Electronic Journal of Statistics
Vol. 18 (2024) 247–300
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2200

A penalised bootstrap estimation procedure for the explained Gini coefficient

Alexandre Jacquemain[1],
Cédric Heuchenne[2], **and Eugen Pircalabelu**[2]

[1] *ISBA, UCLouvain*
e-mail: alexandre.jacquemain@uclouvain.be; eugen.pircalabelu@uclouvain.be

[2] *HEC Liege, University of Liège*
e-mail: C.Heuchenne@ulg.ac.be

Abstract: The Lorenz regression estimates the explained Gini coefficient, a quantity with a natural application in the measurement of inequality of opportunity. Assuming a single-index model, it corresponds to the Gini coefficient of the conditional expectation of a response given some covariates and it can be estimated without having to estimate the link function. However, it is prone to overestimation when many covariates are included. In this paper, we propose a penalised bootstrap procedure which selects the relevant covariates and produces valid inference for the explained Gini coefficient. The obtained estimator achieves the Oracle property. Numerically, it is computed by the SCAD-FABS algorithm, an adaptation of the FABS algorithm to the SCAD penalty. The performance of the procedure is ensured by theoretical guarantees and assessed via Monte-Carlo simulations. Finally, a real data example is presented.

MSC2020 subject classifications: Primary 62G08, 62P20; secondary 62F40, 62E20.
Keywords and phrases: FABS algorithm, Gini coefficient, Lorenz regression, SCAD penalty, single-index models.

Received August 2022.

Contents

1. Introduction

The purpose of the Lorenz regression developed by [10], consists in estimating the explained Gini coefficient, measuring the inequality of an economic outcome Y which can be attributed to a set of covariates $X = (X^1, \dots, X^p)^\intercal$. We assume that $0 < E[Y] < \infty$, where $E[\cdot]$ is the expected value. The Gini coefficient of Y is defined as

$$\mathrm{Gi}_Y \coloneqq \frac{2C[Y, F_Y(Y)]}{E[Y]},$$

where $F_Y(\cdot)$ is the cumulative distribution of Y and $C[\cdot,\cdot]$ is the covariance between the random variables Y and $F_Y(Y)$. The Gini coefficient is a measure of the inequality of Y. In some applications, it might be interesting to measure the inequality of Y that is attributable to X. One way to formalize this idea is to consider the Gini coefficient of the conditional expectation of Y given X.

Throughout this paper, we assume the single-index model

$$E[Y|X = x] = H(x^\intercal \theta_0), \tag{1}$$

where H is a strictly increasing function and θ_0 is a vector of weights, normalized in order to ensure identifiability. The explained Gini coefficient is defined as

$$\begin{aligned} \mathrm{Gi}_{Y,X} &\coloneqq \max_\theta \frac{2C[Y, F_\theta(X^\intercal\theta)]}{E[Y]} && (2)\\ &= \frac{2C[H(X^\intercal\theta_0), F_H(H(X^\intercal\theta_0))]}{E[H(X^\intercal\theta_0)]}, && (3)\end{aligned}$$

where $F_\theta(\cdot)$ is the cumulative distribution function (CDF) of $X^\intercal\theta$ and $F_H(H(X^\intercal\theta_0))$ is the CDF of $H(X^\intercal\theta_0)$. In the economic literature, the objective function in (2) is called the concentration index of Y with respect to $X^\intercal\theta$. This representation opens the door to an estimation procedure that does not depend on $H(\cdot)$. Let $(X_i^\intercal, Y_i)_{i=1,\dots,n}^\intercal$ be an i.i.d sample sharing the same distribution as $(X^\intercal, Y)^\intercal$. The weight vector θ_0 and $\mathrm{Gi}_{Y,X}$ are consistently estimated with

$$\overline{\theta} \coloneqq \arg\max_{\theta} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} Y_i \, \mathbb{1}\{X_i^\intercal \theta \geq X_j^\intercal \theta\}, \tag{4}$$

$$\overline{\mathrm{Gi}}_{Y,X} \coloneqq \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{Y_i}{\overline{Y}} \, \mathbb{1}\{X_i^\intercal \overline{\theta} \geq X_j^\intercal \overline{\theta}\} - \frac{n+1}{n}, \tag{5}$$

where $\overline{Y}$ denotes the empirical mean of $(Y_1, \dots, Y_n)^\intercal$ and $\mathbb{1}\{\cdot\}$ denotes the indicator function. Equation (4) yields a special case of the monotone rank estimator introduced by [3] in the context of the single-index model. Equation (3) is interesting for interpretation purposes. It indicates that the explained Gini coefficient is the Gini coefficient of $H(X^\intercal \theta_0)$. The explained Gini coefficient therefore measures the inequality of the conditional expectation of Y given X, assuming a single-index model with strictly increasing link function $H(\cdot)$.

Facing large micro-datasets covering hundreds of covariates, the Lorenz regression may face several issues. First, the objective function in (4) is non-differentiable, which complicates the numerical solution. [10] proposed a genetic algorithm but its performance may be altered in datasets of large dimension. Another issue is overfitting. In large datasets we might expect some covariates to be irrelevant but, at the same time, to bear some empirical correlation with the response. Much like the R^2 in linear regression, the estimated explained Gini coefficient never decreases as one keeps introducing new covariates. Facing a sparse model, where some of the elements of θ_0 are equal to 0, an estimation of the full model will lead us to overestimate the explained Gini coefficient. These concerns echo the justifications underlying the use of penalised regressions. Indeed, these procedures provide proper inference of the model parameters and automatic selection of the covariates, even in a situation where the number of covariates is large. Procedures based on the SCAD penalty ensure a strong statistical guarantee: the oracle property of the estimator. We refer the reader to [6] for the original procedure and to [11] for an adaptation to the single-index model.

In this paper, we present a penalised Lorenz regression combined with a pairs bootstrap procedure. The proposed methodology presents several advantages. First, it makes it possible to include many covariates without inducing an overestimation of the explained Gini coefficient. Second, the penalisation leads to a selection of the relevant covariates. Finally, the pairs bootstrap allows one to construct confidence intervals for the explained Gini coefficient without having to estimate the link function $H(\cdot)$ of the single-index model, see Equation (1) for the definition of the model. The proposed procedure comes with statistical and numerical guarantees. Through the use of a SCAD penalty, the estimator achieves the oracle property. On the numerical side, we adapt the FABS algorithm proposed by [15] to a SCAD penalty. Hence, we benefit from a procedure enjoying good theoretical properties as well as a fast and efficient algorithm to obtain estimates. From the viewpoint of the estimation of a single-index model, the most obvious competitor is the penalised maximum smoothing rank correlation (PMSRC) estimator proposed by [11]. Compared to this procedure, the advantage of the penalised Lorenz regression consists in the fact that it exploits more information contained in the data. At the level of the response, it uses the observation values, and not only the ranks. It is therefore expected to provide a better tradeoff between flexibility and efficiency. This point is confirmed by the favourable simulation results displayed in Section 4.2.

The Lorenz regression methodology has a natural application in the estimation of inequality of opportunity (IOP). This concept embodies the idea that inequalities in an economic advantage, e.g. earnings, are unfair if and only if they are generated by variables over which individuals have no control, so-called circumstances. As pointed out in [14], the measurement of IOP is characterized by a two-stage nature. First, the advantage variable is fitted in an econometric model where the economic advantage is the response variable and circumstances are covariates. Second, the estimated model is used in combination with a measure of inequality in order to evaluate the extent of IOP. Blending these two stages harmoniously remains an issue in the literature. The first stage is often a log-linear regression, see for example [1] and [7]. A recent approach measures IOP with the Gini coefficient of fitted values obtained via machine learning methods, see for example [2]. Interestingly, this method produces an automatic selection of the relevant circumstances and does not rest on a restrictive parametric model. However, as discussed in [5], it is prone to lead to a biased estimation of IOP because of the absence of robustness of the Gini coefficient with respect to the derivation of the fitted values. Also, the method comes without an inferential procedure. [5] propose a debiased estimator and a valid inference procedure based on orthogonal moments. In practice, the debiased estimator boils down to the empirical concentration index of the economic advantage with respect to the fitted values. This finding unveils a new advantage of our procedure. In the assumed single-index model, the estimated explained Gini coefficient corresponds to the debiased estimator proposed by [5]. By assumption, $\mathbb{1}\{X_i^\intercal\overline{\theta} \geq X_j^\intercal\overline{\theta}\} = \mathbb{1}\{H(X_i^\intercal\overline{\theta}) \geq H(X_j^\intercal\overline{\theta})\}$. Hence (5) is an estimator of the concentration index of the economic advantage with respect to the fitted values, where the fitted value of observation i is $H(X_i^\intercal\overline{\theta})$.

This paper is organized as follows. In Section 2, we present the penalised bootstrap Lorenz regression and provide asymptotic results for the estimated covariate weights and for the explained Gini coefficient. The theoretical framework considered in this paper hinges on a series of conditions on the penalty function, which include the SCAD but exclude the LASSO. We also provide a small discussion of these two methods in our context. The numerical algorithms are presented in Section 3. We recall the main ideas underlying the FABS algorithm and present the SCAD-FABS algorithm more thoroughly, for which we derive convergence properties. Similarly to the FABS algorithm, we show that each solution along the SCAD-FABS path is a δ-approximate solution to the penalised programme. In practice, the FABS algorithm is used to fit the penalised bootstrap Lorenz regression with the LASSO penalty. The SCAD-FABS algorithm is used when the SCAD penalty is considered. Throughout this paper, we call these two methods PLR-LASSO and PLR-SCAD respectively. Simulation results assessing the performance of the procedure are displayed in Section 4. A comparison of the PLR-SCAD with the PLR-LASSO is first provided in Section 4.1. In Section 4.2, we provide a broader comparison with the PMSRC. Section 4.3 illustrates the asymptotic properties of the procedure and Section 4.4 evaluates the performance of the confidence intervals. We confront the method to real data in Section 5. A robustness analysis and an assessment of computing time are provided in Sections 5.2 and 5.3 respectively. Finally, a discussion on the method is provided in Section 6.

2. The penalised bootstrap Lorenz regression

This section develops as follows. First, we introduce the penalised Lorenz regression programme and provide asymptotic results for the estimated weight vector and explained Gini coefficient. Second, we turn to the bootstrap procedure and the choice of the regularisation parameter.

Given n i.i.d samples $(X_i^\intercal, Y_i)^\intercal_{i=1,\dots,n}$, the penalised Lorenz regression solves the following optimization programme

$$\hat{\theta} := \underset{\theta, \|\theta\|=1}{\arg\max} \left\{ G_n(\theta) - \sum_{k=1}^{p} p_\lambda(|\theta_k|) \right\}, \tag{6}$$

where $p_\lambda(\cdot)$ is a nonconcave penalty function, $\lambda > 0$ is a penalty parameter and $\|\cdot\|$ denotes the L2-norm of a vector. The non-penalised objective function is a smooth approximation of the objective function displayed in (5). It is given by

$$G_n(\theta) := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} Y_i K\left(\frac{X_i^\intercal \theta - X_j^\intercal \theta}{h} \right), \tag{7}$$

where $K(\cdot)$ is the integral of a kernel function and h is its corresponding bandwidth. In essence, (6) is an adaptation of the programme characterizing the PMSRC estimator. The difference lies in the original objective function. While the PMSRC is an adaptation of the maximum rank correlation estimator introduced by [9], (6) is related to the monotone rank estimator proposed by [3]. In both cases, the idea consists in combining smoothing techniques and penalisation with the double advantage of facilitating the estimation procedure and avoiding overfitting.

The theoretical results of this section hold for a class of penalty functions, satisfying a number of conditions. One member of this class is the SCAD penalty. It satisfies

$$p'_\lambda(x) = \begin{cases} \lambda & \text{if } x \leq \lambda \\ \frac{a\lambda - x}{a-1} & \text{if } \lambda < x \leq a\lambda \\ 0 & \text{if } x > a\lambda \end{cases} \tag{8}$$

where $x > 0$, $a > 2$ is an arbitrary constant and, for a differentiable function $f(\cdot)$, $f'(x) = df(x)/dx$. Without loss of generality, we order the covariates such that $\{1, \ldots, s\}$ are active and $\{s+1, \ldots, p\}$ are non-active. Let $\theta_0 = (\theta_0^{A\intercal}, \theta_0^{I\intercal})^\intercal$ be the vector of unknown weights. The vector θ_0^A is related to the s active covariates, while $\theta_0^I = 0$ corresponds to the non-active part. In Theorem 2.2, it will also be useful to write $\vartheta_0 = (\theta_{0,1}, \ldots, \theta_{0,s-1})^\intercal$, i.e. $\theta_0^A = (\vartheta_0^\intercal, \theta_{0,s})^\intercal$.

