and Zhang (2010) proposed alternative penalties which lead to sparse estimators.

Logistic regression is a widely studied problem in statistics and has been useful to classify data. It is well known that in the non-sparse scenario the maximum likelihood estimator (MLE) of the regression coefficients is very sensitive to outliers, meaning that we cannot accurately classify a new observation based on these estimators, neither identify those covariates with important information for assignation. Robust methods for logistic regression bounding the deviance have been proposed in Bianco and Yohai (1996). In particular, for the family of estimators defined therein, (Croux and Haesbroeck 2003) introduced a loss function that guarantees the existence of the resulting robust estimator when the maximum likelihood estimators do exist. The proposal due to (Basu et al. 2017) on the basis of minimum divergence can also be seen as a particular case of the (Bianco and Yohai 1996) estimator with a properly defined loss function. Other approaches were given in Cantoni and Ronchetti (2001) and Bondell (2005, 2008). However, all these methods are not reliable under collinearity and they do not allow for automatic variable selection when only a few number of covariates are relevant. The previous ideas on regularization can be directly extended to logistic regression.

In the last decade, some robust estimators for logistic regression in the sparse regressors framework have been proposed in the literature. Among others, we can mention (Chi and Scott 2014) who considered a least squares estimator with a Ridge and Elastic Net penalty and (Kurnaz et al. 2018) who proposed estimators based on a trimmed sum of the deviances with an Elastic Net penalty. It is worth noticing that the least squares estimator in logistic regression corresponds to a particular choice of the loss function considered in Bianco and Yohai (1996). Finally, Tibshirani and Manning (2013) introduced a real-valued shift factor to protect against the possibility of mislabelling, while (Park and Konishi 2016) considered a weighted deviance approach with weights based on the Mahalanobis distance computed over a lower-dimensional principal component space and included an Elastic Net penalty. Most of the asymptotic results for robust sparse estimators have been given under the linear regression model (see, for example, Smucler and Yohai 2017) or when considering a convex loss function (see, for instance, van de Geer and Müller 2012). More recently, (Avella-Medina and Ronchetti 2018) treated the situation of general penalized M-estimators in shrinking neighbourhoods, when the parameter dimension p is fixed. In this setting, they considered penalties that are a deterministic sum of univariate functions and showed that penalized M-estimators based on loss functions with a bounded derivative behave better in a neighbourhood of the model than the classical oracle estimator. Moreover, they showed that the asymptotic bias of penalized M-estimators is of order $O(\epsilon)$ in ϵ contamination neighbourhoods.

In this paper, we introduce a general family of robust estimators for sparse logistic regression models, that involves both a loss and a weight function to control influential points and also a general penalty term to produce sparse estimators. In contrast to (Avella-Medina and Ronchetti 2018), our approach allows for penalties which may be random and not necessarily a deterministic sum of univariate functions. Random penalties give a more realistic scenario than deterministic ones, since the practitioner usually selects the penalty parameter using a data-driven procedure. Furthermore, they provide a general framework to include adaptive LASSO (ADALASSO). At this point, the choice of the penalty does matter. It is worth noticing that, in the objective function defining our estimators, the loss function keeps bounded the terms related to the deviance. For this reason, it seems wise to consider a bounded penalty, otherwise, the regularization term may tend to dominate in the minimization problem. In this sense, SCAD or MCP, due to (Fan and Li 2001) and (Zhang 2010), respectively, are appealing choices. We also consider as regularization the Sign penalty, that is bounded and, unlike SCAD and MCP, does not depend on an extra parameter. This penalty acts like LASSO applied to the direction of the regression vector, that is why, it does not shrink the estimated coefficients to 0 as LASSO does. In the framework of sparse representations in signal analysis, the Sign is known as the ℓ_1/ℓ_2 penalty and some of its algorithmic aspects have been discussed among others in Esser et al. (2013), Rahimi et al. (2019) and Wang et al. (2020). In opposition to our interests, these last papers focus on signal analysis, thus, the statistical properties of the related estimators are not studied. It is worth mentioning that the Sign penalty cannot be written as a sum of univariate deterministic functions, so the asymptotic properties of the penalized estimators cannot be derived from Theorem 2 in Avella-Medina and Ronchetti (2018). In this sense, our results fill the gap.

A primary focus of this paper is to provide a rigorous theoretical foundation for our approach to robust sparse logistic regression when the dimension of the covariates is fixed. It should be highlighted that a similar strategy to the one proposed herein could be followed in the high-dimensional scenario as done for robust quasi-likelihood-type estimators in Avella-Medina and Ronchetti (2018). However, when the dimension p increases with the sample size n, particular considerations and developments are required to obtain theoretical properties. This interesting topic is beyond the scope of the present paper and will be object of future research.

The rest of this paper is organized as follows. In Sect. 2, the robust penalized logistic regression estimators are introduced. In particular, Sect. 2.1 introduces a robust procedure to select the penalty parameter and discusses the importance of considering a bounded loss in the cross-validation criterion. Sections 3 and 4 summarize the asymptotic properties of the proposal. Section 5 reports the results of a Monte Carlo study. In Sect. 6, we present the analysis of a real dataset related to breast cancer diagnosis, while Sect. 7 contains some concluding remarks. Proofs are relegated to the Supplementary file where we also describe an algorithm to effectively compute the estimators and report some complementary simulation results. The analysis of dataset related to tomography images is also presented in the online supplement.

2 Robust penalized estimators

Throughout this paper, we consider a logistic regression model, that is, we have a sample of i.i.d. observations $(y_i, \mathbf{x}_i)$, $1 \le i \le n$ such that $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{0, 1\}$ is a binary variable such that $y_i|\mathbf{x}_i \sim Bi(1, F(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0))$, where $Bi(1, p)$ stands for the Bernoulli distribution with success probability p, $F(t) = \exp(t)\left[1 + \exp(t)\right]^{-1}$ is the logistic function and $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the true logistic regression vector.

In the non-sparse setting, M-estimators were defined in Bianco and Yohai (1996) and Basu et al. (2017), while in order to obtain bounded influence estimators a weighted version was introduced in Croux and Haesbroeck (2003). For the sake of completeness, we briefly recall their definition. Let $\rho : \mathbb{R}_{\ge 0} \to \mathbb{R}$ be a bounded, differentiable and non-decreasing function with derivative $\psi = \rho'$ and define

$$L_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{x}_i)\,, \tag{1}$$

with

$$\phi(y, t) = \rho(d(y, t)) + G(F(t)) + G(1 - F(t))\,, \tag{2}$$

where $d(y, t) = -\log(F(t))y - \log(1 - F(t))(1 - y)$ is the deviance function and $G(t) = \int_0^t \psi(-\log u)\,du$ is the correction factor needed to guarantee Fisher-consistency. The weights $w(\mathbf{x}_i)$ are usually based on a robust Mahalanobis distance of the explanatory variables, that is, they depend on the distance between $\mathbf{x}_i^{\star}$ and a robust centre of the data, where $\mathbf{x} = (1, \mathbf{x}^{\star\mathrm{T}})^{\mathrm{T}}$ when an intercept is included in the model and $\mathbf{x} = \mathbf{x}^{\star}$ when no intercept is considered. The weighted M-estimators are then defined as

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\, L_n(\boldsymbol{\beta})\,. \tag{3}$$

As for the maximum likelihood estimators, the weighted M-estimators do not lead to sparse estimators. This entails that they do not allow to make variable selection and may have a bad performance regarding robustness and efficiency. In this setting, a usual way to improve the behaviour of existing estimators is to include a regularization term that penalizes candidates without few nonzero components. The penalized estimators are defined as

$$\widehat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \phi(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\, w(\mathbf{x}_i) + I_{\lambda_n}(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\, L_n(\boldsymbol{\beta}) + I_{\lambda_n}(\boldsymbol{\beta}), \quad (4)$$

where $L_n(\boldsymbol{\beta})$ is given in (1), ϕ is defined in (2) and $I_{\lambda_n}(\boldsymbol{\beta})$ is a penalty function, chosen by the user, depending on a tuning parameter λ_n which measures the estimated logistic regression model complexity. The intercept is usually not penalized, when the model contains one. For that reason and for the sake of simplicity, when deriving the asymptotic properties of the estimators, we will assume that the model has no intercept. If the penalty function is properly chosen, the penalized M-estimator defined in (4) will lead to sparse models.

It is worth noticing that the estimators introduced in (4) represent a wide family which includes the M-estimators defined in Bianco and Yohai (1996), by taking $w(\mathbf{x}) = 1$ and $I_{\lambda_n}(\boldsymbol{\beta}) = 0$. In particular, the penalized maximum likelihood estimators correspond to $\rho(t) = t$ which is not bounded and a penalized version of the minimum divergence estimators defined in Basu et al. (2017) taking $\rho(t) = \rho_{\mathrm{DIV}}(t) = (1 + 1/c)\{1 - \exp(-ct)\}$. From now on, we denote $\|\boldsymbol{\beta}\|_q^q = \sum_{j=1}^{p} \beta_j$, for $q > 0$. The estimators defined in Chi and Scott (2014) belong to the family (4) just by choosing $\rho(t) = 1 - \exp(-t)$ and $I_\lambda(\boldsymbol{\beta}) = \lambda\left(\theta\|\boldsymbol{\beta}\|_1 + [(1-\theta)/2]\|\boldsymbol{\beta}\|_2^2\right)$, with $\theta \in [0, 1]$, i.e. the Elastic Net penalty. Note that Elastic Net reduces to the LASSO penalty for $\theta = 1$ and to the Ridge penalty for $\theta = 0$. The main drawbacks of this penalization is that it introduces an extra parameter that must be chosen additionally to the penalty factor λ and that it produces estimators of the non-null components with a large bias.

Some other penalties considered in the linear regression model are the Bridge penalty introduced in Frank and Friedman (1993) and defined as $I_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_q^q$. For linear models the Bridge penalty leads to sparse estimations when $0 < q < 1$. Zou (2006) has shown that LASSO may not be an oracle procedure for linear regression models and introduced the adaptive LASSO from an initial consistent estimator $\widetilde{\boldsymbol{\beta}}$. The penalty function for the ADALASSO estimator is chosen as $I_\lambda(\boldsymbol{\beta}) = \lambda I^\star(\boldsymbol{\beta})$, where $I^\star(\boldsymbol{\beta})$ is a random function defined as

$$I^\star(\boldsymbol{\beta}) = \sum_{j=1}^{p} \frac{|\beta_j|}{|\widetilde{\beta}_j|^\gamma}\,, \quad (5)$$

for some $\gamma > 0$, where we understand that $|\beta_j|/|\widetilde{\beta}_j|^\gamma = \infty$ if $|\widetilde{\beta}_j| = 0$ but $|\beta_j| \neq 0$, while $|\beta_j|/|\widetilde{\beta}_j|^\gamma = 0$ if $|\widetilde{\beta}_j| = |\beta_j| = 0$. If we seek for a robust penalized procedure using ADALASSO and to preserve robustness of the final estimator, $\widetilde{\boldsymbol{\beta}}$ can be chosen as the non-penalized robust estimator, that is, the minimizer of $L_n(\boldsymbol{\beta})$.

A distinguishing feature in logistic regression is that the response variable is bounded. This implies that when considering the penalized least squares estimators the first term in (4) is always bounded and hence, the penalty term may dominate the behaviour of the objective function, unless the regularization function is also bounded.

This is the reason why, we will also consider bounded penalties such as the SCAD penalty defined in Fan and Li (2001) as

$$I_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} \lambda|\beta_j|\, \mathbf{1}_{\{|\beta_j|\le\lambda\}} + \sum_{j=1}^{p} \frac{a\lambda|\beta_j| - 0.5(\beta_j^2+\lambda^2)}{a-1}\, \mathbf{1}_{\{\lambda<|\beta_j|\le a\lambda\}}$$
$$+ \sum_{j=1}^{p} \frac{\lambda^2(a^2-1)}{2(a-1)}\, \mathbf{1}_{\{|\beta_j|>a\lambda\}}\,,$$

for $a > 2$, where $\mathbf{1}_A$ is the indicator function of the set A, and the MCP penalty proposed by Zhang (2010) in the linear regression model which is given by

$$I_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left(\lambda|\beta_j| - \frac{\beta_j^2}{2\,a}\right) \mathbf{1}_{\{|\beta_j|\le a\,\lambda\}} + \frac{1}{2}\,a\,\lambda^2\, \mathbf{1}_{\{|\beta_j|>a\,\lambda\}}.$$

Furthermore, a main objective under a sparse setting is variable selection, that is, to identify variables related to non-null coefficients. Hence, it is more relevant to determine the coefficients β_j that are non-null than their size. For that purpose, we also consider a penalty that shrinks the coefficients by pulling the vector $\boldsymbol{\beta}$ to the unit Euclidean ball before applying a LASSO penalty. This results in the so-called Sign penalty, also known as the ℓ_1/ℓ_2 penalization in signal analysis, which is defined as

$$I_\lambda(\boldsymbol{\beta}) = \lambda \frac{\|\boldsymbol{\beta}\|_1}{\|\boldsymbol{\beta}\|_2} \mathbf{1}_{\boldsymbol{\beta}\neq\mathbf{0}} = \lambda\|s(\boldsymbol{\beta})\|_1 \mathbf{1}_{\boldsymbol{\beta}\neq\mathbf{0}}\,,$$

where $s(\boldsymbol{\beta}) = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2$ is the sign function. In multivariate analysis, the sign function has been extensively considered to construct robust estimators. Up to our knowledge, this paper is the first one in deriving the asymptotic properties of penalized estimators based on $s(\boldsymbol{\beta})$. Note that the Sign penalty works like LASSO over all unit vectors and in this sense, it enables the selection of a direction, more than raw variable selection. The Sign penalty produces a thresholding rule, that is, it estimates some coefficients as nonzero. It reaches the minimum when only one of its components is not zero and its maximum when all its components are equal and different from zero. Two important features of this penalty are that it is scale invariant, so it does not shrink the estimated coefficients as the Elastic Net penalty does, and it does not require to select an extra parameter as SCAD and MCP.