Call $f(\cdot)$ the joint density of X and $g(\cdot)$ the density of $Z = X^\intercal\theta_0$. Also, denote by $F_{Y,X}(\cdot)$ and $F_{Y,Z}(\cdot)$ the joint distribution functions of $(Y, X^\intercal)^\intercal$ and $(Y, Z)^\intercal$ respectively. We will use the following regularity conditions in order to prove the main results.

- **(RC1)** X is bounded with compact support $\mathcal{X}$. Also, we denote the compact support of Z by $\mathcal{Z}$.
- **(RC2)** $H(X^\intercal\theta_0)$ has finite first and second moment and $H(\cdot)$ is twice continuously differentiable.
- **(RC3)** $g(\cdot)$ is twice continuously differentiable. Also, for each $l = 1, \ldots, p$ and $m < l$, the joint density of $(X^l, Z)^\intercal$ is four-times continuously differentiable and the joint density of $(X^l, X^m, Z)^\intercal$ is three-times continuously differentiable.
- **(RC4)** $\kappa(\cdot)$ is a density function with compact support $[-1, 1]$ such that $\kappa(-1) = \kappa(1) = 0$, and satisfying $\int v^l\kappa(v)dv = 0$ for $l = 1, 2$ and $\int v^q\kappa(v)dv \neq 0$ for some $q \geq 3$. Also, $\kappa(\cdot)$ is twice continuously differentiable. Denote by $K(\cdot)$ the CDF related to $\kappa(\cdot)$.
- **(RC5)** $nh^6 \to 0$ and $nh^5(\log(h^{-1}))^{-1} \to \infty$.
- **(RC6)** $p_\lambda(\cdot)$ has piecewise continuous second derivatives and bounded third derivatives.
- **(RC7)** $E[YF_\theta(X^\intercal\theta)]$ has a unique maximum with respect to θ. Also, the second derivative of the function $\theta \mapsto F_\theta(x^\intercal\theta)$ is continuous with respect to x and θ.

(RC1)–(RC5) are standard conditions to obtain strong consistency results for a kernel density estimator and its first and second derivatives, and for the numerator of a Nadaraya-Watson estimator, as in the works of [16], [13] and [12]. (RC6) is necessary to perform Taylor expansions of the objective function displayed in (6) and is satisfied for the SCAD. Finally, (RC7) is needed to ensure the local consistency of $\hat{\theta}$. The first part of this assumption is proven under mild conditions in Theorem 1 from [3].

Theorems 2.1 and 2.2 are adaptations of Theorems 1 and 2 from [11] to our context. Their proofs are deferred to Appendix A.

Theorem 2.1. *Let $(X_i^\intercal, Y_i)^\intercal_{i=1,\ldots,n}$ be i.i.d random vectors satisfying Equation (1) with $H(\cdot)$ strictly increasing, $\|\theta_0\| = 1$, and $E[Y_i^2] < \infty$. If $\max_k\{p''_\lambda(|\theta_{0,k}|)\} \to 0$ and (RC1)–(RC7) hold, then there exists a local maximizer $\hat{\theta}$ of Equation (6) such that*

$$\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2} + a_n),$$

where $a_n := \max\{p'_\lambda(|\theta_{0,k}|) : \theta_{0,k} \neq 0\}$.

With the SCAD penalty, $a_n = 0$ with a proper choice of λ. This allows us to obtain a $\sqrt{n}$-consistent estimator for the coefficient vector θ_0. This is an advantage over the LASSO penalty, where the convergence rate is $O_p(n^{-1/2}+\lambda)$. Define next

$$b := [p'_\lambda(|\theta_{0,1}|)\text{sign}(\theta_{0,1}), \dots, p'_\lambda(|\theta_{0,s-1}|)\text{sign}(\theta_{0,s-1})]^\intercal - \frac{p'_\lambda(\theta_{0,s})}{\theta_{0,s}}\vartheta_0. \quad (9)$$

Theorem 2.2. *Let $(X_i^\intercal, Y_i)^\intercal_{i=1,\dots,n}$ be i.i.d random vectors satisfying Equation (1) with $H(\cdot)$ strictly increasing, $\|\theta_0\| = 1$, $\theta_{0,s} > 0$ and $E[Y_i^2] < \infty$. Assume that $a_n = 0$ and*

$$\liminf_{n\to\infty} \liminf_{x\to 0^+} p'_\lambda(x)/\lambda > 0.$$

If $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ as $n \to \infty$, and (RC1)–(RC7) hold, then the $\sqrt{n}$-consistent local maximizers $\hat{\theta} = (\hat{\theta}^{A\intercal}, \hat{\theta}^{I\intercal})^\intercal$ of Theorem 2.1, where $\hat{\theta}^A = (\hat{\vartheta}^\intercal, \hat{\theta}_s)^\intercal$ estimates θ_0^A and $\hat{\theta}^I$ estimates θ_0^I, satisfy

1. *Sparsity: $P(\hat{\theta}^I = 0) \to 1$, as $n \to \infty$.*
2. *Asymptotic normality: $\sqrt{n}[\hat{\vartheta} - \vartheta_0 + \Sigma^{-1}b] \xrightarrow{d} N(0, \Sigma^{-1}\Omega\Sigma^{-1\intercal})$, where Σ is an invertible $(s-1)\times(s-1)$ matrix defined in (29), b is a vector of dimension $s-1$ defined in (9) and Ω is an $(s-1)\times(s-1)$ matrix defined in (30).*

With the SCAD penalty, as $\lambda \to 0$ and in light of Equation (8), each component of the vector b is set to 0. In that case, the penalised Lorenz regression enjoys two major properties. The first is the property of sparsity, which indicates that the estimated weights attached to the non-active covariates are set to zero with a probability tending to one. The second is the asymptotic normality of the estimated weight vector related to the active covariates. The convergence rate is the same as what one would obtain with the monotone rank estimator of [3] computed on the active set of covariates. The selection process does not lead to any loss of efficiency. Hence, our estimator enjoys the oracle property. Finally, notice that due to the constraint $\|\hat{\theta}\| = 1$, the last active covariate can be viewed as a function of the others, namely $\hat{\theta}_s = \sqrt{1 - \hat{\vartheta}^\intercal\hat{\vartheta}}$. As a consequence, the asymptotic normality is obtained on $\hat{\vartheta}$ rather than on $\hat{\theta}^A$.

An estimator $\widehat{\text{Gi}}_{Y,X}$ for the explained Gini coefficient is obtained by plugging $\hat{\theta}$ into Equation (5), where $\hat{\theta}$ is the $\sqrt{n}$-consistent estimator of Theorem 2.2. Theorem 2.3 establishes the asymptotic normality of $\widehat{\text{Gi}}_{Y,X}$. Its proof is deferred to Appendix A.

Theorem 2.3. *Let $(X_i^\intercal, Y_i)^\intercal_{i=1,\dots,n}$ be i.i.d random vectors with $0 < E[Y_i] < \infty$. If the conditions of Theorem 2.2 hold, then $\sqrt{n}[\widehat{\text{Gi}}_{Y,X} - \text{Gi}_{Y,X}] \xrightarrow{d} N(0, \sigma_\zeta^2)$, where $\sigma_\zeta^2 := V[\zeta_i]$ and ζ_i is defined in Equation (31).*

As in Theorem 2.2, we achieve the same rate of convergence as an unpenalised procedure undertaken on the set of active covariates. In the penalised procedure, the asymptotic normality and unbiasedness comes from the estimated index being itself asymptotically normal and unbiased. This, in turn, stems from the use of an unbiased penalty function, i.e. the SCAD. In this construction, an alternative would be to use the debiased LASSO introduced by [20]. However, a more general comment is worth making. From [5], it appears that we could accommodate for some bias in the estimation of the index. This is due to the construction of $\widehat{\mathrm{Gi}}_{Y,X}$, that uses the index only through its ordering structure. Following this route, it would be possible to relax the requirements on the estimation of θ_0. Interestingly, one of the conditions on the estimated index would be that the sign of the difference in the estimated index converges in probability to the sign of the difference in the true index. Thanks to Theorem 2.1, it is easy to show that this is satisfied, even for the LASSO. In light of this, the LASSO is another candidate to provide an unbiased estimation of the explained Gini coefficient.

Even though the variance is difficult to estimate in practice, the asymptotic normality of the estimated explained Gini coefficient opens the door to a hybrid bootstrap procedure where the quantiles of the normal distribution are used and the variance is estimated via bootstrap. Section 4.3 evaluates the asymptotic normality of the estimated explained Gini coefficient for several sample sizes, while section 4.4 compares different bootstrap procedures. To anticipate slightly the discussion, the conclusion will be that the hybrid bootstrap produces the best performance.

The proposed bootstrap procedure is described in Algorithm 1. In a nutshell, the idea consists in constructing training bootstrap resamples $(X^{*\intercal}, Y^*)^\intercal$ drawn with replacement from the pairs $(X_i^\intercal, Y_i)^\intercal$, and validation bootstrap samples $(\tilde{X}^{*\intercal}, \tilde{Y}^*)^\intercal$ corresponding to the out-of-bag (OOB) samples, i.e. the data unused in the construction of the training samples. As we know from bootstrap aggregating techniques, the size of the validation samples will be approximately of $\frac{n}{e}$, where e is Euler's number. The training samples are then used to obtain samples of bootstrap estimators for θ_{Gi} and for $\mathrm{Gi}_{Y,X}$ on a grid of λ values. The validation samples can then be used to determine an optimal λ. At the end of the procedure, one has at disposal samples of bootstrap estimates $\widehat{\mathrm{Gi}}^*(\lambda)$ that one can use to produce confidence intervals.

The choice of the regularisation parameter λ is of great importance. An optimal value should strike the balance between under- and over-penalisation and, hence, between underfit and overfit of the explained Gini coefficient. As we have just stated, the bootstrap procedure offers a first method to select it: λ^{OOB} is the value of λ which performs best in terms of explained Gini coefficient in the out-of-bag resamples. Another possibility lies in the use of a BIC-like criterion, as proposed in [11]. Formally, λ^{BIC} is obtained as the value of λ which maximizes

$$\mathrm{BIC}_\lambda := \log(\widehat{\mathrm{Gi}}_{Y,X;\lambda}) - k_\lambda \frac{\log(n)}{2n},$$

where k_λ is the number of covariates selected using λ. Also, we write $\widehat{\mathrm{Gi}}_{Y,X;\lambda}$ in order to acknowledge the dependence of the estimated explained Gini coefficient on λ through $\hat{\theta}$. As the simulation results from Section 4 and the real

Algorithm 1: Bootstrap procedure

```
Data: $(\mathbf{X}^\intercal, \mathbf{Y})^\intercal \in \mathbb{R}^{n\times(p+1)}$, where $\mathbf{X}$ denotes the matrix of covariates and $\mathbf{Y}$ is the
      response vector.
for $b = 1$ to $B$ do
    Generate the training bootstrap sample $(\mathbf{X}_b^{*\intercal}, \mathbf{Y}_b^*)^\intercal \in \mathbb{R}^{n\times(p+1)}$
    Obtain $\hat{\theta}_b^*(\lambda)$ by solving Equation (6) on $(\mathbf{X}_b^{*\intercal}, \mathbf{Y}_b^*)^\intercal$ via the SCAD-FABS
      algorithm from Section 3, using a $\lambda$ sequence;
    Plug $\hat{\theta}_b^*(\lambda)$ and $(\mathbf{X}_b^{*\intercal}, \mathbf{Y}_b^*)^\intercal$ into Equation (5) and denote the estimator $\widehat{\mathrm{Gi}}_b^*(\lambda)$;
    Generate the validation bootstrap sample $(\tilde{\mathbf{X}}_b^{*\intercal}, \tilde{\mathbf{Y}}_b^*)^\intercal \in \mathbb{R}^{\tilde{n}_b\times(p+1)}$;
    Plug $\hat{\theta}_b^*(\lambda)$ and $(\tilde{\mathbf{X}}_b^{*\intercal}, \tilde{\mathbf{Y}}_b^*)^\intercal$ into Equation (5) and denote the estimator $\widetilde{\mathrm{Gi}}_b^*(\lambda)$
end
Obtain $\lambda^{\mathrm{OOB}}$ as the solution to $\max_\lambda \text{OOB-score}(\lambda) \coloneqq \frac{1}{B}\sum_{b=1}^B \widetilde{\mathrm{Gi}}_b^*(\lambda)$;
For each $\lambda$, retrieve $\widehat{\mathrm{Gi}}^*(\lambda) \coloneqq (\widehat{\mathrm{Gi}}_1^*(\lambda), \dots, \widehat{\mathrm{Gi}}_B^*(\lambda))^\intercal \in \mathbb{R}^B$
```

data example from Section 5 indicate, λ^{OOB} tends to select fuller models, while λ^{BIC} favours sparser ones. In practice, the user might decide between these two options by balancing the OOB-scores attained at λ^{OOB} and λ^{BIC} with the simplicity of the models that they induce.

3. The SCAD-FABS algorithm

The FABS has been developed by [15] to solve penalised problems with differentiable but typically non-convex loss functions and the adaptive LASSO penalty. Its fundamental logic resembles the coordinate-descent algorithm of [8]. Indeed, it starts with a very large value for λ, imposing full sparsity, and then allows at each step the penalty parameter to relax. Hence, it produces a whole solution path ranging from high to low sparsity. For each λ, the optimization problem is solved using a coordinate-descent algorithm. To facilitate the comparison, we review the construction of the FABS algorithm in a pure LASSO setting. Then, we introduce the SCAD-FABS algorithm and present its convergence properties.

3.1. The FABS algorithm

For a grid of penalty parameters, the FABS solves

$$\min_\theta Q(\theta) \coloneqq L(\theta) + \sum_{k=1}^p p_\lambda(|\theta_k|), \tag{10}$$

where $Q(\cdot)$ is the objective function and $L(\cdot)$ is a general loss function. We focus here on a pure LASSO setting, where $p'_\lambda(x) = \lambda$. At each iteration, only one coefficient is updated by a fixed amount. A backward step is undertaken if this operation reduces the value of the objective function. Otherwise, the iteration consists in a forward step. For a given index k, the update is of the form

$$\theta^{t+1} = \theta^t - \mathrm{sign}(\theta_k^t)\mathbf{1}_k\epsilon \qquad \text{(backward step)} \tag{11}$$

$$\theta^{t+1} = \theta^t - \text{sign}\left(\nabla_k L(\theta^t)\right) \mathbf{1}_k \epsilon \qquad \text{(forward step)} \qquad (12)$$

where θ^t (θ^{t+1}) represents the vector of coefficients at iteration t ($t+1$), $\mathbf{1}_k$ denotes the vector of size p taking value 1 at the kth component and 0 everywhere else, ϵ is the step size of the algorithm and ∇ is the gradient vector. The FABS uses a first-order Taylor expansion of $L(\theta^{t+1})$ around θ^t to determine which coordinate should be updated. Formally, the coordinate in $t+1$ is determined as follows

$$\begin{aligned} k &= \underset{l \in \mathcal{A}^t}{\arg\min} \left\{ -\nabla_l L(\theta^t) \text{sign}(\theta^t_l) \right\} && \text{(backward step)} \\ k &= \underset{l=1,\dots,p}{\arg\max} \, |\nabla_l L(\theta^t)| && \text{(forward step)} \end{aligned}$$

where $\mathcal{A}^t \coloneqq \{k \in \{1,\dots,p\} : \theta^t_k \neq 0\}$. Finally, λ^t is updated as follows

$$\begin{aligned} \lambda^{t+1} &= \lambda^t && \text{(backward step)} \\ \lambda^{t+1} &= \min\{\lambda^t, L^{t,t+1}_\epsilon\} && \text{(forward step)} \qquad (13) \end{aligned}$$

where $L^{t,t+1}_\epsilon \coloneqq \frac{L(\theta^t) - L(\theta^{t+1})}{\epsilon}$. The update rule for λ^t generates a path of values for the regularisation parameter that is a decreasing step function. For a given value, the algorithm minimizes the objective function by searching for the best direction. However, at some iteration, the objective function can no longer be decreased with this value of the regularisation parameter. Hence, λ^t needs to make a jump, and the process is reiterated to search for the best direction that minimizes the objective function with this new value of the regularisation parameter. Therefore, (13) ensures that an update occurs only when the objective function can no longer be improved using λ^t. Then, λ^{t+1} is chosen such that $Q(\theta^{t+1}, \lambda^{t+1}) = Q(\theta^t, \lambda^{t+1})$.

3.2. The SCAD-FABS algorithm

The SCAD-FABS solves the minimization problem displayed in (10) using the SCAD penalty function. The vector θ^{t+1} is obtained according to Equations (11) and (12). The form of the forward and backward steps is therefore the same as in the FABS. Also, the index k is chosen optimally, based on Taylor expansions. In a backward step, a Taylor expansion of the objective function around θ^t gives

$$Q(\theta^{t+1}) = Q(\theta^t) - \nabla_k Q(\theta^t) \text{sign}(\theta^t_k)\epsilon + O(\epsilon^2).$$

The index of the updated coefficient at iteration $t+1$ is then given by

$$k = \underset{l \in \mathcal{A}^t}{\arg\min} \left\{ -\nabla_l Q(\theta^t) \text{sign}(\theta^t_l) \right\}.$$

In a forward step, the Taylor expansion gives

$$Q(\theta^{t+1}) = Q(\theta^t) - |\nabla_k L(\theta^t)|\epsilon + p'_\lambda(|\theta^t_k|)\epsilon + O(\epsilon^2),$$

where $p'_\lambda(|\theta_k^t|)$ is defined according to Equation (8) for $|\theta_k^t| > 0$ and $p'_\lambda(|\theta_k^t|)$ is set to λ if $\theta_k^t = 0$. For $k \notin \mathcal{A}^t$, the result is obtained with a Taylor expansion of the loss while, for $k \in \mathcal{A}^t$, the result is obtained using a Taylor expansion of the loss and of the penalty, combined with Lemma 3.1. The index of the updated coefficient is obtained as

$$k = \underset{l=1,\dots,p}{\arg\max} \left\{ |\nabla_l L(\theta^t)| - p'_\lambda(|\theta_l^t|) \right\}.$$

Note that the choice of the direction differs from the FABS. The reason is the following. Since the derivative of the LASSO penalty is constant, coupled with the fact that only one coefficient is updated by a fixed amount, the penalty part does not play a role in the Taylor expansion of the FABS in Section 3.1. Using a SCAD penalty, this no longer applies and the whole objective function must be considered.

The update rule for λ^t follows the same spirit as in the original FABS. An update is conducted when the objective function can no longer be improved using λ^t, i.e. $Q(\theta^{t+1}, \lambda^t) > Q(\theta^t, \lambda^t)$. Consider a forward update on coefficient k. Using a Taylor expansion of $p_{\lambda^t}(|\theta_k^{t+1}|)$ around $|\theta_k^t|$, it approximately holds

$$Q(\theta^{t+1}, \lambda^t) > Q(\theta^t, \lambda^t) \Leftrightarrow L(\theta^t) - L(\theta^{t+1}) < p'_{\lambda^t}(|\theta_k^t|)\epsilon.$$

This occurs when

$$\begin{cases} \lambda^t > L_\epsilon^{t,t+1} & \text{if } |\theta_k^t| \le \lambda^t \\ \lambda^t > \frac{1}{a}\left[(a-1)L_\epsilon^{t,t+1} + |\theta_k^t|\right] & \text{if } \lambda^t < |\theta_k^t| \le a\lambda^t \\ L_\epsilon^{t,t+1} < 0 & \text{if } |\theta_k^t| > a\lambda^t. \end{cases} \tag{14}$$

Notice that, in the last case, λ^t plays no role since $p'_{\lambda^t}(|\theta_k^t|) = 0$ and, hence, should not be updated. We now focus on the form of the update. In the event of an update, λ^{t+1} is chosen to ensure approximately $Q(\theta^{t+1}, \lambda^{t+1}) = Q(\theta^t, \lambda^{t+1})$, which happens whenever

$$\lambda^{t+1} = \lambda_A^{t+1} := L_\epsilon^{t,t+1} \qquad \text{if } |\theta_k^t| \le \lambda^{t+1} \tag{15}$$

$$= \lambda_B^{t+1} := \frac{1}{a}\left[(a-1)L_\epsilon^{t,t+1} + |\theta_k^t|\right] \qquad \text{if } \lambda^{t+1} < |\theta_k^t| \le a\lambda^{t+1}. \tag{16}$$

Using Equation (15), we choose $\lambda^{t+1} = \lambda_A^{t+1}$ if $|\theta_k^t| \le \lambda_A^{t+1}$. Using Equations (15) and (16), we choose $\lambda^{t+1} = \lambda_B^{t+1}$ if $|\theta_k^t| > \lambda_A^{t+1}$ and $\lambda_B^{t+1} \le |\theta_k^t| \le a\lambda_B^{t+1}$. It is easy to prove that this boils down to choosing $\lambda^{t+1} = \max\{\lambda_A^{t+1}, \lambda_B^{t+1}\}$. The situation $|\theta_k^t| > a\lambda_B^{t+1}$ implies $L_\epsilon^{t,t+1} < 0$ and can be disregarded due to the loss improvement check introduced in the SCAD-FABS algorithm, presented in Algorithm 2. Taking our last results together with Equation (14), the update rule for λ^t becomes

$$\lambda^{t+1} = \begin{cases} \min\left(\max\{\lambda_A^{t+1}, \lambda_B^{t+1}\}, \lambda^t\right) & \text{if } |\theta_k^t| \le a\lambda^t \\ \lambda^t & \text{otherwise.} \end{cases} \tag{17}$$

```
Algorithm 2: The SCAD-FABS algorithm
  Data: $(\mathbf{X}^\intercal, \mathbf{Y})^\intercal \in \mathbb{R}^{n\times(p+1)}$
  Initialization: start from the empty solution and compute
      $k = \arg\max_{l=1,\dots,p} |\nabla_l L(0)|; \mathcal{A}^0 = \{k\}$
      $\theta^0 = -\text{sign}(\nabla_k L(0))\mathbf{1}_k\epsilon$
      $\lambda^0 = \frac{1}{\epsilon}[L(0) - L(\theta^0)]$
  Backward step: for each iteration $t$, compute
      $k = \arg\min_{l\in\mathcal{A}^t} \left\{-\nabla_l Q(\theta^t)\text{sign}(\theta_l^t)\right\}$
      $\Delta_k = -\text{sign}(\theta_k^t)\mathbf{1}_k$
  If $L(\theta^t + \Delta_k\epsilon) - L(\theta^t) - \epsilon p'_{\lambda^t}(|\theta_k^t|) < 0$, take a backward step. Then
      $\theta^{t+1} = \theta^t + \Delta_k\epsilon$
      $\lambda^{t+1} = \lambda^t$
  Otherwise, take a forward step.
  Forward step: set $\mathcal{B}^t = \{1,\dots,p\}$ and compute
      $k = \arg\max_{l\in\mathcal{B}^t} \left\{|\nabla_l L(\theta^t)| - p'_\lambda(|\theta_l^t|)\right\}$
  If $L\left(\theta^t - \text{sign}\left(\nabla_k L(\theta^t)\right)\mathbf{1}_k\epsilon\right) > L(\theta^t)$, set $\mathcal{B}^t = \mathcal{B}^t\backslash\{k\}$ and return to the
    computation of the index $k$ to be updated. Otherwise, set
      $\theta^{t+1} = \theta^t - \text{sign}\left(\nabla_k L(\theta^t)\right)\mathbf{1}_k\epsilon$
      $\lambda^{t+1} = \begin{cases} \min\left(\max\{\lambda_A^{t+1}, \lambda_B^{t+1}\}, \lambda^t\right) & \text{if } |\theta_k^t| \le a\lambda^t \\ \lambda^t & \text{otherwise.} \end{cases}$
  Stopping rule: Update $t \to t+1$, repeat the backward and forward
    steps and stop when $\lambda^{t+1} \le 0$ or $\mathcal{B}^t = \emptyset$. The final iteration is then
    given by $T = t$.
```

Notice the presence of a loss improvement check in the forward step. This step is unnecessary in the classical FABS algorithm as proposed by [15]. Indeed, an increase in the loss would imply $L_\epsilon^{t,t+1} < 0$, which would cause $\lambda^{t+1} < 0$ and the algorithm would stop. In the SCAD-FABS, updates on coefficients whose amplitude exceeds $a\lambda^t$ do not yield updates on λ^t. Hence, one needs to ensure that these coefficients are not updated if they yield an increase in the loss.

We use the FABS and the SCAD-FABS algorithms in order to solve (6). We use the notations PLR-LASSO and PLR-SCAD to refer to the two methods respectively. The loss function is given by $L(\theta) = -G_n(\theta)$. As suggested by [6], we set $a = 3.7$. For the PLR-LASSO, we use $h = n^{-1/5.5}$ and $\epsilon = 0.01$. For the PLR-SCAD, the bandwidth and step size are given by $h^* = ch$ and $\epsilon^* = c\epsilon$, where $c > 0$ is a constant. With a simple adaptation of Theorem 2 in [15], it is easy to show that the solution path of the PLR-LASSO would not be influenced by the choice of c. However, this is not the case for the PLR-SCAD. As we explain below, the choice of c can therefore be used to highlight the algorithmic differences between the two methods.