2.1 Selection of the penalty parameter

As it is well known, the selection of the penalty parameter is an important practical issue when fitting sparse models, since in some sense it tunes the complexity of the model. This problem has been discussed, among others, in Efron et al. (2004), Meinshausen (2007) and Chi and Scott (2014). In this paper, a robust K-fold criterion is used to select the penalty parameter.

As usual, first randomly split the dataset into K disjoint subsets of approximately equal sizes, with indices $\mathcal{C}_j$, $1 \le j \le K$, the j-th subset having size $n_j \ge 2$, so that $\bigcup_{j=1}^{K} \mathcal{C}_j = \{1, \ldots, n\}$ and $\sum_{j=1}^{K} n_j = n$. Let $\widetilde{\Lambda} \subset \mathbb{R}$ be the set of possible values for λ to be considered, and let $\widehat{\boldsymbol{\beta}}_\lambda^{(j)}$ be an estimator of $\boldsymbol{\beta}_0$, computed with penalty parameter $\lambda \in \widetilde{\Lambda}$ and without using the observations with indices in $\mathcal{C}_j$. For each $i = 1, \dots, n$, the prediction residuals $\widehat{d}_{i,\lambda}$ are $\widehat{d}_{i,\lambda} = d(y_i, \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\lambda^{(j)})$, for $i \in \mathcal{C}_j$ and $j = 1, \dots, K$. The classical cross-validation criterion constructs adaptive data-driven estimators by minimizing

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \widehat{d}_{i,\lambda} \,, \tag{6}$$

an objective function that is usually employed for the classical estimators which minimize the deviance. However, this criterion is very sensitive to the presence of outliers. In fact, even when $\boldsymbol{\beta}_0$ is estimated by means of a robust method, the traditional cross-validation criterion may lead to poor variable selection results since atypical data may have large prediction residuals that could be very influential on $CV(\lambda)$. To overcome this problem, when using robust estimators, it seems natural to use the same loss function ϕ as in (4). Hence, the robust cross-validation criterion selects the penalty parameter by minimizing over $\widetilde{\Lambda}$

$$RCV(\lambda) = \frac{1}{n} \sum_{1 \le j \le K} \sum_{i \in \mathcal{C}_j} \phi(y_i, \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\lambda^{(j)}) \, w(\mathbf{x}_i) \,. \tag{7}$$

The particular case $K = n$ leads to leave-one-out cross-validation which is a popular choice with a more expensive computational cost. In Section S.8.1 of the supplementary material, we illustrate through a numerical example, the importance of considering a bounded loss in the cross-validation criterion when performing the selection of the penalty parameter in order to achieve reliable prediction.

3 Consistency and order of convergence

In this section, we study the asymptotic behaviour of the estimators defined in (4) when p is fixed. Even though we are mainly concerned with bounded penalties, our results are general and include among others the Bridge and Elastic Net penalties.

3.1 Assumptions

When considering the function ϕ given in (2), the following set of assumptions on the loss function ρ are needed.

R1 $\rho : \mathbb{R}_{\ge 0} \to \mathbb{R}$ is a bounded, continuously differentiable function with bounded derivative ψ and $\rho(0) = 0$.

R2 $\psi(t) \ge 0$ and there exists some $c \ge \log 2$ such that $\psi(t) > 0$ for all $0 < t < c$.

R3 ρ is twice continuously differentiable with bounded derivatives, i.e. ψ and $\psi' = \rho''$ are bounded.

Remark 1 Note that for the function $\phi(y,t)$ defined in (2), $\Psi(y,t) = \partial\phi(y,t)/\partial t = -[y - F(t)]\nu(t)$ with $\nu(t)$ given by

$$\nu(t) = \psi\left(-\log F(t)\right)[1 - F(t)] + \psi\left(-\log[1 - F(t)]\right) F(t)\,. \tag{8}$$

Further, under **R1** and **R2**, the function $\Psi(y,\cdot)$ is continuous and strictly positive.

Denote as $\chi(y,t) = \partial\Psi(y,t)/\partial t = F(t)(1 - F(t))\nu(t) - (y - F(t))\nu'(t)$ and note that $\chi(0,s) = \chi(1,-s)$. The function $\chi(y,t)$ always exists for the minimum divergence estimators and is well defined for any function ρ satisfying **R3**.

It is worth noticing that when $\psi(t) > 0$ the constant c in **R2** may be taken as ∞. For instance, this happens when choosing the loss function $\rho = \rho_{\text{DIV}}$ related to the divergence estimators or the function $\rho = \rho_c$, with $c > 0$, defined as

$$\rho_c(t) = \begin{cases} te^{-\sqrt{c}} & \text{if } \ t \le c \\ -2e^{-\sqrt{t}}\left(1 + \sqrt{t}\right) + e^{-\sqrt{c}}\left(2\left(1 + \sqrt{c}\right) + c\right) & \text{if } \ t > c\,, \end{cases} \tag{9}$$

which has been introduced in Croux and Haesbroeck (2003) to ensure the existence of the M-estimators under the same conditions that guarantee existence for the maximum likelihood estimators. Moreover, when considering the penalized minimum divergence estimators, ρ automatically satisfies conditions **R1**, **R2** and **R3**.

For the results in this section, the following assumptions regarding the distribution of $\mathbf{x}$ are needed.

H1 For all $\boldsymbol{\alpha} \in \mathbb{R}^p$, $\boldsymbol{\alpha} \neq \mathbf{0}$, we have $\mathbb{P}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha} = 0) = 0$.

H2 w is a non-negative bounded function with support $\mathcal{C}_w$ such that $\mathbb{P}(\mathbf{x} \in \mathcal{C}_w) > 0$. Without loss of generality, we assume that $\|w\|_\infty = 1$.

H3 $\mathbb{E}[w(\mathbf{x})\|\mathbf{x}\|^2] < \infty$.

H4 The matrix $\mathbf{A} = \mathbb{E}\left(F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\left[1 - F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]\nu(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\, w(\mathbf{x})\, \mathbf{x}\mathbf{x}^{\mathrm{T}}\right)$, where $\nu(t)$ is defined in (8), is non-singular.

Remark 2 Assumptions **H1** and **H2** entail that the estimators defined in (3) are Fisher-consistent and will allow to derive consistency results for the estimators defined in (4). **H1** holds for instance, when $\mathbf{x}$ has a density with support $\mathcal{S}$ such that $\mathcal{S} \cap \mathcal{C}_w \neq \emptyset$. In fact, the weaker assumption $\mathbb{P}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha} = 0 \cup w(\mathbf{x}) = 0) < 1$ for any $\boldsymbol{\alpha} \neq \mathbf{0}$ is enough for obtaining Fisher-consistency. However, in order to ensure consistency a stronger requirement is needed to guarantee that the infimum is not attained at infinity. It is worth noticing that **H1** and **H2** entail that $\mathbb{E}[w(\mathbf{x})\,\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ is a positive definite matrix. Furthermore, when considering the minimum divergence estimators the matrix $\mathbf{A}$ is non-singular, since $\mathbb{P}(\nu(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) > 0) = 1$, so **H4** holds. Similarly, when $\mathbb{P}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha} = 0) < 1$ for any $\boldsymbol{\alpha} \neq \mathbf{0}$, and ϕ is given by (2) with $\psi(t) > 0$ for all t, as is the case with the loss function introduced in Croux and Haesbroeck (2003), $\mathbf{A}$ is non-singular. On the other hand, when **R2** holds for some finite positive constant $c \geq \log 2$, $\mathbf{A}$ is positive definite when **H1** holds. Moreover, define $\Upsilon(t) = F(t)(1 - F(t))\nu(t)$, straightforward arguments allow to see that $\mathbf{A}$ is also non-singular when $\mathbb{P}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha} = 0) < 1$ holds, for any $\boldsymbol{\alpha} \neq \mathbf{0}$, and at least one of the following conditions is fulfilled:

a) the function $\mathbb{E}[w(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{1}_{\Upsilon(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\geq\eta}]$ is continuous in η or b) there exists some $c > 0$ such that $\mathbb{P}(\Upsilon(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) > c) = 1$.

Remark 3 It is worth mentioning that assumption **H3** is weaker than Condition 3 in Avella-Medina and Ronchetti (2018), while condition **H4** is equivalent to the non-singularity requirement in Condition 2 therein. Regarding Condition 1 of Avella-Medina and Ronchetti (2018), the Fisher-consistency is automatically fulfilled due to the correction factor $G(\cdot)$. Furthermore, instead of the uniformity condition asked by those authors, we only require to the function ψ continuity and boundedness. Note that when $w \equiv 1$ and the covariates are bounded or when considering hard rejection weights, their Condition 1 is satisfied.

3.2 Consistency and rate of convergence

The next theorem states the strong consistency of the estimators defined in (4), when considering as function ϕ the function controlling large values of the *deviance* residuals given in (2).

Theorem 1 *Let $\phi : \mathbb{R}^2 \to \mathbb{R}$ be the function given in* (2)*, where the function ρ satisfies **R1** and **R2**. Then, if $I_{\lambda_n}(\boldsymbol{\beta}_0) \xrightarrow{a.s.} 0$ when $n \to \infty$ and **H1** and **H2** hold, we have that the estimator $\widehat{\boldsymbol{\beta}}_n$ defined in* (4) *is strongly consistent for $\boldsymbol{\beta}_0$.*

It is worth noticing that, in Theorem 1, the penalty function I_{λ_n} may be deterministic or random, since the only requirement is that $I_{\lambda_n}(\boldsymbol{\beta}_0) \xrightarrow{a.s.} 0$. In particular, for the penalties LASSO, Sign, Ridge, Bridge, SCAD and MCP described in Sect. 2 this condition holds when $\lambda_n \xrightarrow{a.s.} 0$. Moreover, for the ADALASSO penalty, the condition $I_{\lambda_n}(\boldsymbol{\beta}_0) \xrightarrow{a.s.} 0$ is fulfilled when the initial estimator $\widetilde{\boldsymbol{\beta}}$ is consistent and $\lambda_n \xrightarrow{a.s.} 0$.

In order to prove the $\sqrt{n}$-consistency of the proposed estimators, we need the following assumption on the penalty function. From now on, $\mathcal{B}(\boldsymbol{\beta}, \epsilon)$ stands for the closed ball, with respect to the usual $\|\cdot\|_2$ norm, centred at $\boldsymbol{\beta}$ with radius ϵ, i.e. $\mathcal{B}(\boldsymbol{\beta}, \epsilon) = \{\mathbf{b} \in \mathbb{R}^p : \|\mathbf{b} - \boldsymbol{\beta}\|_2 \leq \epsilon\}$.

P1 $I_\lambda(\boldsymbol{\beta})/\lambda$ is Lipschitz in a neighbourhood of $\boldsymbol{\beta}_0$, that is, there exists $\epsilon > 0$ a constant K, which does not depend on λ, such that if $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}(\boldsymbol{\beta}_0, \epsilon)$ then $|I_\lambda(\boldsymbol{\beta}_1) - I_\lambda(\boldsymbol{\beta}_2)| \leq \lambda K \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_1$.

Remark 4 Note that penalties Ridge, Elastic Net, SCAD and MCP satisfy **P1**, since $\|\boldsymbol{\beta}\|_2 \leq \|\boldsymbol{\beta}\|_1 \leq \sqrt{p}\,\|\boldsymbol{\beta}\|_2$. Furthermore, the Sign penalty also satisfies **P1** if $\|\boldsymbol{\beta}_0\|_2 \neq 0$. Moreover, if $I_\lambda(\boldsymbol{\beta}) = \lambda \sum_{\ell=1}^{p} J_\ell(|\beta_\ell|)$, where $J_\ell(\cdot)$ is a continuously differentiable function, then I_λ satisfies **P1**, which implies that the Bridge penalty satisfies **P1** for $q \geq 1$.