The difference between the FABS and the SCAD-FABS algorithms stems from the fact that they entail different marginal penalty rates. Take any coefficient θ_k^t, the marginal penalty rate in the FABS is defined as $p'_{\lambda_t}(|\theta_k^t|) = \lambda^t$. In the SCAD-FABS, it is given by

$$p'_{\lambda^t}(|\theta_k^t|) = \begin{cases} \lambda^t & \text{if } |\theta_k^t| \leq \lambda^t & \text{(region 1)} \\ \frac{a\lambda^t - |\theta_k^t|}{a-1} & \text{if } \lambda^t < |\theta_k^t| \leq a\lambda^t & \text{(region 2)} \\ 0 & \text{if } |\theta_k^t| > a\lambda^t & \text{(region 3).} \end{cases}$$

At the first iteration, both algorithms coincide since, for all $k = 1, \dots, p$, one has $\theta_k^0 = 0$ and $p'_{\lambda^0}(|\theta_k^0|) = \lambda^0$. This stays true as long as all coefficients lie in region 1. To put this differently, the difference between the SCAD-FABS and the FABS kicks in as soon as a first coefficient enters region 2. Depending on the choice of bandwidth and step size, this may happen close to the beginning or to the end of the algorithm. To illustrate this, consider a generic coefficient at the first iteration $\theta_k^0 = 0$ and set $\epsilon^* = c\epsilon$ and $h^* = ch$. One can ask the question: keeping $\lambda^t = \lambda^0$ constant, how many consecutive forward updates n_{fwd} would it take on θ_k^t in order to reach the second region? Recall that $\lambda^0 = (L(0) - L(\theta^0))/\epsilon^*$. It is easy to show that the loss function $L(\cdot)$ is independent on the choice of c. Also, the magnitude of θ_k^t after n_{fwd} forward updates is equal to $n_{\text{fwd}}\epsilon^*$. The coefficient θ_k^t enters the second region as soon as $|\theta_k^t| > \lambda^0$, which happens if

$$c > \frac{1}{\epsilon}\sqrt{\frac{L(0) - L(\theta^0)}{n_{\text{fwd}}}}. \tag{18}$$

The smaller the value of n_{fwd}, the sooner the PLR-SCAD and PLR-LASSO paths will differ. In the applications, we choose a grid of values of n_{fwd}. For each value on the grid, we choose

$$c = \frac{1}{\epsilon}\left[\sqrt{\frac{L(0) - L(\theta^0)}{n_{\text{fwd}}}} + \varrho\right],$$

where ϱ is a small positive constant based on machine precision. For a fixed value of n_{fwd}, notice that the couple (ϵ^*, h^*) is unaffected by a multiplication of ϵ and h by the same constant. We therefore recover the property of the FABS that the path is determined only by the ratio between ϵ and h. In conclusion, the FABS and SCAD-FABS algorithms depend on a common parameter, i.e. the ratio between ϵ and h. The SCAD-FABS is more general than the FABS, as it depends on a further tuning parameter materialized by n_{fwd}. With a large enough value of n_{fwd}, the paths obtained by both algorithms coincide, as illustrated in Section 4.1.

3.3. Properties of the SCAD-FABS algorithm

Throughout this section, we assume that $L(\cdot)$ has bounded second-order derivatives. The proofs of the following results are deferred to Appendix B. Lemma 3.1 indicates that a forward step always increases the amplitude of the updated coefficient. As long as no backward step is taken, the solution path is therefore monotone.

Lemma 3.1. *Let $k \in \mathcal{A}^t$ be updated via a forward step. Then, $\theta_k^{t+1} = \theta_k^t + \operatorname{sign}(\theta_k^t)\epsilon$, i.e. $\operatorname{sign}(\nabla_k L(\theta^t)) = -\operatorname{sign}(\theta_k^t)$.*

The SCAD-FABS exploits the differentiability of the loss and penalty functions through Taylor expansions. As such, it produces approximation errors. Lemmas B.2, B.3 and B.4, presented in Appendix B, show that the approximation error is of order $O(\epsilon^2)$, where we remind the reader that ϵ is the step size of the algorithm. Proposition 3.2 implies that the objective function never increases from t to $t+1$, up to an approximation error of order ϵ^2. This stems from the fact that λ^t is updated whenever it becomes impossible to improve the score further with that value of the regularisation parameter.

Proposition 3.2. *The SCAD-FABS algorithm ensures $Q(\theta^{t+1}, \lambda^{t+1}) \leq Q(\theta^t, \lambda^t)$. More precisely, if k is the index of the updated coefficient, then it holds*

$$\begin{aligned} Q(\theta^{t+1}, \lambda^{t+1}) &< Q(\theta^t, \lambda^t) - \frac{d_k^t \epsilon^2}{2(a-1)} && \textit{(backward step)} \\ Q(\theta^{t+1}, \lambda^{t+1}) &\leq Q(\theta^t, \lambda^t) - \frac{c_k^t \epsilon^2}{2(a-1)}, && \textit{(forward step)} \end{aligned}$$

where d_k^t and $c_k^t \in [0,1]$.

In the remaining of this section, we show that the SCAD-FABS enjoys the same δ-optimality property as the FABS. A candidate is called a δ-approximate solution if it satisfies the KKT conditions associated to the constrained optimization problem, up to a tolerance δ. We adapt the definition introduced by [15] to the SCAD penalty.

Definition 3.3. The parameter vector $\theta = (\theta_1, \dots, \theta_p)^\intercal$ is called a δ-approximate solution with regularisation parameter λ if the following two conditions are met

$$\begin{aligned} |\nabla_l Q(\theta, \lambda)| &\leq \delta && \text{if } \theta_l \neq 0 \\ |\nabla_l L(\theta)| &\leq p'_\lambda(|\theta_l|) + \delta && \text{if } \theta_l = 0. \end{aligned}$$

In what follows, the path refers to the collection of iterations $\{t = 1, \dots, T : \lambda^{t+1} < \lambda^t\}$, where T is the last iteration.

Theorem 3.4. *Every solution θ^t, with $t = 1, \dots, T$, along the SCAD-FABS path is a δ-approximate solution with regularisation parameter λ^t and $\delta = m\epsilon$, where m is the upper bound of the second-order derivatives of $L(\cdot)$.*

Theorem 3.4 shows that any point along the SCAD-FABS path is a δ-approximate solution, with δ proportional to the step size ϵ. The SCAD-FABS algorithm enjoys therefore the same optimality condition as the FABS. As $\epsilon \to 0$, the KKT conditions are recovered and each point along the path converges to a stationary point of (10).

4. Monte-Carlo simulations

In this section, we evaluate the performance of the penalised Lorenz regression combined with the proposed bootstrap procedure by means of Monte-Carlo simulations. As a first step, we focus on the quality of the estimation of the weight vector and of the explained Gini coefficient, and on the performance of model selection. In Section 4.1, we compare the PLR-SCAD with the PLR-LASSO and illustrate the convergence of the former to the latter when n_{fwd} increases. Section 4.2 provides a more thorough comparison, including the PMSRC. We evaluate the impact of changes in the explained Gini coefficient and in the sample size. In Section 4.3, we evaluate the consistency and asymptotic normality of the estimated explained Gini coefficient. Finally, we turn to the coverage of the confidence intervals in Section 4.4.

Throughout the simulations, we will use the following data generating process (DGP)

$$Y_i = \mathcal{Q}(F_{\theta_0}(X_i^\intercal \theta_0))\epsilon_i,$$

where $i = 1\dots, n$ and $\|\theta_0\| = 1$. $\mathcal{Q}(\cdot)$ is the quantile function of the lognormal distribution, with parameters tailored to ensure an expected value of 2400 and an explained Gini coefficient of 0.15. This choice is justified by the popular use of the lognormal distribution in empirical work concerning income distributions, see [4]. Two scenarios are considered for the distribution of X. In a first case, X follows a multivariate normal distribution with mean 0, unit variance and a correlation matrix following an AR(1) process with correlation parameter $\rho = 0.3$. In a second case, X follows a multivariate Student distribution with 3 degrees of freedom, mean 0, variance of 3 and the same correlation matrix as before. The variable ϵ_i is a lognormal noise with mean 1 and a variance set to ensure that $V[Y_i]/V[H(X_i^\intercal \theta_0)] = 3/2$. Concerning the kernel, we use

$$K(u) = \begin{cases} 0 & \text{if } u < -1 \\ \frac{9}{8}u - \frac{5}{8}u^3 + \frac{1}{2} & \text{if } u \in [-1, 1] \\ 1 & \text{if } u > 1, \end{cases}$$

which corresponds to a fourth-order kernel constructed from an Epanechnikov kernel, and which matches the conditions required by the theory. As mentioned previously, we set $h = n^{-1/5.5}$, $\epsilon = 0.01$ and use the grid $(5, 20, 50, 1000)$ for the values of n_{fwd}. In Section 4.1, we display the results of the whole grid in order to provide a comparison between the PLR-LASSO and PLR-SCAD methods. In Sections 4.2 to 4.4, the constant n_{fwd} is chosen from the grid as the value that maximizes the OOB-score (bootstrap procedure) or that maximizes the BIC criterion (BIC procedure).

We consider two setups. The first is low-dimensional with $n = 100$ and $p = 20$ (Setup 1) while the second is high-dimensional with $n = 100$ and $p = 120$ (Setup 2). In both cases, $s = 5$ of the covariates are active and the weight vector is $\theta_0 = (-\frac{\sqrt{3}}{5}, \frac{3}{5}, 0, 0, -\frac{\sqrt{7}}{5}, \frac{1}{5}, \frac{\sqrt{5}}{5}, 0^\intercal)^\intercal$, where $0^\intercal$ is a vector of zeroes of size 13 in Setup 1 and of size 113 in Setup 2. Unless specified otherwise, we sample $M = 400$ different datasets from the proposed DGPs and, for each simulation run, $B = 400$ bootstrap resamples are used.

The accuracy of the estimation is evaluated by the square-root of the empirical mean squared error (MSE) of the explained Gini coefficient, defined as

$$\text{RMSE.Gini} := \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\widehat{\text{Gi}}_{Y,X;m} - \text{Gi}_{Y,X})^2},$$

where $\widehat{\text{Gi}}_{Y,X;m}$ is the estimated explained Gini coefficient in simulation run m and $\text{Gi}_{Y,X}$ is the true value. We also compute the empirical mean of the L2-distance between the estimated weight vector and the true weight vector, i.e.

$$\text{Distance.}\theta := \frac{1}{M}\sum_{m=1}^{M}\|\hat{\theta}_m - \theta_0\|,$$

where $\hat{\theta}_m$ is the estimated weight vector in simulation run m. In order to assess the performance of the model selection, we compute the false positive rate (FPR) and false negative rate (FNR), defined as

$$\text{FPR} = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{p-s}\sum_{k=1}^{p}\mathbb{1}\{\hat{\theta}_{m,k} \neq 0, \theta_{0,k} = 0\}\right)$$
$$\text{FNR} = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{s}\sum_{k=1}^{p}\mathbb{1}\{\hat{\theta}_{m,k} = 0, \theta_{0,k} \neq 0\}\right).$$

4.1. Comparison of the PLR-SCAD and the PLR-LASSO

Tables 1 and 2 provide a comparison of the PLR-SCAD and the PLR-LASSO in terms of estimation accuracy and of quality of the selection process respectively. The regularisation parameter λ is either selected using the BIC-like criterion or the bootstrap procedure. As a first general comment, the results obtained with the PLR-SCAD and $n_{\text{fwd}} = 1000$ are extremely similar, often identical, to those obtained with the PLR-LASSO. This illustrates the argument made in Section 3.2. As n_{fwd} increases, the path provided by the PLR-SCAD converges to the path provided by the PLR-LASSO.

We move to an analysis of Table 1. Focusing on the PLR-LASSO, a clear pattern emerges. For both metrics of estimation accuracy, the bootstrap yields better results in the low sparsity setup (Setup 1), while the BIC performs best in the situation of high sparsity (Setup 2). In almost all cases, and for all values of n_{fwd}, the PLR-SCAD outperforms the PLR-LASSO in terms of estimation of θ_0. This is not surprising as the switch from the LASSO to the SCAD penalty yields better statistical guarantees concerning the weight vector θ_0, as highlighted by Theorem 2.2. Turning to the estimation of the explained Gini coefficient, the

Table 1
SCAD vs LASSO: accuracy of the estimation.