Theorem 2 *Let $\widehat{\boldsymbol{\beta}}_n$ be the estimator defined in* (4) *with $\phi(y, t)$ given in* (2)*, where the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfies **R3**. Furthermore, assume that $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$ and that assumptions **H2** to **H4** hold.*

(a) *If assumption* **P1** *holds,* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\lambda_n + 1/\sqrt{n})$*. Hence, if* $\lambda_n = O_{\mathbb{P}}(1/\sqrt{n})$*, we have that* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1/\sqrt{n})$*, while if* $\lambda_n\sqrt{n} \to \infty$*,* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\lambda_n)$*.*

(b) *Suppose* $I_{\lambda_n}(\boldsymbol{\beta}) = \sum_{\ell=1}^{p} J_{\ell,\lambda_n}(|\beta_\ell|)$ *where the functions* $J_{\ell,\lambda_n}(\cdot)$ *are twice continuously differentiable in* $(0,\infty)$*, take non-negative values,* $J'_{\ell,\lambda_n}(|\beta_{0,\ell}|) \geq 0$ *and* $J_{\ell,\lambda_n}(0) = 0$*, for all* $1 \leq \ell \leq p$*. Let*

$$a_n = \max\left\{J'_{\ell,\lambda_n}(|\beta_{0,\ell}|) : 1 \leq \ell \leq p \;\; and \;\; \beta_{0,\ell} \neq 0\right\} \quad and \quad \alpha_n = \frac{1}{\sqrt{n}} + a_n.$$

In addition, assume that there exists some $\delta > 0$ *such that*

$$\sup\{|J''_{\ell,\lambda_n}(|\beta_{0,\ell}| + \tau\delta)| : \tau \in [-1,1]\,,\; 1 \leq \ell \leq p \;\; and \;\; \beta_{0,\ell} \neq 0\} \xrightarrow{p} 0.$$

Then, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\alpha_n)$*.*

Remark 5 Theorem 2(a) shows that, when the penalty satisfies assumption **P1**, the estimator rate of convergence depends on the convergence rate of λ_n to 0. In particular, if $\lambda_n\sqrt{n}$ is bounded in probability, then the robust penalized consistent estimator has rate $\sqrt{n}$, while if $\lambda_n\sqrt{n} \to \infty$, the convergence rate of $\widehat{\boldsymbol{\beta}}_n$ is slower than $\sqrt{n}$. This result is analogous to the one obtained, under a linear regression model, in Zou (2006) for the penalized least squares estimator when a LASSO penalty is considered. Note that, for the LASSO penalty, the convergence rates obtained in (a) and (b) are equal since $J_{\ell,\lambda_n}(v) = \lambda_n\, v$, for any $1 \leq \ell \leq p$, which entails that $a_n = \lambda_n$ and for any $\beta_{0,\ell} \neq 0$, $\tau \in [-1,1]$, $J''_{\ell,\lambda_n}(|\beta_{0,\ell}| + \tau\delta) = 0$ for a small enough $\delta > 0$.

Penalties SCAD and MCP are not only Lipschitz, but also based on univariate twice continuously differentiable functions $J_{\ell,\lambda_n}(t) = J_{\lambda_n}(t)$, for all $1 \leq \ell \leq p$, satisfying the requirements asked in Theorem 2(b) when $\lambda_n \to 0$. Indeed, for these penalties $J'_{\lambda_n}(t)$ and $J''_{\lambda_n}(t)$ are 0 if $t > a\,\lambda_n$ where a is their second tuning constant which is assumed to be fixed. Hence, if $\lambda_n \xrightarrow{p} 0$ for any $\delta > 0$ there exists n_0 such that, for any $n \geq n_0$, we have that $\mathbb{P}(a\lambda_n < m_0) > 1 - \delta$ with $m_0 = \min\{|\beta_{0,\ell}|) : 1 \leq \ell \leq p \;$ and $\; \beta_{0,\ell} \neq 0\}$. Thus, for $n \geq n_0$, $\mathbb{P}(a_n = 0 \text{ and } b_n = 0) > 1 - \delta$ and therefore, $\alpha_n = O_{\mathbb{P}}(1/\sqrt{n})$, implying that the root-n rate may be achieved only assuming only that $\lambda_n \xrightarrow{p} 0$. It is worth noticing that, even when, the Ridge penalty is Lipschitz and it is also based on univariate twice continuously differentiable functions, $J'_{\lambda_n}(|\beta_{0,\ell}|) = \lambda_n|\beta_{0,\ell}|$, so that $a_n = O(1/\sqrt{n} + \lambda_n)$, leading to root-$n$ consistency rate with the additional requirement $\lambda_n = O_{\mathbb{P}}(1/\sqrt{n})$. The different behaviour of the estimators related to Lipschitz penalties or penalties related to twice continuously differentiable functions with null first derivative for n large enough plays an important role regarding the variable selection properties of the procedure.

Furthermore, when considering the ADALASSO estimators, root-n estimators are obtained when the initial estimator $\widetilde{\boldsymbol{\beta}}$ is consistent and $\sqrt{n}\lambda_n = O_{\mathbb{P}}(1)$, since in this case $a_n = \lambda_n \max_{j\in\mathcal{A}} |\widetilde{\beta}_j|^{-\gamma}$, with $\mathcal{A} = \{j : \beta_{0,j} \neq 0\}$. In particular, for deterministic bandwidths, this result holds if $\sqrt{n}\lambda_n \to 0$ in concordance with Theorem 2 from (Zou 2006).

4 Asymptotic distribution results

The first result in this section concerns the variable selection properties for our estimator. As shown below, the result depends on the behaviour of the penalty function. Without loss of generality, assume that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,A}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\beta}_{0,A} \in \mathbb{R}^k$, $k \geq 1$, is the subvector with **active** coordinates of $\boldsymbol{\beta}_0$ (i.e. the subvector of nonzero elements of $\boldsymbol{\beta}_0$). We will make use of the notation $\boldsymbol{\beta} = (\boldsymbol{\beta}_A^{\mathrm{T}}, \boldsymbol{\beta}_B^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_A \in \mathbb{R}^k$ with $k \geq 1$ and $\boldsymbol{\beta}_B \in \mathbb{R}^{p-k}$.

When the estimator automatically selects variables, we will be able to show an oracle property, that is, that the penalized M-estimator of the non-null components of $\boldsymbol{\beta}_0$, $\widehat{\boldsymbol{\beta}}_{n,A}$ has the same asymptotic distribution as that of the estimator obtained assuming that the last components of $\boldsymbol{\beta}_0$ are equal to 0 and using this restriction in the logistic regression model. It is worth noticing that in the non-sparse scenario, the asymptotic behaviour of the estimators $\widehat{\boldsymbol{\beta}}$ defined in (3) has been studied in Bianco and Martínez (2009), while Basu et al. (2017) consider the particular case of the minimum divergence estimators and $w(\mathbf{x}) \equiv 1$. More precisely, the above mentioned authors have shown that $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, where

$$\mathbf{B} = \mathbb{E}\left(F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\left[1 - F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]\nu^2(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\, w^2(\mathbf{x})\, \mathbf{x}\mathbf{x}^{\mathrm{T}}\right) . \tag{10}$$

with $\nu(t)$ defined in (8) and the matrix $\mathbf{A}$ is given in assumption **H4**.

For the sake of simplicity, throughout this section, we will assume that the parameter λ_n is deterministic. Similar results may be obtained when the penalty parameter is random. However, we also admit $I_\lambda(\boldsymbol{\beta})$ to be random, so in Sect. 4.1, we will treat separately the case in which $I_\lambda(\boldsymbol{\beta})$ is a deterministic or random function, leading to Theorems 3 and 4, respectively.

4.1 Variable selection property

Theorem 3 *Let $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n,A}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}_{n,B}^{\mathrm{T}})^{\mathrm{T}}$ be the estimator defined in (4), where $\phi(y,t)$ is given in (2) and the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfies **R3**. Furthermore, assume that **H2** and **H3** hold and that $\sqrt{n}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$. Moreover, assume that for every $C > 0$ and $\ell \in \{k+1, \ldots, p\}$, there exist a constant $K_{C,\ell}$ and $N_{C,\ell} \in \mathbb{N}$ such that if $\|\mathbf{u}\|_2 \leq C$ and $n \geq N_{C,\ell}$, then*

$$I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}^{(-\ell)}}{\sqrt{n}}\right) \geq K_{C,\ell}\,\frac{\lambda_n}{\sqrt{n}}\,|u_\ell|, \tag{11}$$

where $\mathbf{u}^{(-\ell)}$ is obtained by replacing the ℓ-th coordinate of $\mathbf{u}$ with zero and u_ℓ is the ℓ-th coordinate of $\mathbf{u}$.

(a) For every $\tau > 0$, there exists $b > 0$ and $n_0 \in \mathbb{N}$ such that if $\lambda_n = b/\sqrt{n}$, we have that, for any $n \geq n_0$, $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,B} = \mathbf{0}_{p-k}) \geq 1 - \tau$.

(b) If $\lambda_n\sqrt{n} \to \infty$, then $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,B} = \mathbf{0}_{p-k}) \to 1$.

To prove variable selection properties for our estimators, it only remains to show that condition (11) holds for the different penalties mentioned above. First note that (11) is clearly satisfied for the LASSO penalty. In the proof of Corollary 1, we show that SCAD, MCP and the Sign penalty also verify (11).

Corollary 1 *Let $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n,A}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}_{n,B}^{\mathrm{T}})^{\mathrm{T}}$ be the estimator defined in (4) with $\phi(y,t)$ given by (2) where the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfies* ***R3***. *Assume that* ***H2*** *and* ***H3*** *hold and $\sqrt{n}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$.*

(a) If $I_{\lambda_n}(\boldsymbol{\beta})$ is the Sign penalty, then for every $\tau > 0$ there exist $b > 0$ and $n_0 \in \mathbb{N}$ such that if $\lambda_n = b/\sqrt{n}$, we have that, for any $n \geq n_0$, $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,B} = \mathbf{0}_{p-k}) \geq 1 - \tau$.

(b) If $I_{\lambda_n}(\boldsymbol{\beta})$ is taken as the SCAD or MCP penalties and $\sqrt{n}\lambda_n \to \infty$, then $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,B} = \mathbf{0}_{p-k}) \to 1$.

Remark 6 It is noteworthy that when the penalty function $I_\lambda \boldsymbol{\beta})$ is deterministic and can we written as a sum of continuously differentiable univariate functions, inequality (11) is equivalent to Condition 4 in Avella-Medina and Ronchetti (2018).

A consequence of Corollary 1 is that the penalties SCAD and MCP have the property of automatically selecting variables when $\sqrt{n}\lambda_n \to \infty$. This states a difference with (Avella-Medina and Ronchetti 2018) who require stronger rates on λ_n, see Remark 9. In contrast, when using the LASSO and Sign penalties, we cannot ensure the variable selection property when the estimator is root-n consistent. Recall that, for these two penalties, Theorem 2 entails that the estimator converges at a rate slower than $\sqrt{n}$ when $\lambda_n\sqrt{n} \to \infty$. For that reason, we can only guarantee that for a given $0 < \tau < 1$, we can choose a sequence of penalty parameters $\lambda_n = b/\sqrt{n}$ (in order to ensure that the estimator has a root-n rate) and such that the penalized M-estimator selects variables with probability larger than $1 - \tau$.

The results in the asymptotic distribution given below will allow to conclude that, for the LASSO and Sign penalties, when the estimator has convergence rate $\sqrt{n}$, then $\limsup_n \mathbb{P}(\mathcal{A}_n = \mathcal{A}) < 1$, where $\mathcal{A} = \{j : \beta_{0,j} \neq 0\} = \{1, \ldots, k\}$ and $\mathcal{A}_n = \{j : \widehat{\beta}_{n,j} \neq 0\}$ are the set of indexes related to the active components of $\boldsymbol{\beta}_0$ and to the non-null coordinates of $\widehat{\boldsymbol{\beta}}_n$, respectively. This result is analogous to Proposition 1 in Zou (2006), which shows that the LASSO estimator leads to inconsistent variable selection in the linear regression model, when $\lambda_n = O(1/\sqrt{n})$.

It is worth noticing that $\widehat{\boldsymbol{\beta}}_{n,B} = \mathbf{0}_{p-k}$ if and only if $\mathcal{A}_n \subset \mathcal{A}$, hence, if $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,B} = \mathbf{0}_{p-k}) \to 1$ we have that $\mathbb{P}(\mathcal{A}_n \subset \mathcal{A}) \to 1$. Note that when $\mathcal{A}_n \subsetneq \mathcal{A}$, the penalized M-estimator may select a submodel with less predictors than the original one, shrinking the estimation of some of the active to 0; however, the oracle property of the estimators based on SCAD or MCP given in Theorem 8 will allow to conclude that $\mathbb{P}(\mathcal{A}_n = \mathcal{A}) \to 1$.

To derive the variable selection property for random penalties such as the ADALASSO constructed from a root-n consistent initial estimator, we state the following result whose proof is omitted since it follows using similar arguments to those considered in the proof of Theorem 3. As mentioned above, this property is crucial to obtain the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{n,A}$ in Sect. 5. Note that for the ADALASSO penalty, the constant $\gamma > 0$ in Theorem 4 corresponds to the value of γ involved in its definition in (5).

Theorem 4 *Let $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n,A}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}_{n,B}^{\mathrm{T}})^{\mathrm{T}}$ be the estimator defined in (4), where $\phi(y,t)$ is given in (2) and the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfies **R3**. Assume that $\sqrt{n}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$ and that for some $\gamma > 0$, $n^{(1+\gamma)/2}\lambda_n \to \infty$. Furthermore, assume that for every $C > 0$, $\ell \in \{k+1, \ldots, p\}$ and $\tau > 0$, there exist a constant $K_{C,\ell}$ and $N_{C,\ell} \in \mathbb{N}$ such that if $\|\mathbf{u}\|_2 \leq C$ and $n \geq N_{C,\ell}$, we have that*

$$\mathbb{P}\left(I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}^{(-\ell)}}{\sqrt{n}}\right) \geq K_{C,\ell}\,\frac{\lambda_n}{\sqrt{n^{1-\gamma}}}\,|u_\ell| \right) > 1 - \tau, \quad (12)$$

*where $\mathbf{u}^{(-\ell)}$ and u_ℓ are defined as in Theorem 3. Then, under **H2** and **H3**, $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,B} = \mathbf{0}_{p-k}) \to 1$.*

4.2 Asymptotic distribution

In this section, we derive separately the asymptotic distribution of our estimator depending on the choice of the penalty. As the rate of convergence to 0 of λ_n required to obtain root$-n$ estimators for the Sign is different from that of SCAD or MCP penalties, we will study these two situations separately. Even though most results on penalized estimators assume that the sequence of penalty parameters is deterministic, in this section, as in Theorem 2, we will allow random penalty parameters λ_n, having in this sense a more realistic point of view.