		PLR-SCAD				PLR-LASSO
	n_{fwd}	5	20	50	1000	
		RMSE.Gini expressed in %				
Setup 1 Normal	BIC	1.48	1.49	1.51	1.52	1.52
	Bootstrap	1.50	1.52	1.46	1.47	1.47
Setup 1 Student	BIC	1.57	1.57	1.60	1.62	1.62
	Bootstrap	1.50	1.57	1.50	1.51	1.51
Setup 2 Normal	BIC	1.46	1.47	1.46	1.46	1.46
	Bootstrap	1.53	1.48	1.88	1.59	1.58
Setup 2 Student	BIC	1.52	1.63	1.72	1.72	1.72
	Bootstrap	1.77	1.73	1.91	1.81	1.81
		Distance.θ				
Setup 1 Normal	BIC	0.24	0.25	0.27	0.29	0.29
	Bootstrap	0.27	0.25	0.23	0.26	0.26
Setup 1 Student	BIC	0.26	0.26	0.29	0.31	0.31
	Bootstrap	0.29	0.27	0.25	0.29	0.29
Setup 2 Normal	BIC	0.30	0.29	0.36	0.37	0.37
	Bootstrap	0.34	0.29	0.33	0.37	0.37
Setup 2 Student	BIC	0.36	0.33	0.42	0.42	0.42
	Bootstrap	0.41	0.34	0.41	0.43	0.43

situation is less clear. This echoes the discussion following Theorem 2.3. As the estimation of the explained Gini coefficient uses the index only through its ordering structure, the LASSO could already provide a good estimation performance. Still, in our experiments, one can always find at least one value of n_{fwd} for which the PLR-SCAD outperforms the PLR-LASSO, both using the bootstrap and using the BIC. For Setup 1, the value $n_{\text{fwd}} = 50$ yields the best performance whereas for Setup 2, it is $n_{\text{fwd}} = 5$. This indicates that a suitable value of n_{fwd} should be data-driven.

The quality of the selection process is assessed by the FPR and FNR, displayed in Table 2. For the PLR-LASSO, the bootstrap procedure yields a high FPR and low FNR, while the contrary goes for the BIC. With a low FNR, the bootstrap procedure is therefore optimal at detecting the active covariates. This explains why it performs best in a scenario of low sparsity. In contrast, the BIC procedure has a low FPR. It is therefore suitable at ruling out the non-active covariates and it performs best in the scenario of high sparsity. The PLR-SCAD outperforms the PLR-LASSO when the BIC criterion is used, both in terms of FPR and FNR. More specifically, the value $n_{\text{fwd}} = 20$ is optimal in terms of FPR and the value $n_{\text{fwd}} = 5$ is optimal in terms of FNR. For the bootstrap procedure, the situation is less clear. For Setup 1, one cannot find a value of n_{fwd} that outperforms the PLR-LASSO both in terms of FPR and FNR. However, the PLR-SCAD yields a better tradeoff. For most values, we observe a slightly larger FNR but a substantially lower FPR. For Setup 2, the PLR-SCAD outperforms the PLR-LASSO for most values of n_{fwd}.

Table 2
SCAD vs LASSO: accuracy of the selection.

		PLR-SCAD				PLR-LASSO
	n_{fwd}	5	20	50	1000	
				FPR		
Setup 1 Normal	BIC	1.83	1.45	2.22	2.28	2.28
	Bootstrap	18.48	12.55	16.48	25.50	25.10
Setup 1 Student	BIC	2.62	2.55	4.03	4.18	4.18
	Bootstrap	20.95	13.70	19.43	27.15	26.73
Setup 2 Normal	BIC	0.84	0.50	0.71	0.70	0.70
	Bootstrap	2.56	2.32	11.27	6.13	6.10
Setup 2 Student	BIC	1.41	0.98	1.25	1.24	1.24
	Bootstrap	3.72	3.04	11.46	6.51	6.51
				FNR		
Setup 1 Normal	BIC	12.75	13.95	13.20	14.70	14.70
	Bootstrap	5.60	6.00	4.15	3.05	3.05
Setup 1 Student	BIC	12.55	13.20	13.25	14.55	14.55
	Bootstrap	6.30	7.75	4.65	4.95	4.95
Setup 2 Normal	BIC	14.7	18.70	20.90	21.3	21.3
	Bootstrap	12.6	12.85	7.05	12.3	12.3
Setup 2 Student	BIC	16.20	18.45	23.45	23.60	23.60
	Bootstrap	16.50	15.30	11.25	16.15	16.15

4.2. Comparison with competitors

We now compare the performance of the PLR-SCAD and of the PLR-LASSO with that of the PMSRC, obtained either via the procedure proposed by [11], denoted as PMSRC (LP), or using the FABS algorithm of [15], denoted as PMSRC (FABS). In both cases, the optimal regularisation parameter is obtained via the BIC-like criterion proposed by [11]. All the results that follow were obtained on the Gaussian scenario. The results obtained on the Student distribution are relegated to Appendix C.

We start the comparison with a baseline, where the sample size is fixed to $n = 100$ and the explained Gini coefficient is set to $\text{Gi}_{Y,X} = 0.15$. Table 3 provides the results obtained with the different estimation procedures on both setups. The first two columns evaluate the estimation accuracy for the explained Gini coefficient (RMSE.Gini expressed in percentages) and for the parameter vector (Distance. θ) respectively. The last two columns assess the quality of the selection process through the FPR and FNR. In line with the results outlined in Table 2, the procedures where the penalty parameter is selected via the BIC are well performing where ruling out the non active covariates (low FPR), at the cost of a poorer selection performance for the active ones (high FNR). A relatively clear ranking emerges. The PLR-SCAD performs best, followed by the PLR-LASSO. The PMSRC obtained with the FABS algorithm yields a performance close to the PLR-LASSO in Setup 1. However, it yields inferior results in Setup 2. The PMSRC obtained with the procedure of [11] yields the worst performance. At the other side of the tradeoff, and as we already discussed, the PLR-SCAD and the PLR-LASSO based on the bootstrap offer good performance in terms of

TABLE 3

Comparison of the estimation procedures – Gaussian scenario with $n = 100$ and $Gi_{Y,X} = 0.15$

		PLR-SCAD				PLR-LASSO
	n_{fwd}	5	20	50	1000	
		RMSE.Gini expressed in %				
Setup 1 Normal	BIC	1.48	1.49	1.51	1.52	1.52
	Bootstrap	1.50	1.52	1.46	1.47	1.47
Setup 1 Student	BIC	1.57	1.57	1.60	1.62	1.62
	Bootstrap	1.50	1.57	1.50	1.51	1.51
Setup 2 Normal	BIC	1.46	1.47	1.46	1.46	1.46
	Bootstrap	1.53	1.48	1.88	1.59	1.58
Setup 2 Student	BIC	1.52	1.63	1.72	1.72	1.72
	Bootstrap	1.77	1.73	1.91	1.81	1.81
		Distance.θ				
Setup 1 Normal	BIC	0.24	0.25	0.27	0.29	0.29
	Bootstrap	0.27	0.25	0.23	0.26	0.26
Setup 1 Student	BIC	0.26	0.26	0.29	0.31	0.31
	Bootstrap	0.29	0.27	0.25	0.29	0.29
Setup 2 Normal	BIC	0.30	0.29	0.36	0.37	0.37
	Bootstrap	0.34	0.29	0.33	0.37	0.37
Setup 2 Student	BIC	0.36	0.33	0.42	0.42	0.42
	Bootstrap	0.41	0.34	0.41	0.43	0.43

FNR at the cost of a higher FPR. Table 10 in Appendix C displays the results obtained with the Student distribution and offers similar conclusions.

In the baseline scenario, the variance of the error term was set to ensure $V[Y_i]/V[H(X_i^\intercal\theta_0)] = 3/2$. In the following two scenarios, we use the same value for the variance but we change the value of the explained Gini coefficient. Either we decrease it to $\text{Gi}_{Y,X} = 0.05$, effectively decreasing the strength of the signal, or we increase it to $\text{Gi}_{Y,X} = 0.25$, effectively increasing the strength of the signal. Table 4 gathers the results from these two scenarios. When $\text{Gi}_{Y,X} = 0.05$, the signal is extremely low and the PMSRC performs the best. Since it does not use the value of the response vector but only the ranks, it is more robust to extreme noise level. When $\text{Gi}_{Y,X} = 0.25$, the PMSRC obtained with the procedure of [11] yields the poorest performance. In terms of estimation, all the remaining procedures offer similar results. Interestingly however, they yield different tradeoffs in terms of model selection. The PLR fitted using the BIC criterion performs the best in terms of FPR, while the PLR fitted using the bootstrap procedure yields the lowest FNR. The PMSRC obtained with the FABS offers a middleground with relatively low FPR and FNR. The results obtained with the Student distribution are displayed in Table 11 in Appendix C and offer similar interpretations. Notice however that in the $\text{Gi}_{Y,X} = 0.25$ situation, the PLR obtained with the bootstrap procedure stands out as the best performer.

As a last comparison, we focus on Setup 2 and evaluate the performance of the procedures as the n/p ratio changes. Recall that the number of covariates is $p = 120$ and, in the baseline, the sample size is $n = 100$. We now examine the performance using $n = 50$ and $n = 200$. The results are gathered in Ta-

Table 4

Comparison of the estimation procedure – Gaussian scenario with $n = 100$ and $Gi_{Y,X} \in \{0.05, 0.25\}$.

	Setup 1				Setup 2			
	$\mathrm{Gi}_{Y,X}$	θ_0	FPR	FNR	$\mathrm{Gi}_{Y,X}$	θ_0	FPR	FNR
	Low explained Gini coefficient ($\mathrm{Gi}_{Y,X} = 0.05$)							
PLR-SCAD (BIC)	1.67	0.77	26.25	29.85	3.50	1.08	7.80	54.10
PLR-SCAD (Bootstrap)	1.80	0.79	48.45	23.40	4.42	1.06	26.79	39.80
PLR-LASSO (BIC)	1.58	0.74	30.03	26.35	2.80	1.00	7.59	50.80
PLR-LASSO (Bootstrap)	1.80	0.79	49.27	22.70	4.28	1.06	25.43	42.15
PMSRC (FABS)	1.29	0.83	9.23	52.35	1.37	1.06	1.51	73.50
PMSRC (LP)	1.34	0.90	5.30	62.65	1.28	1.08	1.25	76.90
	High explained Gini coefficient ($\mathrm{Gi}_{Y,X} = 0.25$)							
PLR-SCAD (BIC)	2.16	0.19	0.00	12.95	2.05	0.19	0.00	13.95
PLR-SCAD (Bootstrap)	2.14	0.14	9.62	0.60	2.06	0.16	2.73	1.90
PLR-LASSO (BIC)	2.19	0.20	0.48	10.35	2.15	0.25	0.14	14.30
PLR-LASSO (Bootstrap)	2.12	0.15	18.80	0.35	2.02	0.21	4.73	2.20
PMSRC (FABS)	2.15	0.15	3.97	2.35	2.16	0.22	0.47	9.35
PMSRC (LP)	2.43	0.26	1.63	13.10	2.48	0.32	0.40	22.20

Table 5

Comparison of the estimation procedure – Gaussian scenario and Setup 2 with $n \in \{50, 200\}$.

	$n = 50$				$n = 200$			
	$\mathrm{Gi}_{Y,X}$	θ_0	FPR	FNR	$\mathrm{Gi}_{Y,X}$	θ_0	FPR	FNR
PLR-SCAD (BIC)	2.13	0.56	1.15	35.50	1.05	0.17	0.21	8.20
PLR-SCAD (Bootstrap)	2.84	0.61	10.44	23.15	1.07	0.19	1.22	5.20
PLR-LASSO (BIC)	2.35	0.66	1.31	41.95	1.07	0.21	0.40	8.60
PLR-LASSO (Bootstrap)	2.92	0.68	13.12	24.20	1.07	0.23	1.49	6.25
PMSRC (FABS)	3.50	0.75	0.77	54.30	1.10	0.22	0.68	7.90
PMSRC (LP)	4.47	0.86	0.39	65.45	1.21	0.27	0.29	18.00

ble 5 and are compared with the baseline, see Table 3. When $n = 50$, the n/p ratio decreases and this has several consequences on what a suitable procedure should yield. First, the ability to extract information from the data becomes crucial. This explains why the performance gap between the PLR and the PMSRC increases. Second, it becomes increasingly important to keep a low FPR. Therefore, the performance gap between the PLR fitted with the BIC and the PLR fitted with the bootstrap procedure also increases. Notice finally that the PLR-SCAD remains superior to the PLR-LASSO. Moving to the $n = 200$ scenario, the n/p ratio increases and all the procedures, except PMSRC (LP), offer a similar estimation accuracy. In terms of model selection, we observe the same tradeoff as before. The results obtained with the Student distribution are shown in Table 12 in Appendix C and offer similar conclusions.