It is worth noticing that, under **H4**, the matrix **A** defined in assumption **H4** is positive definite, so the submatrix corresponding to the active coordinates of $\boldsymbol{\beta}_0$ is also positive definite.

From now on, $\mathbf{e}_\ell$ stands for the ℓ-th canonical vector and $\mathrm{sign}(z)$ is the univariate sign function, that is, $\mathrm{sign}(z) = z/|z|$ when $z \neq 0$ and $\mathrm{sign}(0) = 0$.

Theorem 5 *Let $\widehat{\boldsymbol{\beta}}_n$ be the estimator defined in (4) with $\phi(y,t)$ given in (2), where the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfies **R3**. Assume that **H2** to **H4** hold, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = O_{\mathbb{P}}(1)$ and $\sqrt{n}\,\lambda_n \xrightarrow{p} b$. Consider the Sign penalty given by $I_\lambda(\boldsymbol{\beta}) = \lambda\,\|\boldsymbol{\beta}\|_1/\|\boldsymbol{\beta}\|_2$. Then, if $\|\boldsymbol{\beta}_0\| \neq 0$, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} \mathrm{argmin}_{\mathbf{z}}\, R(\mathbf{z})$, where the process $R : \mathbb{R}^p \to \mathbb{R}$ is defined as $R(\mathbf{z}) = \mathbf{z}^{\mathrm{T}}\mathbf{w} + (1/2)\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z} + b\ \mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z})$, with $\mathbf{w} \sim N_p(\mathbf{0}, \mathbf{B})$, **A** and **B** are given in assumption **H4** and in equation (10), respectively, $\mathbf{q}(\mathbf{z}) = \sum_{\ell=1}^{p} \nabla_\ell(\boldsymbol{\beta}_0)\boldsymbol{I}_{\{\beta_{0,\ell}\neq 0\}} + \left(\mathrm{sign}(z_\ell)/\|\boldsymbol{\beta}_0\|_2\right)\boldsymbol{I}_{\{\beta_{0,\ell}=0\}}\,\mathbf{e}_\ell$ and $\nabla_\ell(\boldsymbol{\beta}) = -\left(|\beta_\ell|/\|\boldsymbol{\beta}\|_2^3\right)\boldsymbol{\beta} + (\mathrm{sign}(\beta_\ell)/\|\boldsymbol{\beta}\|_2)\,\mathbf{e}_\ell$.*

The following result generalizes Theorem 5 to differentiable penalties and includes, among others, the LASSO and Ridge penalties, and any convex combination of them, in particular the Elastic Net.

Theorem 6 *Let $\widehat{\boldsymbol{\beta}}_n$ be the estimator defined in (4) with $\phi(y,t)$ given by (2), where the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfies **R3** and let **A** and **B** be the matrices defined in assumption **H4** and in equation (10), respectively. Let us consider the penalty*

$$I_\lambda(\boldsymbol{\beta}) = \lambda\left\{(1-\alpha)\sum_{\ell=1}^{p} J_\ell(|\beta_\ell|) + \alpha\sum_{\ell=1}^{p}|\beta_\ell|\right\}, \quad (13)$$

where $J_\ell(\cdot)$ is a continuously differentiable function such that $J_\ell'(0) = 0$. Assume that ***H2*** *to* ***H4*** *hold,* $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = O_{\mathbb{P}}(1)$ *and that* $\sqrt{n}\,\lambda_n \xrightarrow{p} b$. *Then, if* $\|\boldsymbol{\beta}_0\| \neq 0$, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} \operatorname{argmin}_{\mathbf{z}} R(\mathbf{z})$ *where the process* $R : \mathbb{R}^p \to \mathbb{R}$ *is defined as* $R(\mathbf{z}) = \mathbf{z}^{\mathrm{T}}\mathbf{w} + (1/2)\,\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z} + b\ \ \mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z})$, *with* $\mathbf{w} \sim N_p(\mathbf{0}, \mathbf{B})$ *and* $\mathbf{q}(\mathbf{z}) = (q_1(\mathbf{z}), \dots, q_p(\mathbf{z}))^{\mathrm{T}}$ *being* $q_\ell(\mathbf{z}) = (1-\alpha)J_\ell'(|\beta_{0,\ell}|)\ \operatorname{sign}(\beta_{0,\ell}) + \alpha\left\{\operatorname{sign}(\beta_{0,\ell})\boldsymbol{I}_{\{\beta_{0,\ell}\neq 0\}} + \operatorname{sign}(z_\ell)\boldsymbol{I}_{\{\beta_{0,\ell}=0\}}\right\}$.

Remark 7 Note that when $\sqrt{n}\lambda_n \xrightarrow{p} 0\ (b=0)$, the penalized estimators based on the Sign penalty or on a penalty of the form (13) have the same asymptotic distribution as the M-estimators defined through (3). If $b > 0$ and $\alpha > 0$ in (13), analogous arguments to those considered in linear regression by Knight and Fu (2000), allow to show that the asymptotic distribution of the coordinates of $\widehat{\boldsymbol{\beta}}_n$ corresponding to null coefficients of $\boldsymbol{\beta}_0$, that is, the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{n,B}$ puts positive probability at zero. On the other hand, if $\alpha = 0$ and $b > 0$, the amount of shrinkage of the estimated regression coefficients increases with the magnitude of the true regression coefficients. Hence, for "large" parameters, the bias introduced by the differentiable penalty $J_\ell(\cdot)$ may be large.

It is worth noticing that Theorem 6 implies that, when $I_\lambda(\boldsymbol{\beta}) = \lambda \sum_{\ell=1}^{p} J_\ell(|\beta_\ell|)$ and $\sqrt{n}\lambda_n \xrightarrow{p} b$, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} \mathbf{A}^{-1}(\mathbf{w} + b\mathbf{a})$, where $\mathbf{a} = (a_1, \dots, a_p)^{\mathrm{T}}$ is such that $a_\ell = J_\ell'(|\beta_{0,\ell}|)\ \operatorname{sign}(\beta_{0,\ell})$, which shows the existing asymptotic bias introduced in the limiting distribution, unless $b = 0$. In particular, the robust Ridge M-estimator, that provides a robust alternative under collinearity, is asymptotically distributed as $N_p(2\,b\,\mathbf{A}^{-1}\boldsymbol{\beta}_0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})$.

When considering the Sign and LASSO penalties, analogous arguments to those considered in the proof of Proposition 1 in Zou (2006), together with Theorems 5 and 6 allow to see that, if the penalized M-estimator has a root$-n$ rate of convergence, then it is inconsistent for variable selection (see Corollary 2). Furthermore, from the proof we may conclude that if $\sqrt{n}\lambda_n \xrightarrow{p} 0$, then $\mathbb{P}(\mathcal{A}_n = \mathcal{A}) \to 0$, that is, we need regularization parameters that converge to 0, but not too fast in order to select variables with non-null probability.

Corollary 2 *Let* $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n,A}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}_{n,B}^{\mathrm{T}})^{\mathrm{T}}$ *be the estimator defined in (4), where $\phi(y,t)$ is given through (2) with the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfying* ***R3****. Assume that* $\|\boldsymbol{\beta}_0\| \neq 0$, $\sqrt{n}\lambda_n \xrightarrow{p} b$, $\sqrt{n}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$ *and that* ***H2*** *to* ***H4*** *hold. Then, for the Sign or LASSO penalties, there exists $c < 1$ such that* $\limsup_n \mathbb{P}(\mathcal{A}_n = \mathcal{A}) \leq c < 1$, *where* $\mathcal{A} = \{j : \beta_{0,j} \neq 0\}$ *is the set of indexes corresponding to the active coordinates of* $\boldsymbol{\beta}_0$ *and* $\mathcal{A}_n = \{j : \widehat{\beta}_{n,j} \neq 0\}$.

Similar arguments to those used in the proof of Theorem 5, allow to obtain the asymptotic distribution of the penalized M-estimator with Sign penalty, when $\sqrt{n}\lambda_n \to \infty$. A similar result holds for penalizations satisfying (13), as the LASSO one.

Theorem 7 *Let* $\widehat{\boldsymbol{\beta}}_n$ *be the estimator defined in (4), where $\phi(y,t)$ is given through (2) with the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfying* ***R3****. Assume that* $\|\boldsymbol{\beta}_0\| \neq 0$,

*$\sqrt{n}\lambda_n \xrightarrow{p} \infty$, $\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 = O_{\mathbb{P}}(\lambda_n)$ and that **H2** to **H4** hold. Let $\mathbf{A}$ be the matrix defined in assumption **H4** and consider the Sign penalty $I_\lambda(\boldsymbol{\beta}) = \lambda\, \|\boldsymbol{\beta}\|_1/\|\boldsymbol{\beta}\|_2$. Then, $(1/\lambda_n)\ (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{p} \operatorname{argmin}_{\mathbf{z}} R(\mathbf{z})$, where the function $R : \mathbb{R}^p \to \mathbb{R}$ is defined through $R(\mathbf{z}) = (1/2)\, \mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z} + \mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z})$, with $\mathbf{q}(\mathbf{z})$ the function defined in Theorem 5.*

Remark 8 Under a linear regression model, Lemma 3 in Zou (2006) provides a result analogous to Theorem 7 for the LASSO least squares estimator. As in the referred result, the rate of convergence of $\widehat{\boldsymbol{\beta}}_n$ is slower than $\sqrt{n}$ and the limit is a non-random quantity. As noted in Zou (2006), the optimal rate for $\widehat{\boldsymbol{\beta}}_n$ is obtained when $\lambda_n = O_{\mathbb{P}}(1/\sqrt{n})$, but at expenses of not selecting variables.

Finally, the following theorem gives the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{n,A}$ when the penalty is consistent for variable selection, that is, when $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,B} = \mathbf{0}_{p-k}) \to 1$. For that purpose, recall that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,A}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}$ where $\boldsymbol{\beta}_{0,A} \in \mathbb{R}^k$, $k \geq 1$, is the vector of active coordinates of $\boldsymbol{\beta}_0$ and for $\mathbf{b} \in \mathbb{R}^k$, define

$$\nabla I_\lambda(\mathbf{b}) = \frac{\partial I_\lambda\left((\mathbf{b}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}\right)}{\partial \mathbf{b}}\,.$$

Theorem 8 *Let $\widehat{\boldsymbol{\beta}}_n$ be the estimator defined in (4) with $\phi(y,t)$ given in (2), where the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfies **R3** and assume that **H2** and **H3** hold. Suppose that there exists some $\delta > 0$ such that*

$$\sup_{\|\boldsymbol{\beta}_A - \boldsymbol{\beta}_{0,A}\|_2 \leq \delta} \|\nabla I_{\lambda_n}(\boldsymbol{\beta}_A)\|_2 = o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right), \tag{14}$$

*$\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,B} = \mathbf{0}_{p-k}) \to 1$ and $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$. Let $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ be the $k \times k$ submatrices of $\mathbf{A}$ and $\mathbf{B}$, respectively, corresponding to the first k coordinates of $\boldsymbol{\beta}_0$, where $\mathbf{A}$ and $\mathbf{B}$ were defined in assumption **H4** and in equation (10), respectively. Then, if $\widetilde{\mathbf{A}}$ is invertible, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{n,A} - \boldsymbol{\beta}_{0,A}) \xrightarrow{D} N_k(\mathbf{0}, \widetilde{\mathbf{A}}^{-1}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^{-1})$.*

Remark 9 Penalties SCAD and MCP fulfil (14) when $\lambda_n \to 0$. Effectively, recall that any of them may be written as $I_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} J_\lambda(|\beta_j|)$, where $J_\lambda(t)$ is constant in $[a\lambda, \infty)$, with $a > 0$ the second tuning constant of these penalties. Using that $J_\lambda(0) = 0$, we obtain that, for any $\mathbf{b} \in \mathbb{R}^k$, $I_\lambda\left((\mathbf{b}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}\right) = \sum_{j=1}^{k} J_\lambda(|b_j|)$ and $\nabla I_\lambda(\mathbf{b}) = \sum_{j=1}^{k} J_\lambda'(|b_j|)$. Since $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}\left(1/\sqrt{n}\right)$, given $\delta > 0$ there exists $C_1 > 0$ such that $\mathbb{P}(\mathcal{D}_n) > 1 - \delta$ for $n \geq n_0$, with $\mathcal{D}_n = \{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq C_1/\sqrt{n}\}$.

Let n_1 be such that $C_1/\sqrt{n} \leq m_0/2$. Then, for any $\omega \in \mathcal{D}_n$, $n \geq n_1$ and $1 \leq j \leq k$, we have that $|\widehat{\beta}_j| \geq |\beta_{0,j}| - |\widehat{\beta}_j - \beta_{0,j}| \geq m_0 - C_1 n^{-1/2} \geq m_0/2$. Using that $\lambda_n \to 0$ we get that for $n \geq \max\{n_0, n_1\}$, we have that $j = 1, \ldots, k$, $|\widehat{\beta}_j| > a\lambda_n$, implying that $\mathcal{D}_n \subset \{\|\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_A)\|_2 = 0\}$ as desired. Hence, using Corollary 1, we get that the penalized M-estimators defined through (4) have the oracle property when using SCAD or MCP and $\lambda_n \to 0$ with $\sqrt{n}\,\lambda_n \to \infty$ which are the same convergence rates required in Fan and Li (2001).

In contrast, when considering the ADALASSO regularization, the penalized M-estimators have the oracle property when $\sqrt{n}\,\lambda_n \to 0$ and $n^{(1+\gamma)/2}\,\lambda_n \to \infty$, which coincide with the penalty parameter rates required in Zou (2006).

Summarizing, in our results, the rates of convergence of the penalty parameter are in concordance with those required in Zou (2006) or Fan and Li (2001), when considering ADALASSO or MCP, respectively. In particular, for SCAD and MCP penalties we only require $\lambda_n \to 0$ to obtain rates of convergence and $\sqrt{n}\lambda_n \to \infty$ to derive variable selection results and asymptotic distribution (see Corollary 1 and Theorem 8), while (Avella-Medina and Ronchetti 2018) need that the penalty parameter goes faster to 0 ($\sqrt{n}\lambda_n \to 0$ and $n\ \lambda_n \to \infty$) mainly due to the fact that they obtain results in shrinking neighbourhoods of the true model.

5 Monte Carlo study

In this section, we present the results of a Monte Carlo study designed to compare the small sample performance of classical and robust penalized estimators. Section S.6 of the supplementary file describes the algorithm used to compute the estimators. Complementary results of the numerical experiment presented here are given in Section S.8 of the supplementary file.

To compare the different proposals, throughout our numerical study, we considered a training sample $\mathcal{M}$ of i.i.d. observations $(y_i, \mathbf{x}_i)$, $1 \le i \le n$, $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i|\mathbf{x}_i \sim Bi(1, F(\gamma_0 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0))$, where the intercept $\gamma_0 = 0$ and we vary the values of n, p and $\boldsymbol{\beta}_0$. For clean samples, the covariates distribution is $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where two choices for $\boldsymbol{\Sigma}$ are taken. For brevity purposes, we report here the situation where $\boldsymbol{\Sigma} = \mathbf{I}_p$, while the case of correlated covariates is described in Sect. S.8.3. This last case is of particular interest since correlation among predictors may impact the variable selection performance of a given penalized estimator, see for instance (Wang et al. 2020).

5.1 Numerical settings

To confront our estimators with some challenging situations, we considered cases where the ratio p/n is large. More precisely, we choose the pairs (n, p), with $n \in \{150, 300\}$ and $p \in \{40, 80, 120\}$. In order to generate a sparse scenario, we chose the true regression parameter with only a few nonzero components. Herein, we present the results corresponding to $\boldsymbol{\beta}_0 = (1, 1, 1, 1, 1, 0, \dots, 0)^{\mathrm{T}} \in \mathbb{R}^p$, i.e. the regression parameter has only five nonzero components. In Section S.8.3, we consider a regression parameter with coordinates of different sizes combined with a non-diagonal matrix $\boldsymbol{\Sigma}$. Note that with these selections of the simulation parameters $\mathbb{E}(y_i)$ equals 0.50. In all cases, the number of Monte Carlo replications was $NR = 500$.

Henceforth, the clean samples setting is denoted **C0**. To study the impact of contamination, we have explored two settings by adding a proportion ε of atypical points. In the first contamination scheme, namely outliers of class **A**, we generated misclassified points $(\widetilde{y}, \widetilde{\mathbf{x}})$, where $\widetilde{\mathbf{x}} \sim N_p(0, 20\,\mathbf{I})$ and $\widetilde{y} = 1$ when $\gamma_0 + \widetilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\beta}_0 < 0$ and $\widetilde{y} = 0$, otherwise. Besides, outliers of class **B**, were obtained as in Croux and Haesbroeck (2003).

This means that given $m > 0$, we fixed $\widetilde{\mathbf{w}} = m\sqrt{p}\,\boldsymbol{\beta}_0/5$ and set $\widetilde{\mathbf{x}} = \widetilde{\mathbf{w}} + \widetilde{\mathbf{u}}$, where $\widetilde{\mathbf{u}} \sim N_p(\mathbf{0}, \mathbf{I}/100)$ is introduced so as to get distinct covariate values. The response $\widetilde{y}$, related to $\widetilde{\mathbf{x}}$, is always taken equal to 0. It is worth noticing that $\widetilde{\mathbf{w}}^{\mathrm{T}}\boldsymbol{\beta}_0 \approx m\sqrt{p}$, thus the leverage of the added points increases with m. The selected values of m are 0.5, 1, 2, 3, 4 and 5.

Summarizing, we consider the scenarios **CA1** and **CA2** which correspond to adding, respectively, a proportion $\varepsilon = 0.05$ and 0.10 of outliers of class **A** and **CB** where we add only 5% of outliers of class **B**, as in Croux and Haesbroeck (2003).

We compare the performance of the estimators based on the deviance, that is, when $\rho(t) = t$, labelled ML in all tables, with those obtained by bounding the deviance and also with their robust weighted versions constructed to control the leverage. The three bounded loss functions considered are $\rho(t) = 1 - \exp(-t)$ that leads to the least squares estimators, the loss functions ρ_c introduced by Croux and Haesbroeck (2003), given in (9), and $\rho(t) = (c+1)(1+\exp(-ct))$ related to the divergence estimators. For the last two loss functions, the tuning constant equals $c = 0.5$. These estimators are indicated with the subscript LS, M and DIV, respectively. To consider weighted versions of these estimators, define $D^2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$, the square of the Mahalanobis distance. We take weights $w(\mathbf{x}) = W(D^2(\mathbf{x}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}^{-1}))$, where to adjust for robustness $\widehat{\boldsymbol{\mu}}$ is the ℓ_1-median, $\widehat{\boldsymbol{\Sigma}}^{-1}$ is an estimator of $\boldsymbol{\Sigma}^{-1}$ computed using a robust graphical LASSO and W is the hard rejection weight function $W(t) = \mathbf{1}_{[0,c_w]}(t)$. The tuning constant c_w is adaptive and based on the quantiles of $D^2(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}^{-1})$. To compute $\widehat{\boldsymbol{\Sigma}}^{-1}$ we used the procedure defined in Öllerer and Croux (2015) and Tarr et al. (2016). More precisely, let $\boldsymbol{\Sigma}_{ij} = \sigma_i\sigma_j\rho_{ij}$, where $\rho_{ii} = 1$. On one hand, to estimate σ_j we used the median of the absolute deviations with respect to the median (MAD) of the j-th component, that is, the MAD of $\{x_{1j}, \ldots, x_{nj}\}$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. On the other hand, to estimate ρ_{ij} we use the Spearman correlation. The matrix $\widehat{\boldsymbol{\Sigma}}$ is defined element-wise as $\widehat{\boldsymbol{\Sigma}}_{ij} = \widehat{\sigma}_i\widehat{\sigma}_j\widehat{\rho}_{ij}$. Finally, we apply graphical LASSO (Friedman et al. 2008) to the matrix $\widehat{\boldsymbol{\Sigma}}$ in order to obtain $\widehat{\boldsymbol{\Sigma}}^{-1}$. These weighted estimators are labelled with the subscript WLS, WM or WDIV, according to the loss function considered.

For each loss function, different penalties are considered: LASSO, Sign and MCP, labelled with the superscript L, S and MCP, respectively. The non-sparse estimators without any penalization term are indicated with no superscript. In Sect. S.8.2, we include a comparison between SCAD and MCP penalties since in some regression settings when considering the classical estimators, the first one outperforms the latter. However, in our framework, as shown in the supplementary file, the results obtained for both penalties are similar. For that reason, we do not report here the results obtained with the SCAD penalty.

Under **C0** and scenarios **CA1** and **CA2**, we compare all described estimators. However, in view of the results obtained for these three situations and for the sake of brevity, under **CB** we only report the results for $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$ with penalties S and MCP.

To evaluate the performance of a given estimator $\widehat{\boldsymbol{\beta}}$, we consider three summary measures. In the following, let $\mathcal{T} = \{(y_{i,\mathcal{T}}, \mathbf{x}_{i,\mathcal{T}}), i = 1, \ldots, n_{\mathcal{T}}\}$, $n_{\mathcal{T}} = 100$, be a new sample generated independently from the training sample $\mathcal{M}$ and distributed as

C0. Given estimates $\widehat{\boldsymbol{\beta}}$ of the slope and $\widehat{\gamma}$ of the intercept computed from $\mathcal{M}$ and to compare the performance of the estimators, we compute the probability mean squared errors (PMSE), the true positive proportion (TPP) and the true null proportion (TNP) defined, respectively, as

$$\text{PMSE} = \frac{1}{n_{\mathcal{T}}} \sum_{i=1}^{n_{\mathcal{T}}} \left(F(\mathbf{x}_{i,\mathcal{T}}^{\mathrm{T}} \boldsymbol{\beta}_0 + \gamma_0) - F(\mathbf{x}_{i,\mathcal{T}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \widehat{\gamma}) \right)^2$$

$$\text{TPP} = \frac{\#\{j : 1 \le j \le p,\ \beta_{0,j} \neq 0,\ \widehat{\beta}_j \neq 0\}}{\#\{j : 1 \le j \le p,\ \beta_{0,j} \neq 0\}} \quad \text{and}$$

$$\text{TNP} = \frac{\#\{j : 1 \le j \le p,\ \beta_{0,j} = 0,\ \widehat{\beta}_j = 0\}}{\#\{j : 1 \le j \le p,\ \beta_{0,j} = 0\}} \,.$$

In all tables, we report the mean of the summary measures over 500 replications.

5.2 Results of the numerical study

Tables 1 and 2 sum up the results corresponding to **C0**, Tables 3, 4and 5 summarize contaminations **CA1** and **CA2**, while Tables 6, 7and 8 present the results obtained under scenario **CB**.

Table 1 shows that, for samples without contamination, the estimators penalized with MCP tend to achieve lower PMSE values than with the other penalties. In particular, for samples of size $n = 300$, the maximum likelihood estimators using the MCP penalty come to have PMSE values that are less than a half of those obtained with the LASSO penalty. That difference is even greater for the least squares estimator and for the M-estimators calculated with the function $\rho = \rho_c$ given in (9). Under **C0**, the robust weighted estimators give similar results to the unweighted ones with respect to all the considered measures (see Tables 1 and 2), showing that the weights do not impact on the procedure performance when samples are not contaminated.

As Table 1 reveals, the M-estimator penalized with LASSO loses more efficiency in terms of prediction than with the other penalties, reaching PMSE values that at least double those obtained with $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{\mathrm{L}}$. Indeed, when $n = 300$ and the sample is clean, the Sign and MCP penalties give lower PMSE values than the LASSO penalty. This fact can be explained by the non-negligible bias, already discussed in this paper, introduced by the LASSO penalty even when the ratio n/p is large. For both bounded penalties, all loss functions give very similar results.

As expected, in all situations, the non-penalized estimators give worse results than those obtained by regularizing the estimation procedure. In addition, the PMSE errors grow when the dimension increases. In particular, this growth is greater when using the Sign penalty for $n = 150$ and $p = 120$, where PMSE values for the M-estimates almost double those obtained with $n = 150$ and $p = 40$ for most estimators. As mentioned above, the case $(n, p) = (150, 120)$ poses a great challenge to the estimation of the regression parameter and to the selection of variables, as well.