4.3. Asymptotic properties of the PLR-SCAD

In what follows, we provide an assessment of the asymptotic properties of the PLR-SCAD, using the SCAD-FABS on Setup 1 with multivariate Gaussian

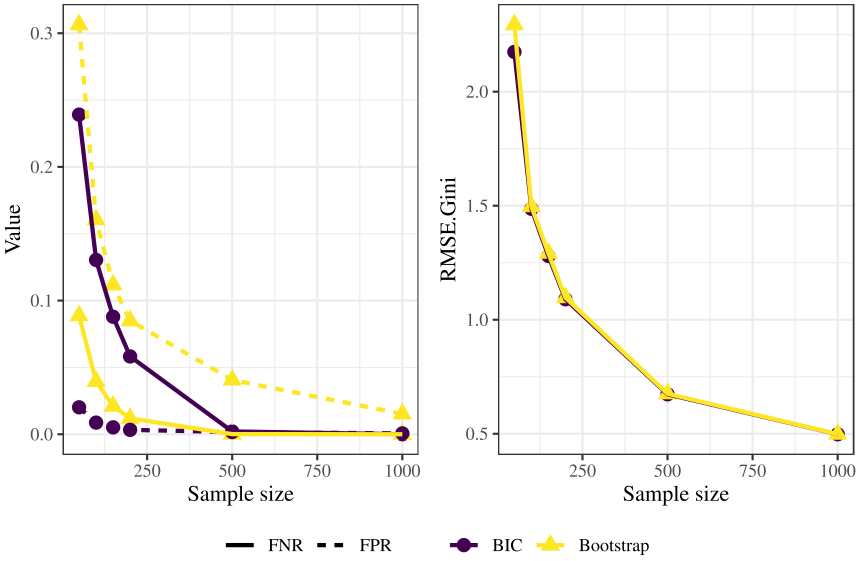

Fig 1. *Consistency of the PLR-SCAD on Setup 1.*

covariates. These results were obtained using $M = 2000$ datasets from the proposed DGP. Figure 1 displays the performance of the BIC and bootstrap procedures for sample sizes ranging from 50 to 1000. The left side of the figure shows the evolution of the FNR (solid lines) and of the FPR (dashed lines) with an increasing sample size. The decay of the FPR is expected due to the property of sparsity stated in Theorem 2.2. The figure displays a similar decay for the FNR. Therefore, as the sample size increases, the PLR-SCAD tends to identify correctly both the active and the non-active sets. The figure also confirms that the BIC criterion yields the best performance for the identification of the zero coefficients (best FPR). On the other hand, the bootstrap procedure is the best at correctly identifying the active set (best FNR). Concerning the explained Gini coefficient, the right hand side of the figure indicates that both procedures yield a consistent estimator and follow very close trajectories in terms of root mean squared error. The asymptotic normality obtained in Theorem 2.3 is confirmed by Figure 8, which shows a qqplot and a histogram of $\widehat{\mathrm{Gi}}_{Y,X}$, obtained on samples of size 1000. This figure is relegated to Appendix C.

4.4. Coverage of the confidence intervals

We construct 95%-confidence intervals for the explained Gini coefficient using three different methods: the basic bootstrap, the percentile bootstrap and the hybrid bootstrap. The basic bootstrap is based on bootstrapping the whole distribution of $\widehat{\mathrm{Gi}}_{Y,X}$. The hybrid bootstrap uses the asymptotic normality and only bootstraps the asymptotic variance. Finally, the percentile bootstrap is obtained by plugging the quantiles of the bootstrap distribution of $\widehat{\mathrm{Gi}}_{Y,X}$. More precisely, $(1-\alpha)$-level confidence intervals for the explained Gini coefficient are given by

$$\mathrm{CI}_{\mathrm{Basic}} := \left[2\widehat{\mathrm{Gi}}_{Y,X} - q_{\widehat{\mathrm{Gi}}^*_{Y,X};1-\frac{\alpha}{2}}; 2\widehat{\mathrm{Gi}}_{Y,X} - q_{\widehat{\mathrm{Gi}}^*_{Y,X};\frac{\alpha}{2}}\right],$$

$$\mathrm{CI}_{\mathrm{Percentile}} := \left[q_{\widehat{\mathrm{Gi}}^*_{Y,X};\frac{\alpha}{2}}; q_{\widehat{\mathrm{Gi}}^*_{Y,X};1-\frac{\alpha}{2}}\right],$$

$$\mathrm{CI}_{\mathrm{Hybrid}} := \left[\widehat{\mathrm{Gi}}_{Y,X} \pm z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}^*_{\xi}}{\sqrt{n}}\right],$$

where $\widehat{\mathrm{Gi}}^*_{Y,X}$ is the estimator of $\mathrm{Gi}_{Y,X}$ in the bootstrap sample. Moreover, $q_{\widehat{\mathrm{Gi}}^*_{Y,X};a}$ is the bootstrap estimator of the a-quantile of the distribution of $\widehat{\mathrm{Gi}}^*_{Y,X}$ and $\hat{\sigma}^*_{\xi}$ is the bootstrap estimator of the standard deviation of $\widehat{\mathrm{Gi}}^*_{Y,X}$. Finally, z_a is the a-quantile of the standard normal distribution. We focus on Setup 1 with multivariate normal covariates and perform our simulations on $M = 2000$ different datasets. We display results using the bootstrap and BIC procedures. The quality of the confidence intervals is assessed in terms of their lengths and coverages.

Figure 2 displays the evolution of the coverage and length for different sample sizes. Solid lines refer to the BIC procedure while dashed lines correspond to the bootstrap procedure. Several observations are worth making. With low sample sizes, all confidence intervals undercover the true parameter. However, the coverages approach the target level of 95% as the sample size increases. In terms of types of bootstrap, a clear ranking emerges. The basic bootstrap performs the worst. The percentile bootstrap comes second while the hybrid bootstrap provides the best performance. Finally, the BIC yields slightly wider confidence intervals with better coverages compared to the bootstrap procedure.

5. Real data example

Our discussion is based on data resulting from the Young Men's Cohort of the National Longitudinal Survey (NLS-Y), a survey started in 1966 on individuals of ages 14-24. The excerpt we use is available in the dataset `Griliches` contained in the `R` package `Ecdat`. Besides wage, schooling (*Schooling*) and experience (*Exp*), we include the following variables in the analysis: age (*Age*), ability (*IQ*), marital status (*Married*), degree of urbanisation (*Urban*) and a dummy indicating whether the individual lives in a Southern State (*South*). Ability was computed as IQ scores collected in a school survey conducted in 1968. All the remaining variables were observed in 1980. The degree of urbanisation is a dummy determining whether the individual lives in a metropolitan area. Before delving into modelling, we note that the considered wage distribution exhibits a mean of 1000, a median of 948 and a Gini coefficient of 22.03%. The concentration indices of wage with respect to age, schooling, experience and ability are respectively of 4.34%, 7.21%, 0.71% and 6.14%.

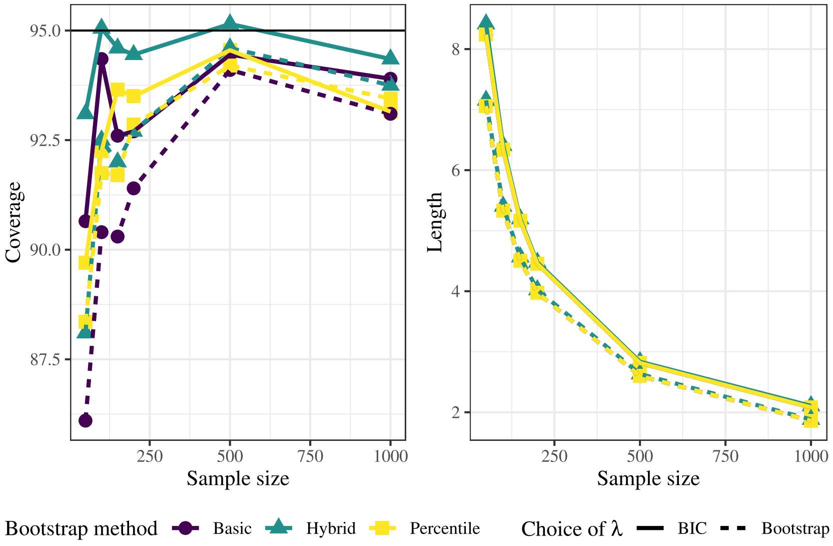

FIG 2. *Coverages and Lengths of confidence intervals for $Gi_{Y,X}$ on Setup 1.*

A PLR analysis of the data is provided in Section 5.1. More specifically, we display the results obtained with the PLR-SCAD for several values of n_{fwd}, and with the PLR-LASSO. The robustness to the number of bootstrap samples and to the choice of the kernel function is evaluated in Section 5.2. Finally, a study of the evolution of the computing time in function of several parameters is provided in Section 5.3.

5.1. PLR analysis of the Griliches data

We run the PLR-LASSO and the PLR-SCAD, with n_{fwd} on the grid 5, 20, 50, on a model including all the covariates as well as the interactions between the binary and the numeric ones. We let the procedure select the variables to include in the final model. In essence, it does so by balancing the complexity of the model with the extra amount of explained inequality that a deshrink in the coefficient would bring. Figure 3 plots the trace of the SCAD-FABS algorithm, with the value $n_{\text{fwd}} = 20$. The horizontal axis displays the evolution of the penalty parameter (in terms of $-\log(\lambda)$), while each line corresponds to the value taken by a given coefficient. At any time, the vector of coefficients is normalized in order to have a unit $L2$-norm. When the algorithm starts, penalisation is the highest and schooling is the first covariate to enter the model. This makes sense since, as we have seen, this is the variable for which we observe the highest concentration index. As the algorithm proceeds, new variables enter the model but they may also shrink back to 0. We observe this phenomenon for several variables throughout the algorithm. The path of the FABS algorithm, related to the PLR-LASSO, is relegated to Appendix D.

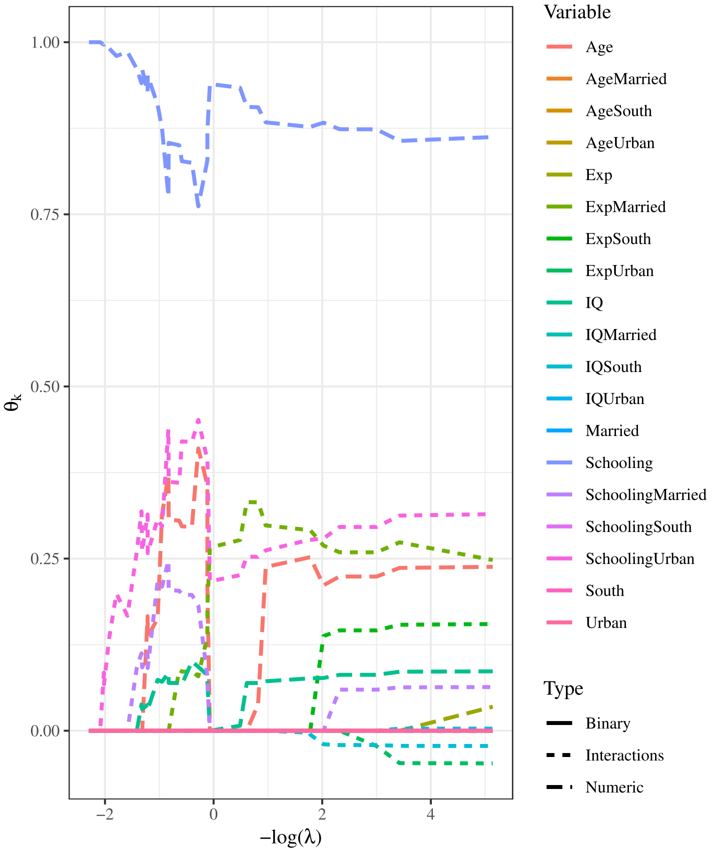

FIG 3. *Trace plot of the SCAD-FABS ($n_{fwd} = 20$) on the Griliches data.*

In order to obtain an estimate for the explained Gini coefficient, one needs to choose a value of λ. As in Section 4, we use the BIC and bootstrap procedures. However, it turns out that both scores tend to reach a plateau after a certain value of λ. This feature is illustrated for the SCAD-FABS (with $n_{\text{fwd}} = 20$) in Figure 4, which displays the evolution of the BIC (solid lines) and OOB (dashed lines) scores with the evolution of the penalty parameter (in terms of $-\log(\lambda)$). The scores are normalized such that the optimum is attained at 1. Taking this into consideration, one may wish to take a value of λ higher than the optimum, at a small cost in terms of optimality, but at a higher benefit in terms of sparsity and interpretability of the model. In practice and for each procedure, we choose the highest value of λ that leads to at most 1% loss in the relative score. In Figure 4, this means to take the first value of $-\log\lambda$ to reach the horizontal line, set at 0.99.