Regarding the proportion of correct classifications and the proportions of true positive and null coefficients, all penalized estimators give similar results. It should be

Table 1 Mean over replications of PMSE under **C0**

n	150			300				150			300		
p	40	80	120	40	80	120		40	80	120	40	80	120
$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$	0.097	0.163	0.172	0.033	0.093	0.161	$\widehat{\boldsymbol{\beta}}_{\mathrm{WML}}$	0.098	0.163	0.172	0.033	0.094	0.161
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{ML}}$	0.023	0.028	0.034	0.012	0.014	0.015	$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WML}}$	0.023	0.028	0.035	0.012	0.014	0.015
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.028	0.050	0.090	0.008	0.010	0.011	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WML}}$	0.029	0.051	0.092	0.008	0.010	0.011
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.023	0.027	0.033	0.005	0.006	0.007	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WML}}$	0.023	0.027	0.034	0.005	0.006	0.007
$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$	0.135	0.146	0.157	0.113	0.126	0.135	$\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}$	0.135	0.147	0.157	0.113	0.126	0.134
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{LS}}$	0.047	0.048	0.052	0.038	0.038	0.039	$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WLS}}$	0.047	0.048	0.052	0.038	0.038	0.039
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{LS}}$	0.049	0.061	0.078	0.010	0.013	0.014	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WLS}}$	0.049	0.062	0.079	0.010	0.013	0.014
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{LS}}$	0.029	0.035	0.049	0.006	0.008	0.010	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WLS}}$	0.029	0.035	0.049	0.006	0.008	0.010
$\widehat{\boldsymbol{\beta}}_{\mathrm{DIV}}$	0.139	0.153	0.164	0.069	0.132	0.140	$\widehat{\boldsymbol{\beta}}_{\mathrm{WDIV}}$	0.139	0.152	0.163	0.071	0.132	0.141
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{DIV}}$	0.025	0.031	0.037	0.015	0.017	0.018	$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	0.026	0.031	0.037	0.015	0.017	0.018
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.035	0.050	0.079	0.009	0.011	0.012	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.035	0.050	0.079	0.009	0.011	0.012
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.025	0.029	0.039	0.006	0.007	0.007	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.025	0.029	0.039	0.006	0.007	0.007
$\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$	0.141	0.159	0.167	0.039	0.140	0.148	$\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$	0.142	0.159	0.167	0.040	0.140	0.149
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{M}}$	0.052	0.053	0.056	0.042	0.043	0.043	$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.052	0.053	0.056	0.042	0.043	0.044
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.031	0.041	0.059	0.008	0.011	0.012	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.032	0.041	0.058	0.008	0.011	0.012
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.024	0.030	0.040	0.005	0.008	0.009	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.024	0.030	0.041	0.006	0.008	0.009

Table 2 True positive proportion/true null proportion. No contamination model: scenario **C0**. Means over 500 replications

p	$n = 150$			$n = 300$		
	40	80	120	40	80	120
$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{ML}}$	1.00/0.61	1.00/0.68	0.99/0.73	1.00/0.69	1.00/0.71	1.00/0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.96/0.86	0.95/0.80	0.93/0.74	1.00/0.92	1.00/0.92	1.00/0.90
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.95/0.95	0.93/0.93	0.88/0.92	1.00/0.97	1.00/0.92	1.00/0.92
$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$	1.00/0.10	1.00/0.19	1.00/0.25	1.00/0.05	1.00/0.10	1.00/0.14
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{LS}}$	0.98/0.84	0.98/0.85	0.97/0.86	1.00/0.92	1.00/0.92	1.00/0.92
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{LS}}$	0.86/0.95	0.85/0.96	0.81/0.95	1.00/0.95	1.00/0.96	1.00/0.96
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{LS}}$	0.89/0.98	0.89/0.97	0.84/0.97	1.00/0.99	1.00/0.98	0.99/0.97
$\widehat{\boldsymbol{\beta}}_{\mathrm{DIV}}$	1.00/0.05	1.00/0.05	1.00/0.04	1.00/0.01	1.00/0.06	1.00/0.05
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{DIV}}$	1.00/0.61	1.00/0.67	0.99/0.71	1.00/0.68	1.00/0.71	1.00/0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.93/0.91	0.92/0.86	0.90/0.82	1.00/0.93	1.00/0.94	1.00/0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.93/0.95	0.91/0.93	0.86/0.93	1.00/0.98	1.00/0.94	1.00/0.93
$\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$	1.00/0.03	1.00/0.02	1.00/0.01	1.00/0.00	1.00/0.03	1.00/0.02
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{M}}$	0.97/0.88	0.97/0.88	0.95/0.88	1.00/0.94	1.00/0.94	0.99/0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.94/0.94	0.91/0.96	0.82/0.96	1.00/0.95	1.00/0.96	1.00/0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.92/0.98	0.91/0.97	0.85/0.97	1.00/0.99	1.00/0.98	0.99/0.97
$\widehat{\boldsymbol{\beta}}_{\mathrm{WML}}$	1.00/0.00	1.00/0.00	1.00/0.01	1.00/0.00	1.00/0.00	1.00/0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WML}}$	1.00/0.62	1.00/0.68	0.99/0.72	1.00/0.69	1.00/0.71	1.00/0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WML}}$	0.95/0.86	0.95/0.80	0.94/0.74	1.00/0.92	1.00/0.92	1.00/0.90
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WML}}$	0.94/0.95	0.92/0.93	0.88/0.92	1.00/0.97	1.00/0.92	1.00/0.92
$\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}$	1.00/0.11	1.00/0.20	1.00/0.25	1.00/0.05	1.00/0.10	1.00/0.15
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WLS}}$	0.98/0.84	0.98/0.85	0.97/0.86	1.00/0.92	1.00/0.92	1.00/0.92
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WLS}}$	0.85/0.96	0.85/0.96	0.79/0.95	1.00/0.95	1.00/0.96	1.00/0.96
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WLS}}$	0.89/0.98	0.89/0.97	0.83/0.97	1.00/0.99	1.00/0.98	0.99/0.98
$\widehat{\boldsymbol{\beta}}_{\mathrm{WDIV}}$	1.00/0.05	1.00/0.05	1.00/0.04	1.00/0.01	1.00/0.06	1.00/0.05
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	1.00/0.61	1.00/0.67	0.99/0.71	1.00/0.68	1.00/0.71	1.00/0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.93/0.90	0.92/0.85	0.89/0.82	1.00/0.93	1.00/0.94	1.00/0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.93/0.96	0.91/0.93	0.86/0.93	1.00/0.98	1.00/0.93	1.00/0.93
$\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$	1.00/0.03	1.00/0.02	1.00/0.01	1.00/0.00	1.00/0.03	1.00/0.03
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.97/0.87	0.97/0.88	0.95/0.88	1.00/0.94	1.00/0.94	1.00/0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.93/0.94	0.91/0.96	0.82/0.96	1.00/0.95	1.00/0.96	1.00/0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.92/0.98	0.91/0.97	0.85/0.97	1.00/0.99	1.00/0.98	1.00/0.97

Table 3 Means over replications of PMSE under **CA1** and **CA2**

p	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
	$n = 150$			$n = 300$			$n = 150$			$n = 300$		
	40	80	120	40	80	120	40	80	120	40	80	120
$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$	0.094	0.194	0.205	0.059	0.094	0.179	0.104	0.217	0.230	0.079	0.103	0.158
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{ML}}$	0.077	0.079	0.081	0.070	0.072	0.071	0.104	0.104	0.106	0.102	0.102	0.104
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.082	0.100	0.145	0.076	0.082	0.078	0.101	0.106	0.135	0.097	0.101	0.102
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.074	0.078	0.086	0.063	0.069	0.071	0.105	0.106	0.109	0.104	0.106	0.107
$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$	0.148	0.166	0.178	0.112	0.138	0.151	0.174	0.195	0.207	0.101	0.163	0.175
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{LS}}$	0.064	0.066	0.071	0.050	0.051	0.054	0.095	0.098	0.100	0.082	0.085	0.089
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{LS}}$	0.069	0.079	0.091	0.023	0.028	0.034	0.088	0.097	0.102	0.055	0.062	0.075
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{LS}}$	0.049	0.062	0.078	0.017	0.021	0.027	0.079	0.093	0.104	0.040	0.050	0.065
$\widehat{\boldsymbol{\beta}}_{\mathrm{DIV}}$	0.154	0.172	0.183	0.056	0.145	0.157	0.168	0.198	0.212	0.064	0.171	0.181
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{DIV}}$	0.049	0.053	0.059	0.033	0.037	0.040	0.089	0.091	0.093	0.077	0.079	0.081
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.061	0.077	0.107	0.024	0.028	0.033	0.087	0.095	0.113	0.066	0.067	0.075
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.047	0.056	0.068	0.021	0.022	0.027	0.080	0.088	0.096	0.051	0.055	0.064
$\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$	0.141	0.178	0.191	0.048	0.150	0.167	0.126	0.204	0.220	0.064	0.130	0.191
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{M}}$	0.078	0.080	0.083	0.067	0.068	0.069	0.103	0.104	0.106	0.100	0.100	0.101
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.072	0.080	0.089	0.035	0.037	0.047	0.094	0.099	0.103	0.087	0.089	0.095
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.054	0.065	0.082	0.024	0.027	0.035	0.094	0.104	0.108	0.078	0.086	0.096

Table 3 continued

p	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
	$n = 150$			$n = 300$			$n = 150$			$n = 300$		
	40	80	120	40	80	120	40	80	120	40	80	120
$\widehat{\boldsymbol{\beta}}_{\mathrm{WML}}$	0.097	0.171	0.187	0.033	0.094	0.165	0.097	0.180	0.199	0.033	0.094	0.169
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WML}}$	0.023	0.029	0.034	0.012	0.014	0.016	0.023	0.029	0.035	0.012	0.014	0.016
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WML}}$	0.030	0.053	0.090	0.008	0.010	0.011	0.033	0.052	0.087	0.010	0.011	0.011
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WML}}$	0.023	0.031	0.039	0.005	0.007	0.008	0.025	0.036	0.047	0.006	0.009	0.010
$\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}$	0.137	0.160	0.175	0.113	0.132	0.145	0.140	0.171	0.193	0.115	0.138	0.156
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WLS}}$	0.047	0.048	0.052	0.038	0.038	0.039	0.046	0.048	0.052	0.038	0.039	0.039
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WLS}}$	0.050	0.067	0.079	0.010	0.013	0.015	0.053	0.071	0.085	0.011	0.014	0.015
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WLS}}$	0.030	0.042	0.056	0.006	0.010	0.012	0.032	0.051	0.065	0.007	0.012	0.016
$\widehat{\boldsymbol{\beta}}_{\mathrm{WDIV}}$	0.140	0.164	0.179	0.070	0.137	0.150	0.142	0.174	0.195	0.070	0.141	0.158
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	0.025	0.031	0.037	0.015	0.017	0.019	0.025	0.031	0.038	0.015	0.017	0.019
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.035	0.052	0.077	0.009	0.011	0.012	0.039	0.054	0.072	0.010	0.012	0.012
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.025	0.035	0.045	0.006	0.008	0.009	0.026	0.039	0.052	0.006	0.009	0.011
$\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$	0.143	0.170	0.182	0.040	0.142	0.155	0.144	0.177	0.197	0.039	0.145	0.161
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.052	0.053	0.055	0.042	0.043	0.044	0.052	0.053	0.056	0.042	0.043	0.045
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.037	0.049	0.061	0.009	0.011	0.012	0.049	0.055	0.066	0.016	0.012	0.013
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.025	0.037	0.049	0.006	0.010	0.012	0.025	0.042	0.057	0.006	0.011	0.014

Table 4 True positive proportion/true null proportion. 5% contamination model: scenario **CA1**. Means over replications

p	$n = 150$			$n = 300$		
	40	80	120	40	80	120
$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{ML}}$	0.77/0.69	0.75/0.76	0.72/0.79	0.90/0.65	0.89/0.71	0.90/0.74
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.55/0.83	0.72/0.71	0.86/0.66	0.60/0.92	0.60/0.86	0.61/0.80
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.69/0.79	0.69/0.82	0.67/0.84	0.78/0.78	0.75/0.81	0.73/0.82
$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$	1.00/0.08	1.00/0.17	0.99/0.25	1.00/0.03	1.00/0.08	1.00/0.12
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{LS}}$	0.91/0.70	0.91/0.73	0.88/0.76	0.97/0.75	0.99/0.76	0.99/0.75
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{LS}}$	0.70/0.95	0.68/0.95	0.62/0.95	0.98/0.87	0.98/0.94	0.97/0.94
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{LS}}$	0.77/0.96	0.74/0.96	0.64/0.96	0.96/0.97	0.97/0.97	0.94/0.97
$\widehat{\boldsymbol{\beta}}_{\mathrm{DIV}}$	1.00/0.05	1.00/0.06	1.00/0.05	1.00/0.00	1.00/0.05	1.00/0.08
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{DIV}}$	0.96/0.50	0.95/0.57	0.94/0.63	0.99/0.48	1.00/0.52	1.00/0.54
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.78/0.85	0.80/0.80	0.82/0.78	0.98/0.79	0.98/0.86	0.96/0.84
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.83/0.89	0.82/0.89	0.78/0.89	0.96/0.90	0.97/0.89	0.95/0.88
$\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$	1.00/0.02	1.00/0.03	1.00/0.03	1.00/0.00	1.00/0.02	1.00/0.03
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{M}}$	0.83/0.74	0.82/0.77	0.81/0.79	0.94/0.75	0.95/0.77	0.96/0.76
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.59/0.96	0.61/0.96	0.55/0.96	0.89/0.90	0.95/0.93	0.89/0.94
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.70/0.96	0.67/0.96	0.57/0.96	0.90/0.96	0.95/0.97	0.91/0.96
$\widehat{\boldsymbol{\beta}}_{\mathrm{WML}}$	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WML}}$	1.00/0.61	1.00/0.68	1.00/0.72	1.00/0.69	1.00/0.72	1.00/0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WML}}$	0.95/0.87	0.93/0.81	0.92/0.76	1.00/0.92	1.00/0.92	1.00/0.90
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WML}}$	0.95/0.95	0.90/0.92	0.85/0.92	1.00/0.97	1.00/0.93	0.99/0.92
$\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}$	1.00/0.10	1.00/0.19	1.00/0.22	1.00/0.05	1.00/0.10	1.00/0.13
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WLS}}$	0.99/0.85	0.98/0.85	0.97/0.85	1.00/0.93	1.00/0.92	1.00/0.91
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WLS}}$	0.86/0.95	0.81/0.95	0.76/0.95	1.00/0.95	1.00/0.96	1.00/0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WLS}}$	0.89/0.98	0.84/0.97	0.77/0.97	1.00/0.99	0.99/0.98	0.99/0.97
$\widehat{\boldsymbol{\beta}}_{\mathrm{WDIV}}$	1.00/0.05	1.00/0.04	1.00/0.02	1.00/0.01	1.00/0.06	1.00/0.05
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	1.00/0.61	1.00/0.67	0.99/0.71	1.00/0.69	1.00/0.71	1.00/0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.93/0.90	0.89/0.87	0.87/0.83	1.00/0.92	1.00/0.94	1.00/0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.94/0.95	0.89/0.93	0.83/0.92	1.00/0.98	1.00/0.94	0.99/0.93
$\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$	1.00/0.03	1.00/0.02	1.00/0.01	1.00/0.00	1.00/0.03	1.00/0.02
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.97/0.88	0.97/0.88	0.96/0.88	1.00/0.95	1.00/0.94	1.00/0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.86/0.89	0.83/0.95	0.77/0.95	0.99/0.90	1.00/0.93	1.00/0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.92/0.97	0.87/0.96	0.79/0.97	1.00/0.99	0.99/0.97	0.99/0.97