Table 6 summarizes the model fits obtained with the different procedures, using the bootstrap procedure to select λ. As the first line indicates, all procedures

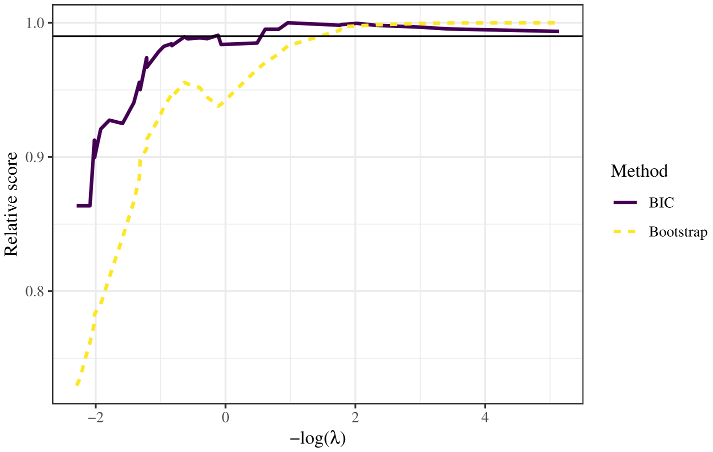

Fig 4. *Evolution of the BIC and OOB-scores for the SCAD-FABS ($n_{fwd} = 20$) on the Griliches data.*

Table 6
Results of the PLR-SCAD and PLR-LASSO on the Griliches data.

	PLR-SCAD			PLR-LASSO
	$n_{\text{fwd}} = 5$	$n_{\text{fwd}} = 20$	$n_{\text{fwd}} = 50$	
$\widehat{\text{Gi}}_{Y,X}$	10.37	10.51	10.50	10.46
95% CI	[8.72, 20.02]	[8.87, 12.15]	[8.83, 12.16]	[8.79, 12.13]
# variables	7	6	5	6
OOB score	9.75	9.78	9.73	9.76

provide an estimated explained Gini coefficient around the same value, ranging from 10.37% to 10.51%. The second line provides 95% hybrid confidence intervals for the explained Gini coefficient. Again, all procedures provide intervals of similar sizes. The third line displays the number of active covariates in the selected model. Notice that it is the lowest for the PLR-SCAD ($n_{\text{fwd}} = 50$) even though it yields the highest estimated explained Gini coefficient. Conversely, it is the highest for the PLR-SCAD ($n_{\text{fwd}} = 5$) while it yields the lowest estimate for the explained Gini coefficient. Finally, the last line shows that all procedures perform similarly in terms of OOB-score. Table 7 shows the estimated coefficients of the active covariates for each procedure. Four covariates are always included: *Age*, *Schooling*, *IQ* and the interaction between *Schooling* and *Urban*. As argued before, as n_{fwd} increases, the coefficients obtained with the PLR-SCAD converge to those obtained with the PLR-LASSO. The summary of the model fits and the estimated coefficients obtained when the BIC procedure is used to select λ are shown in Tables 13 and 14, relegated to Appendix D. The conclusions remain the same, except that the estimated explained Gini coefficient now ranges from 10.19% to 10.33%.

Table 7

Coefficients estimated by the PLR-SCAD and PLR-LASSO on the Griliches data.

	PLR-SCAD			PLR-LASSO
	$n_{\text{fwd}} = 5$	$n_{\text{fwd}} = 20$	$n_{\text{fwd}} = 50$	
Age	0.277	0.252	0.197	0.356
Schooling	0.927	0.877	0.890	0.826
Exp	0.122	/	/	/
IQ	0.063	0.076	0.078	0.107
SchoolingMarried	/	/	/	0.158
SchoolingUrban	0.131	0.277	0.255	0.336
ExpMarried	/	0.291	0.313	0.206
ExpUrban	0.166	/	/	/
IQMarried	0.030	/	/	/
IQSouth	/	−0.003	/	/

Table 8

Evolution of the estimated explained Gini with the number of bootstrap samples.

	B				
	50	100	200	400	1000
PLR-SCAD ($n_{\text{fwd}} = 5$)	10.37	10.37	10.61	10.37	10.37
PLR-SCAD ($n_{\text{fwd}} = 20$)	10.51	10.51	10.51	10.51	10.51
PLR-SCAD ($n_{\text{fwd}} = 50$)	10.51	10.51	10.49	10.50	10.49
PLR-LASSO	10.46	10.44	10.46	10.46	10.44

5.2. Robustness analysis

In what follows, we evaluate the robustness of the results, first to changes in the number of bootstrap samples B and, second, to the specific choice of the kernel. Notice that the parameter B influences the results in two manners. First, it drives the OOB-score and, therefore, influences the estimated explained Gini coefficient obtained with the bootstrap procedure. More directly, it also affects the confidence intervals. Table 8 shows the evolution of the estimated explained Gini coefficient obtained with the bootstrap procedure, for each estimation method, and for a grid of values of B. In general, the results are very stable. In terms of the computation of the OOB-score, it seems therefore that $B = 50$ is already sufficient. Some instability may however occur, see for example PLR-SCAD ($n_{\text{fwd}} = 5$) with $B = 200$. This issue is related to the existence of the plateau mentioned in Section 5.1 and calls for a more refined procedure to select the penalty parameter λ.

Figure 5 displays, for each bootstrap method, the evolution of the confidence intervals with the number of bootstrap samples. Concerning the estimation method, we focus on the PLR-SCAD with ($n_{\text{fwd}} = 20$) but similar pictures are obtained with the other procedures as well. Notice that the hybrid bootstrap (small dashes) leads to more stable results than basic bootstrap (solid) and percentile bootstrap (large dashes). This is expected since the hybrid exploits the bootstrap only to estimate the asymptotic variance. Therefore, it does not require as many bootstrap samples as the other two procedures which rely on the bootstrap to approximate quantiles. In all cases, it seems that $B = 400$ already provides stable results.

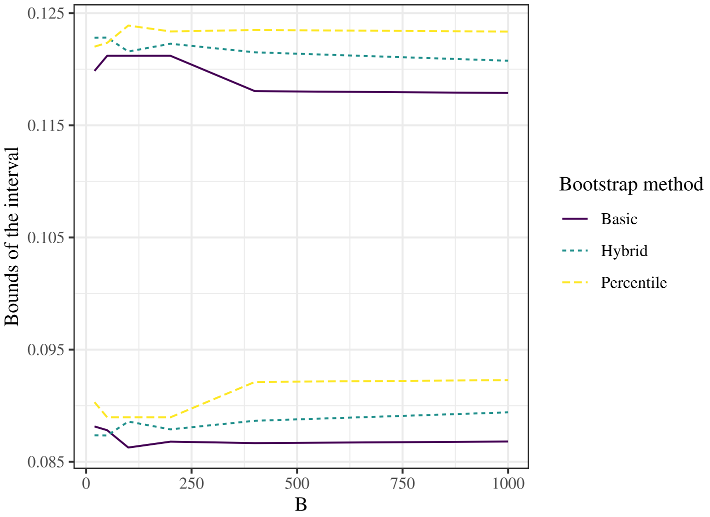

FIG 5. *Evolution of the confidence interval for $Gi_{Y,X}$ for the SCAD-FABS ($n_{fwd} = 20$) on the Griliches data.*

TABLE 9
Results of the PLR-SCAD and PLR-LASSO on the Griliches data (transformed biweight kernel).

		PLR-SCAD		PLR-LASSO
	$n_{\text{fwd}} = 5$	$n_{\text{fwd}} = 20$	$n_{\text{fwd}} = 50$	
$\widehat{\text{Gi}}_{Y,X}$	10.41	10.51	10.42	10.40
95% CI	[8.73, 12.09]	[8.86, 12.16]	[8.77, 12.07]	[8.74, 12.06]
# variables	8	6	7	6
OOB-score	9.73	9.77	9.73	9.76

Let us now turn to the choice of the kernel. The baseline choice is the same as in Section 4, i.e. a fourth-order kernel constructed from an Epanechnikov kernel. As a robustness check, we switch to a fourth-order kernel constructed from a biweight kernel. It satisfies

$$K(u) = \begin{cases} 0 & \text{if } u < -1 \\ \frac{45}{32}u - \frac{25}{16}u^3 + \frac{21}{32}u^5 + \frac{1}{2} & \text{if } u \in [-1, 1] \\ 1 & \text{if } u > 1. \end{cases}$$

Table 9 displays the model fits obtained with the transformed biweight kernel. It is to be compared with Table 6, showing the model fits obtained with the transformed Epanechnikov kernel. The results obtained for the PLR-SCAD with $n_{\text{fwd}} = 20$ and the PLR-LASSO are almost unchanged. For the PLR-SCAD with $n_{\text{fwd}} \in 5, 50$, the use of the transformed biweight kernel induces a slightly larger estimated explained Gini coefficient and an extra included covariate, see the third line of Table 9.

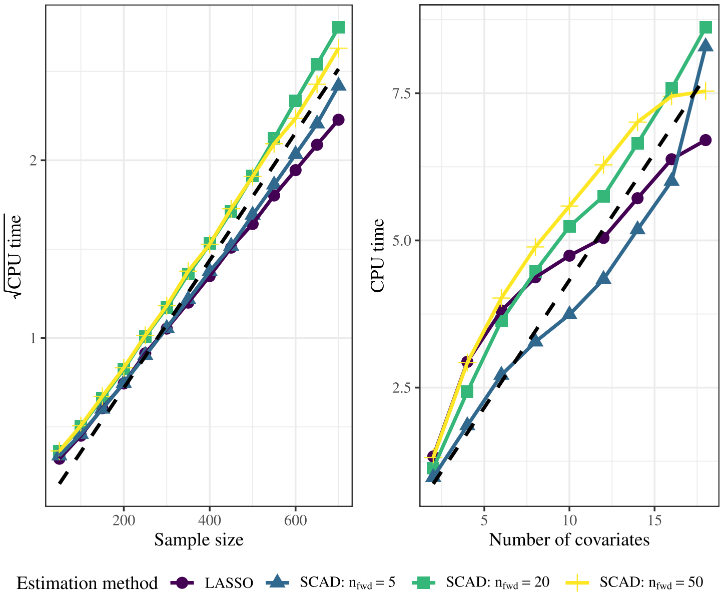

FIG 6. *Evolution of the run time as a function of n and p on the Griliches data.*

5.3. Computation time

We compare next the CPU time of the PLR-SCAD, with $n_{\text{fwd}} \in 5, 20, 50$, and of the PLR-LASSO, and evaluate the impact of a few parameters. More specifically, we examine the influence of the sample size n, of the number of covariates p and of the bandwidth h. To maintain comparability of results, all computations were performed on AMD Epyc-Rome 2.9 GHz CPUs, with 1 GB RAM reserved. In each of the scenario, the computation is repeated $M = 400$ times. Concerning the impact of n, a subsample of the original observations is drawn in each iteration and the bandwidth is fixed to the value used on the full dataset. Concerning p, a subsample of the original covariates is drawn in each iteration. For the other scenarios, the full dataset is used each time.

The evolution of the square-root of the CPU time in seconds as a function of the sample size is displayed in the left part of Figure 6. All the methods follow closely a $O(n^2)$ tendency, even though the LASSO seems to perform slightly more favourably than the SCAD when the sample size increases. The evolution of the CPU time as a function of the number of covariates is represented in the right part of Figure 6. The different methods follow roughly a $O(p)$ trajectory. This $O(pn^2)$ tendency is coherent with the computational complexity involved in the FABS and SCAD-FABS algorithms. Indeed, as stressed out by [15], each iteration is characterized by the computation of the gradient of the loss function, which requires $O(pn^2)$ computations.

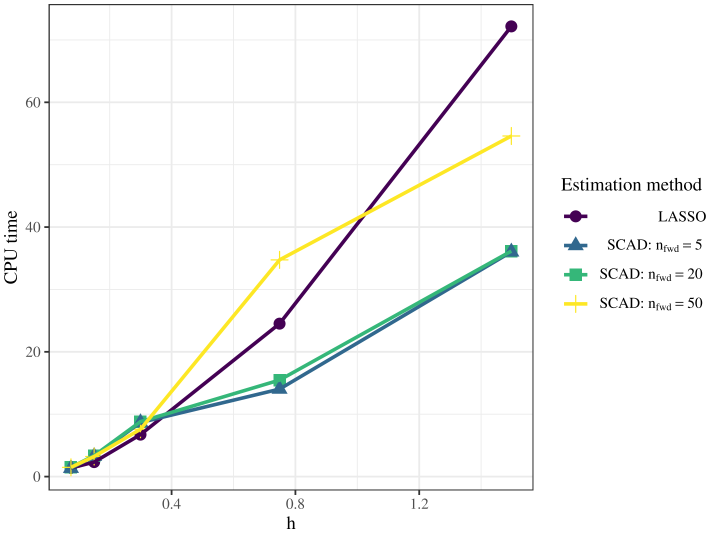

Fig 7. *Evolution of the CPU time in function of h on the Griliches data.*

The impact of the bandwith h is displayed in Figure 7. Overall, the computing time increases with h for all methods. Recall from Section 3 that an increase in h is equivalent to a proportional decrease in ϵ. As the bandwidth increases, the path becomes smoother and the CPU time increases. However, we observe noticeable differences between the methods. The increase in computation time is the largest for the LASSO. As we move to the PLR-SCAD, and especially if we consider low values of n_{fwd}, the increase is less pronounced.