Table 5 True positive proportion/true null proportion. 10% contamination model: scenario **CA2**. Means over replications

p	$n = 150$			$n = 300$		
	40	80	120	40	80	120
$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{ML}}$	0.45/0.78	0.45/0.84	0.42/0.87	0.56/0.76	0.58/0.83	0.55/0.86
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.34/0.93	0.41/0.86	0.58/0.77	0.47/0.98	0.44/0.94	0.37/0.91
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.36/0.80	0.36/0.87	0.35/0.88	0.41/0.82	0.35/0.89	0.37/0.91
$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$	0.99/0.06	0.99/0.14	0.99/0.20	1.00/0.01	1.00/0.06	1.00/0.09
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{LS}}$	0.59/0.77	0.58/0.82	0.57/0.84	0.76/0.70	0.78/0.74	0.77/0.76
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{LS}}$	0.48/0.96	0.42/0.97	0.43/0.96	0.79/0.86	0.79/0.94	0.73/0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{LS}}$	0.53/0.94	0.45/0.95	0.40/0.96	0.84/0.93	0.83/0.95	0.73/0.95
$\widehat{\boldsymbol{\beta}}_{\mathrm{DIV}}$	1.00/0.03	1.00/0.07	1.00/0.07	1.00/0.00	1.00/0.04	1.00/0.06
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{DIV}}$	0.65/0.63	0.66/0.70	0.65/0.74	0.83/0.55	0.81/0.61	0.80/0.65
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.49/0.89	0.51/0.86	0.56/0.82	0.71/0.79	0.72/0.86	0.67/0.86
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.64/0.81	0.59/0.86	0.55/0.88	0.85/0.77	0.86/0.83	0.82/0.85
$\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$	1.00/0.00	1.00/0.04	1.00/0.05	1.00/0.00	1.00/0.00	1.00/0.02
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{M}}$	0.49/0.83	0.53/0.85	0.49/0.88	0.61/0.81	0.66/0.83	0.66/0.84
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.38/0.98	0.38/0.97	0.31/0.97	0.49/0.98	0.60/0.97	0.52/0.96
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.36/0.94	0.34/0.95	0.33/0.96	0.52/0.94	0.48/0.95	0.48/0.96
$\widehat{\boldsymbol{\beta}}_{\mathrm{WML}}$	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WML}}$	1.00/0.61	1.00/0.67	1.00/0.71	1.00/0.69	1.00/0.71	1.00/0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WML}}$	0.93/0.88	0.88/0.83	0.90/0.77	1.00/0.90	1.00/0.92	1.00/0.91
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WML}}$	0.94/0.94	0.88/0.93	0.81/0.92	1.00/0.97	1.00/0.95	0.99/0.93
$\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}$	1.00/0.10	1.00/0.18	1.00/0.19	1.00/0.05	1.00/0.10	1.00/0.12
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WLS}}$	0.98/0.85	0.99/0.85	0.98/0.85	1.00/0.93	1.00/0.92	1.00/0.91
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WLS}}$	0.85/0.93	0.78/0.94	0.71/0.95	1.00/0.94	1.00/0.95	1.00/0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WLS}}$	0.88/0.97	0.80/0.97	0.70/0.97	1.00/0.99	0.99/0.97	0.98/0.96
$\widehat{\boldsymbol{\beta}}_{\mathrm{WDIV}}$	1.00/0.05	1.00/0.05	1.00/0.02	1.00/0.01	1.00/0.06	1.00/0.06
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	1.00/0.60	1.00/0.66	0.99/0.70	1.00/0.68	1.00/0.70	1.00/0.71
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.91/0.91	0.87/0.87	0.84/0.85	1.00/0.91	1.00/0.93	1.00/0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.93/0.94	0.87/0.93	0.80/0.92	1.00/0.98	1.00/0.95	0.99/0.94
$\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$	1.00/0.03	1.00/0.02	1.00/0.00	1.00/0.00	1.00/0.03	1.00/0.03
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.97/0.88	0.97/0.87	0.96/0.87	1.00/0.95	1.00/0.94	1.00/0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.75/0.88	0.74/0.94	0.69/0.95	0.93/0.87	0.98/0.89	0.98/0.91
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.92/0.97	0.84/0.96	0.73/0.96	1.00/0.99	1.00/0.97	0.99/0.96

Table 6 Means over replications of PMSE under **CB**

m	0.5	1	2	3	4	5		0.5	1	2	3	4	5
	$n=150$												
	$p=40$							$p=80$					
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.047	0.080	0.095	0.105	0.112	0.116	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.074	0.096	0.112	0.136	0.149	0.136
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.040	0.068	0.104	0.121	0.129	0.130	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.061	0.098	0.131	0.138	0.137	0.127
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.047	0.071	0.085	0.089	0.089	0.086	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.073	0.087	0.095	0.093	0.091	0.093
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.038	0.051	0.067	0.068	0.062	0.056	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.057	0.076	0.076	0.066	0.060	0.062
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.048	0.071	0.086	0.036	0.039	0.040	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.074	0.088	0.096	0.054	0.053	0.052
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.039	0.051	0.068	0.027	0.026	0.026	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.058	0.076	0.076	0.043	0.043	0.043
	$n=300$												
	$p=80$							$p=120$					
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.032	0.073	0.099	0.111	0.114	0.115	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.050	0.082	0.109	0.113	0.115	0.115
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.030	0.080	0.113	0.123	0.122	0.115	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.046	0.100	0.122	0.125	0.115	0.114
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.022	0.036	0.041	0.034	0.034	0.045	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.027	0.048	0.045	0.039	0.054	0.064
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.019	0.023	0.020	0.015	0.013	0.022	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.025	0.031	0.018	0.014	0.026	0.046
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.022	0.036	0.041	0.011	0.011	0.011	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.026	0.049	0.046	0.012	0.012	0.012
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.019	0.023	0.021	0.010	0.010	0.010	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.025	0.033	0.018	0.012	0.012	0.013

Table 7 True positive proportions/true null proportions for scenario **CB** with $n = 150$. Means over replications

m	0.5	1	2	3	4	5
	$p = 40$					
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.83/0.92	0.53/0.95	0.40/0.95	0.36/0.95	0.37/0.94	0.40/0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.85/0.92	0.66/0.90	0.50/0.84	0.47/0.77	0.45/0.74	0.45/0.73
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.78/0.94	0.57/0.96	0.44/0.97	0.42/0.96	0.44/0.96	0.47/0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.82/0.96	0.70/0.96	0.57/0.96	0.57/0.95	0.61/0.95	0.68/0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.77/0.94	0.56/0.96	0.44/0.97	0.92/0.91	0.89/0.92	0.88/0.92
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.81/0.96	0.70/0.96	0.57/0.96	0.92/0.97	0.93/0.96	0.92/0.97
	$p = 80$					
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.62/0.93	0.45/0.93	0.38/0.91	0.46/0.84	0.54/0.81	0.51/0.86
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.70/0.91	0.54/0.88	0.46/0.81	0.40/0.80	0.45/0.80	0.47/0.86
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.64/0.96	0.50/0.96	0.41/0.96	0.43/0.96	0.44/0.96	0.44/0.97
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.69/0.96	0.50/0.96	0.52/0.95	0.58/0.95	0.65/0.95	0.61/0.96
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.64/0.96	0.50/0.96	0.40/0.96	0.78/0.95	0.79/0.95	0.79/0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.69/0.96	0.51/0.96	0.51/0.95	0.83/0.96	0.83/0.96	0.83/0.96

Table 8 True positive proportions/true null proportions for scenario **CB** with $n = 300$. Means over replications

m	0.5	1	2	3	4	5
	$p = 80$					
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.97/0.95	0.73/0.96	0.57/0.96	0.60/0.95	0.65/0.95	0.62/0.97
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.94/0.95	0.72/0.91	0.47/0.83	0.46/0.76	0.52/0.80	0.49/0.92
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.98/0.95	0.90/0.95	0.83/0.95	0.84/0.95	0.84/0.95	0.78/0.96
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.96/0.97	0.93/0.97	0.91/0.95	0.95/0.94	0.96/0.94	0.90/0.96
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.97/0.95	0.90/0.95	0.81/0.95	1.00/0.95	1.00/0.95	1.00/0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.96/0.97	0.93/0.97	0.91/0.94	1.00/0.97	0.99/0.97	0.99/0.97
	$p = 120$					
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{ML}}$	0.90/0.95	0.65/0.94	0.48/0.92	0.62/0.93	0.61/0.95	0.64/0.96
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{ML}}$	0.87/0.93	0.63/0.89	0.41/0.82	0.48/0.81	0.45/0.92	0.50/0.96
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.95/0.95	0.83/0.95	0.80/0.95	0.82/0.95	0.75/0.96	0.75/0.96
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.94/0.96	0.89/0.96	0.92/0.93	0.96/0.93	0.89/0.96	0.81/0.97
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.96/0.95	0.83/0.95	0.81/0.95	1.00/0.95	1.00/0.95	1.00/0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.94/0.96	0.88/0.96	0.93/0.93	0.98/0.97	0.98/0.97	0.98/0.97

mentioned that, when the LASSO penalty is used, lower TNP values are obtained than with other penalties, giving rise to less sparse estimators. This procedure seems to be less skilled than MCP to identify as 0 those coefficients associated with explanatory variables that are not involved in the model. This drawback is also observed, although to a lesser extent, when considering the divergence estimator or the maximum likelihood one, both combined with the Sign penalty (see Table 2).

The sensitivity to atypical data of estimators based on $\rho(t) = t$ and $w \equiv 1$, combined with any of the considered penalties, becomes evident all along the tables. On one hand, Table 3 shows that, when outliers following schemes **CA1** or **CA2** are introduced, the obtained PMSE are at least three times those obtained for uncontaminated samples. Note that, for instance, when $n = 300$ the reported values for PMSE may be even 10 times larger under this contamination scheme than for clean samples. The only exception is when $n = 150$ and $p = 120$, where as mentioned above the maximum likelihood estimators combined with the Sign penalty already leads to large values of PMSE under **C0**. Table 3 reveals that, under contamination patterns **CA1** and **CA2**, the best behaviour, in terms of stability, is attained by the penalized weighted M-estimators. In fact, their probability mean squared errors (PMSE) are close to those obtained for clean samples with the bounded penalties Sign and MCP. The benefits of using weighted estimators are also reflected in the proportions of true positives and zeros, as illustrated in Tables 4 and 5. In the case of these latter measures, the LASSO penalty gives the highest values of the probability of true positives generally in detriment of the TNP values since, as we mentioned, this penalty has more difficulties in the identification of non-active explanatory variables.

It is worth noticing that, under **CA1** and **CA2**, the unweighted estimators have higher PMSE values than their weighted versions, especially when $n = 150$ (see Table 3). Under **CA2**, these values can double those obtained with the estimators that control the leverage of the covariates. Among the estimators with $w \equiv 1$, those that give lower PMSE values are the procedures corresponding to $\rho = \rho_{\text{DIV}}$ and those based on the least squares method when combined with the Sign and MCP penalties, in particular when $n = 300$.

In scenario **CA1**, the most stable estimators are those based on bounded loss functions. For example, Table 4 shows that the procedure based on $\rho(t) = t$ is the only one having problems with this level of contamination. On the other hand, the loss function introduced by Croux and Haesbroeck (2003) leads to more sparse estimators than those obtained with $\rho = \rho_{\text{DIV}}$ and $\rho(t) = 1 - \exp(-t)$.

Table 5 shows that as the level of contamination increases (scheme **CA2**), all estimators seem to become more sparse since the values of TPP tend to decrease. This effect directly impacts on measure TPP that decreases almost by half in unweighted estimators. As expected, this behaviour is more pronounced when using the Sign and MCP penalties combined with $\rho(t) = t$. Although to a lesser extent, the M-estimators with $\rho = \rho_c$ given in (9) are also affected by this contamination scheme. With respect to the ability to detect active variables, in most cases, weighted estimators achieve similar results to those obtained under **C0**.

With respect to the effect of the contamination **CB** on the penalized maximum likelihood estimator, Table 6, which reports the PMSE under this scheme, illustrates that the PMSE of the penalized maximum likelihood estimators is much larger than those obtained for the weighted or unweighted M-estimators. This effect is more evident when m is larger than 3. In contrast, the weighted M-estimators are more stable. As expected, for mild outliers ($m = 1, 2$) the PMSE of the weighted M-estimators increases and then decreases for larger values of the slope, attaining values similar to those reported for clean samples. In all cases, Table 6 also shows the advantage of combining weighted M-estimators with the MCP penalty. For $n = 300$, the performance of weighted M-estimators is very similar when combined either MCP or the Sign penalty.

Regarding the performance under **CB** in terms of measures TPP and TNP, as observed in Tables 7 and 8, the true positive proportions are reduced compared to those obtained for clean samples, attaining, in some cases, proportions smaller than 0.5. Similar conclusions are valid for the M-estimators, $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}^{\mathrm{MCP}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}^{\mathrm{S}}$. It is worth noticing that the effect of adding outliers on the non-penalized maximum likelihood estimator, $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$, has been studied in Croux et al. (2002) who observed that $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$ never explodes to infinity, but rather breaks down to zero when adding severe outliers to a dataset. This fact may explain the TPP behaviour observed in Tables 7 and 8. Indeed, similar arguments to those considered in the proof of Theorem 2 in Croux et al. (2002) allow to show that the penalized maximum likelihood estimator also shrinks to zero when adding outliers, which explains the behaviour of the measure TPP.

With respect to the weighted M-estimators, $\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}^{\mathrm{S}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}^{\mathrm{MCP}}$, the measure TPP shows some sensitivity for small values of the slope m ($m = 1, 2$) when $n = 150$, but recovers values close to 1 when the slope m increases. Notice that the intermediate values $m = 1, 2$ correspond to mild outliers that are the most difficult ones to be detected. It is worth mentioning that the TNP values obtained under **CB** are similar to those obtained for uncontaminated samples, except for estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{\mathrm{MCP}}$ that seems to be the most affected by this type of contamination.

Summarizing, for the studied contaminations, the weighted M-estimators based on the function $\rho = \rho_c$ given in (9) combined with the MCP and Sign penalties, turn out to be the most stable and reliable among the considered procedures.

6 Real data analysis

In this section, we study a dataset corresponding to the Diagnostic Wisconsin Breast Cancer Database which is available at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. Based on the results obtained in the numerical experiments reported in Sect. 5.1, we only illustrate the performance of the M-estimators computed with the (Croux and Haesbroeck 2003) loss function and of the classical ones by using different penalties. For the robust estimators, the tuning constants are equal to those considered in Sect. 5.1.

Ten real-valued features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass and they describe characteristics of the cell nuclei present in the image. Measured attributes are related to: radius (mean of distances from centre to points on the perimeter), texture (standard deviation of grey-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness

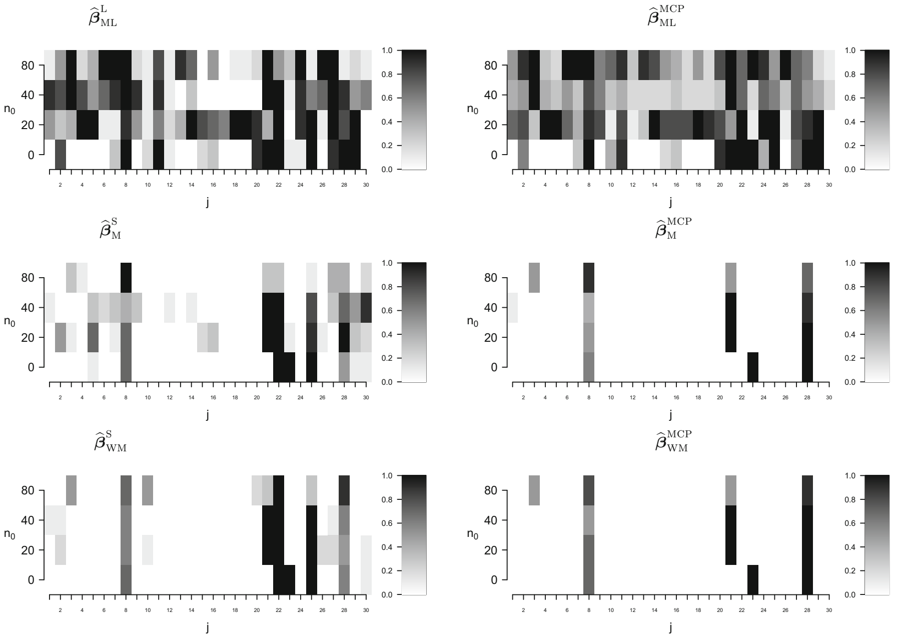

Fig. 1 Grey-scale representation of measures $\Pi_{a,j}$, $1 \leq j \leq 30$ for each method and number of atypical points introduced artificially

($perimeter^2/area - 1.0$), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry and fractal dimension. For each of these features the mean, the standard deviation and the maximum among all the nuclei of the image were computed, generating a total of $p = 30$ covariates for each image. From the $n = 569$ tumours, 357 were benign and 212 malignant and the goal is to predict the type of tumour from the $p = 30$ covariates.

From this dataset, we want to assess the impact of artificial outliers on the variable selection capability of different methods. For this purpose, we add n_0 atypical observations artificially. Each outlier $(\widetilde{y}, \widetilde{\mathbf{x}})$ was generated as follows. In a first step we compute the weighted M-estimator with MCP penalty, $(\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}, \widehat{\gamma}^{\mathrm{MCP}}_{\mathrm{WM}})$, with the original points and then, we generate $\widetilde{\mathbf{x}} \sim N_p(\mathbf{0}, 100\,\mathbf{I})$ and define a bad classified observations as $\widetilde{y} = 1$ when $\widetilde{\mathbf{x}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}} + \widehat{\gamma}^{\mathrm{MCP}}_{\mathrm{WM}} < 0$ and 0, otherwise. We add $n_0 = 0, 20, 40$ and 80 outliers. Given each contaminated set, we split the data in 10 folds of approximately the same size. For each estimation method and each subset i ($1 \leq i \leq 10$), we obtain $\widehat{\boldsymbol{\beta}}^{(-i)}$ and $\widehat{\gamma}^{(-i)}$, the slope and intercept estimates computed without the observations that lie in the i-th subset. Then, for each variable, we evaluate the fraction of times that it is detected as active among the 10 folds as $\Pi_{a,j} = \#\{i : \widehat{\boldsymbol{\beta}}^{(-i)}_j \neq 0\}/10$ for $1 \leq j \leq 30$. Note that this quantity depends on the estimator that is used and on n_0 and, regarding variable selection, it attempts to capture the stability of each method against outliers. In each row of the plots of Fig. 1, for each estimator and each value of n_0, we show a grey-scale representation of the measures $\Pi_{a,1}, \ldots, \Pi_{a,30}$.

As illustrated in Fig. 1, for the considered contamination, the non-robust estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{\mathrm{L}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{\mathrm{MCP}}$ show a very unstable and erratic variable selection, making evident their sensitivity to outliers. The results regarding $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{\mathrm{S}}$ are not included just for brevity since they lead to similar conclusions. In contrast, the robust procedures based on the (Croux and Haesbroeck 2003) loss function select approximately the same subset of covariates, regardless of the amount n_0 of added outliers, showing a stable identification of active variables. In particular, the hard rejection weighted estimators are more stable than their unweighted counterparts, when using the Sign penalty. The robust estimators with MCP penalty are more sparse than when using the Sign penalty, which can be explained by means of the theoretical properties studied in Sect. 4.1.

7 Concluding remarks

The logistic regression model may be used for classification purposes when covariates with predictive capability are observed. When the regression coefficients are assumed to be sparse, i.e. when only a few explanatory variables are active, the problem of joint estimation and automatic variable selection needs to be considered. In these circumstances, the statistical challenge of obtaining sparse and robust estimators that are computationally feasible and provide variable selection should be complemented with the study of their asymptotic properties. For this reason, under a logistic regression model, we accomplished the goal of obtaining more reliable estimators in the presence of atypical data, which automatically selects variables, using weighted penalized M-procedures. The obtained results are derived for a broad family of penalty functions, which include the LASSO, ADALASSO, Ridge, SCAD and MCP penalties. Besides these known penalties, we also consider the Sign penalization, which has an intuitive motivation, a simple expression and has not been exploited in the framework of robust variable selection.

An in-depth study of the theoretical properties of the proposed methods is presented. In particular, under very general conditions, we establish consistency results for a wide family of penalty functions. Besides, to study variable selection and oracle properties, we distinguish the case of Lipschitz functions, such as the Sign, from that of penalties that can be written as a sum of twice differentiable univariate functions, eventually random. These two points make a difference with respect to Sect. 2 in Avella-Medina and Ronchetti (2018) where the conditions to obtain general results regarding sparsity and asymptotic normality are more restrictive than those given herein for the logistic regression model.

In addition to obtaining variable selection properties of the proposed estimators, we derive expressions for their asymptotic distribution. In particular, it is shown that the choice of the penalty function plays a fundamental role in this case. Specifically, we obtain that by using the random penalty ADALASSO or penalties which are constant from one point onwards (such as SCAD or MCP), the estimators have the desired oracle property. The assumptions required to derive these results are very undemanding, which shows that these methods can be applied in very diverse contexts.

We also proposed a robust cross-validation procedure and numerically showed its advantage over the classical one. Through an extensive simulation study, we compared the behaviour of classical and robust estimators for different choices for the loss function and penalty. The obtained results illustrate that robust methods have a performance similar to the classical ones for clean samples and behave much better in contaminated scenarios, showing greater reliability. On the other hand, we showed that the results obtained when using penalties bounded as the Sign or MCP were remarkably better than those obtained when using convex penalties such as LASSO. The penalized weighted M-estimators based on the function $\rho = \rho_c$ defined in Croux and Haesbroeck (2003) combined with the MCP and Sign penalties were the most stable and reliable among the considered procedures. Finally, the proposed methods are applied to two datasets, where the robust estimators combined with bounded penalties showed their advantages over the classical ones.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11749-021-00792-w.

Funding This research was partially supported by Grant 20020170100022BA from the Universidad de Buenos Aires and PICT 2018-00740 from ANPCYT at Buenos Aires, Argentina, and also by the Spanish Project MTM2016-76969P from the Ministry of Economy, Industry and Competitiveness (MINECO/AEI/FEDER, UE) (Ana Bianco and Graciela Boente).

References

Avella-Medina M, Ronchetti E (2018) Robust and consistent variable selection in high-dimensional generalized linear models. Biometrika 105:31–44

Basu A, Gosh A, Mandal A, Martin N, Pardo L (2017) A Wald-type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator. Electr J Stat 11:2741–2772

Bianco A, Martínez E (2009) Robust testing in the logistic regression model. Comput Stat Data Anal 53:4095–4105

Bianco A, Yohai V (1996) Robust estimation in the logistic regression model. Lecture Notes Stat 109:17–34

Bondell HD (2005) Minimum distance estimation for the logistic regression model. Biometrika 92:724–731

Bondell HD (2008) A characteristic function approach to the biased sampling model, with application to robust logistic regression. J Stat Plann Inference 138:742–755

Cantoni E, Ronchetti E (2001) Robust inference for generalized linear models. J Am Stat Assoc 96:1022–1030

Chi EC, Scott DW (2014) Robust parametric classification and variable selection by a minimum distance criterion. J Comput Graph Stat 23:111–128

Croux C, Flandre C, Haesbroeck G (2002) The breakdown behavior of the maximum likelihood estimator in the logistic regression model. Stat Probabil Lett 60:377–386

Croux C, Haesbroeck G (2003) Implementing the Bianco and Yohai estimator for logistic regression. Comput Stat Data Anal 44:273–295

Efron B, Hastie T (2016) Computer age statistical inference. Cambridge University Press, Cambridge

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least Angle Regression. Annals Stat 32:407–499

Esser E, Lou Y, Xin J (2013) A method for finding structured sparse solutions to nonnegative least squares problems with applications. SIAM J Imag Sci 6:2010–2046

Fan J, Li R (2001) Variable selection via non-concave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360

Frank LE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics 35:109–135

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9:432–441

Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the Lasso and generalizations. Chapman and Hall, London

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67

Knight K, Fu W (2000) Asymptotics for Lasso-type estimators. Annals Stat 28:1356–1378

Kurnaz FS, Hoffmann I, Filzmoser P (2018) Robust and sparse estimation methods for high-dimensional linear and logistic regression. Chemomet Intell Lab Syst 172:211–222

Meinshausen N (2007) Relaxed Lasso. Comput Stat Data Anal 52:374–393

Öllerer V, Croux C (2015) Robust high-dimensional precision matrix estimation. In: Nordhausen K, Taskinen S (eds) Modern nonparametric, robust and multivariate methods. Springer, Cham, pp 325–350

Park H, Konishi S (2016) Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection. J Stat Comput Simul 86:1450–1461

Rahimi Y, Wang C, Dong H, Lou Y (2019) A scale invariant approach for sparse signal recovery. SIAM J Sci Comput 41:3649–3672

Smucler E, Yohai VJ (2017) Robust and sparse estimators for linear regression models. Comput Stat Data Anal 111:116–130

Tarr G, Müller S, Weber NC (2016) Robust estimation of precision matrices under cellwise contamination. Comput Stat Data Anal 93:404–420

Tibshirani J, Manning CD (2013) Robust Logistic Regression using Shift Parameters. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, pp. 124-129

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

van de Geer S, Müller P (2012) Quasi-likelihood and/or robust estimation in high dimensions. Stat Sci 27:469–480

Wang C, Yan M, Rahimi Y, Lou Y (2020) Accelerated schemes for the L_1/L_2 minimization. IEEE Trans Signal Process 68:2660–2669

Wang F, Mukherjee S, Richardson S, Hill S (2020) High-dimensional regression in practice: an empirical study of finite-sample prediction, variable selection and ranking. Stat Comput 30:697–719

Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. Annals Stat 38:894–942

Zou H (2006) The adaptive Lasso and its oracle properties. J Am Stat Assoc 101:1418–1429

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Royal Stat Soc: Series B 67:301–320

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.