6. Discussion

The penalised Lorenz regression is proposed as a method to produce estimation and inference for the explained Gini coefficient in a high-dimensional setting. Several features make it a suitable choice. First, it shares the good properties of the Lorenz regression developed by [10]. The statistical model underlying the procedure is a single-index and is therefore more flexible than fully parametric models. Also, the link function $H(\cdot)$, which corresponds to the nonparametric part of the model, does not need to be estimated.

Second, the penalised Lorenz regression provides some improvements to the original procedure. The addition of a SCAD penalty ensures that irrelevant covariates are discarded with a probability tending to one and avoids to overestimate the explained Gini coefficient. The presence of a differentiable approximation in the objective function allows the use of the SCAD-FABS algorithm, for which convergence properties are established. The procedure also enjoys strong statistical guarantees. It yields an asymptotically normal estimator for the explained Gini coefficient, with a convergence rate unaffected by the selection process. The inference for the explained Gini coefficient and the selection of the regularisation parameter can be dealt with in a single bootstrap procedure. In the simulations, the procedure proved to be at least as good as the proposed competitors.

What is still missing is an assessment of the contribution of each covariate to the explained Gini coefficient. The estimated weight vector $\hat{\theta}$ provides such a contribution but it has no direct interpretation in terms of inequality. Also, it is computed on the full model and therefore measures a contribution of one covariate given the others. Further research should focus on filling this gap.

Appendix A: Asymptotic properties of the penalised Lorenz regression

In this appendix, we provide the proofs of Theorems 2.1, 2.2 and 2.3. We also state and prove some technical results related to the asymptotic properties of the penalised Lorenz regression.

We rewrite (7) as

$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} Y_i \hat{F}_\theta(X_i^\intercal \theta),$$

where

$$\hat{F}_\theta(t) \coloneqq \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{t - X_i^\intercal \theta}{h}\right).$$

In what follows, for any function $\varphi_\theta(z)$, denote $\varphi'_\theta(z) = \partial\varphi_\theta(z)/\partial z$ and $\nabla_{\theta_0}\varphi_\theta(y) = [\partial\varphi_\theta(y)/\partial\theta_1, \dots, \partial\varphi_\theta(y)/\partial\theta_p]^\intercal|_{\theta=\theta_0}$. Also, $\nabla^2_{\theta_0}\varphi_\theta(y)$ is a $[p \times p]$ matrix whose $[k,l]$ element is given by $[\partial^2\varphi_\theta(y)/\partial\theta_k\partial\theta_l]|_{\theta=\theta_0}$.

Lemma A.1. *Assume (RC1)–(RC4). Then, for any* $0 < \delta < 1$*,*

$$\sup_{z\in\mathcal{Z}} \left|\hat{F}_{\theta_0}(z) - F_{\theta_0}(z)\right| = o_p(1) \tag{19}$$

$$\sup_{z\in\mathcal{Z}} \left|\hat{F}'_{\theta_0}(z) - F'_{\theta_0}(z)\right| = o_p(1) \tag{20}$$

$$\sup_{z,z'\in\mathcal{Z}} \frac{\left|\hat{F}'_{\theta_0}(z) - F'_{\theta_0}(z) - \hat{F}'_{\theta_0}(z') + F'_{\theta_0}(z')\right|}{|z - z'|^\delta} = O_p(1). \tag{21}$$

Proof. Relations (19) and (20) are direct consequences of Theorem 3 in [19] and Theorem A in [16] respectively. Hence, we focus on proving (21). Let $d_n(z) \coloneqq \hat{F}_{\theta_0}(z) - F_{\theta_0}(z)$ and $\beta_n(z,z') \coloneqq (d'_n(z) - d'_n(z'))/|z-z'|^\delta$. We have that $d'_n(z) - d'_n(z') = (z - z')d''_n(z^*)$ for some z^* between z and z'. Hence, it holds

$$\begin{aligned}|\beta_n(z,z')| &\leq |z - z'|^{1-\delta} \sup_{z\in\mathcal{Z}} |d''_n(z)| \\ &\leq C \sup_{z\in\mathcal{Z}} |d''_n(z)|,\end{aligned}$$

for some constant C. The supremum above is $O_p(1)$ by Theorem C in [16]. This finishes the proof. □

Proposition A.2. *Assume (RC1)–(RC4). Then,*

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{Y_i\hat{F}_{\theta_0}(Z_i) - Y_iF_{\theta_0}(Z_i) - E[Y\hat{F}_{\theta_0}(Z)|\mathcal{X}_n] + E[YF_{\theta_0}(Z)]\Big\} = o_p(n^{-1/2}),$$

where $\mathcal{X}_n = \{(X_i^\intercal, Y_i)^\intercal, i = 1, \dots, n\}$ *and* $Z_i = X_i^\intercal\theta_0$*.*

Proof. We are going to use Lemma 19.24 from [17], with

$$\begin{aligned}\hat{f}_n &: (y, z) \mapsto y\hat{F}_{\theta_0}(z) = yd_n(z) + yF_{\theta_0}(z)\\ f_0 &: (y, z) \mapsto yF_{\theta_0}(z),\end{aligned}$$

where $d_n(z) \coloneqq \hat{F}_{\theta_0}(z) - F_{\theta_0}(z)$. Define the class

$$\mathcal{F} = \{(z, y) \to y[d(z) + F_{\theta_0}(z)], d \in C_1^{1+\delta}(\mathcal{Z})\},$$

where $C_1^{1+\delta}(\mathcal{Z})$ is the class of differentiable functions d defined on $\mathcal{Z}$ satisfying $\|d\|_{1+\delta} \leq 1$, where $\|d\|_{1+\delta} = \max\{\sup_z |d(z)|, \sup_z |d'(z)|\} + \sup_{z,z'} |d'(z) - d'(z')|/|z - z'|^\delta$. From Lemma A.1, we have that $P(d_n \in C_1^{1+\delta}(\mathcal{Z})) \to 1$ as $n \to \infty$. We start by showing that $\mathcal{F}$ is Donsker. From Theorem 19.5 in [17], this reduces to showing that

$$\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{P_{Y,Z}})}d\epsilon < \infty, \tag{22}$$

where $N_{[]}$ is the bracketing number, $P_{Y,Z}$ is the probability measure related to $F_{Y,Z}$ and $\|\cdot\|_{P_{Y,Z}}$ denotes the $L2(P_{Y,Z})$-norm. By Corollary 2.7.2 in [18], $O(\exp(\epsilon^{-2/(1+\delta)}))$ brackets are needed for $d(z)$. Hence, the bracketing entropies $\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{P_{Y,Z}})$ are growing with order slower than ϵ^{-2}. Hence, (22) holds and $\mathcal{F}$ is Donsker. To be able to use Lemma 19.24 from [17], it remains to show that

$$\int\int y^2[\hat{F}_{\theta_0}(z) - F_{\theta_0}(z)]^2 dF_{Y,Z}(y, z) = o_p(1).$$

This is a direct consequence of Lemma A.1 and $E[Y^2] < \infty$. □

Proposition A.3. *Assume (RC1)–(RC4). Then,*

$$\frac{1}{n}\sum_{i=1}^{n} Y_i\hat{F}_{\theta_0}(Z_i) = \frac{1}{n}\sum_{i=1}^{n}\Big\{Y_iF_{\theta_0}(Z_i) + \int\int y\eta(Z_i, z)dF_{Y,Z}(y, z)\Big\} + o_p(n^{-1/2}),$$

with $\eta(Z_i, z) = K((z - Z_i)/h) - F_{\theta_0}(z)$*. Also,* $E[\eta(Zi, z)] = O(h^3)$*.*

Proof. Using Proposition A.2, we have

$$\frac{1}{n}\sum_{i=1}^{n} Y_i\hat{F}_{\theta_0}(Z_i) = \frac{1}{n}\sum_{i=1}^{n} Y_iF_{\theta_0}(Z_i) + E\big[Y[\hat{F}_{\theta_0}(Z) - F_{\theta_0}(Z)]|\mathcal{X}_n\big] + o_p(n^{-1/2}).$$

Also, we can write

$$E\big[Y[\hat{F}_{\theta_0}(Z) - F_{\theta_0}(Z)]|\mathcal{X}_n\big] = \frac{1}{n}\sum_{i=1}^{n}\int\int y\eta(Z_i,z)dF_{Y,Z}(y,z).$$

To finish the proof, we show that $E[\eta(Z_i,z)] = O(h^3)$. Let $\mathcal{Z} = [z_1, z_2]$, we have

$$\begin{aligned}
&E\left[K\left(\frac{z-Z_i}{h}\right)\right]\\
&= \int_{z_1}^{z_2} K\left(\frac{z-t}{h}\right)dF_{\theta_0}(t)\\
&= \left[K\left(\frac{z-t}{h}\right)F_{\theta_0}(t)\right]_{z_1}^{z_2} + \int_{z_1}^{z_2}\frac{1}{h}\kappa\left(\frac{z-t}{h}\right)F_{\theta_0}(t)dt\\
&= K\left(\frac{z-z_2}{h}\right) + \int_{z-h}^{z+h}\frac{1}{h}\kappa\left(\frac{t-z}{h}\right)F_{\theta_0}(t)dt - \left[1 - K\left(\frac{z_2-z}{h}\right)\right]\\
&= \int_{-1}^{1}\kappa(u)F_{\theta_0}(z+hu)du\\
&= F_{\theta_0}(z) + O(h^3),
\end{aligned}$$

where the last equality comes from a Taylor expansion of $F_{\theta_0}(z+hu)$ around z. This closes the proof. □

Define next

$$\hat{g}(z) := \frac{1}{nh}\sum_{i=1}^{n}\kappa\left(\frac{z-Z_i}{h}\right)$$
$$\hat{r}(z) := \frac{1}{nh}\sum_{i=1}^{n}\kappa\left(\frac{z-Z_i}{h}\right)X_i^{\intercal}.$$

Hence, $\hat{g}(z)$ is a kernel density estimator of $g(z)$ and $\hat{r}(z)$ is an estimator of $r(z) := g(z)E[X|Z=z]^{\intercal}$. Let $\hat{r}_k(z)$ denote the kth component of vector $\hat{r}(z)$, $\hat{r}_k(z)$ then corresponds to the numerator of a Nadaraya-Watson estimator of X^k on Z. These objects will play a role in the asymptotic representation of $\nabla G_n(\theta_0)$. Indeed, we can write

$$\hat{D}(x, x^{\intercal}\theta_0) := \nabla_{\theta_0}\hat{F}_{\theta}(x^{\intercal}\theta) = x\hat{g}(x^{\intercal}\theta_0) - \hat{r}(x^{\intercal}\theta_0)$$
$$D(x, x^{\intercal}\theta_0) := \nabla_{\theta_0}F_{\theta}(x^{\intercal}\theta) = xg(x^{\intercal}\theta_0) - r(x^{\intercal}\theta_0).$$

Notice that $\sup_{z\in\mathcal{Z}}\big|\hat{g}(z) - g(z)\big| = o_p(1)$ by (20).

Lemma A.4. *Assume (RC1)–(RC4). Then, for any* $0 < \delta < 1$,

$$\sup_{z\in\mathcal{Z}} \left|\hat{g}'(z) - g'(z)\right| = o_p(1) \tag{23}$$

$$\sup_{z,z'\in\mathcal{Z}} \frac{\left|\hat{g}'(z) - g'(z) - \hat{g}'(z') + g'(z')\right|}{|z - z'|^{\delta}} = O_p(1) \tag{24}$$

Proof. Relation (23) is proven in Theorem C from [16]. The proof of (24) follows that of (21) and uses again Theorem C from [16], applied to the second derivative of $\hat{g}(z)$. □

Lemma A.5. *Assume (RC1)–(RC4). Then, for any* $0 < \delta < 1$,

$$\sup_{z\in\mathcal{Z}} \left|\hat{r}(z) - r(z)\right| = o_p(1) \tag{25}$$

$$\sup_{z\in\mathcal{Z}} \left|\hat{r}'(z) - r'(z)\right| = o_p(1) \tag{26}$$

$$\sup_{z,z'\in\mathcal{Z}} \frac{\left|\hat{r}'(z) - r'(z) - \hat{r}'(z') + r'(z')\right|}{|z - z'|^{\delta}} = O_p(1) \tag{27}$$

Proof. We start with the proof of (25). Write $|\hat{r}(z) - r(z)| \leq |\hat{r}(z) - E[\hat{r}(z)]| + |E[\hat{r}(z)] - r(z)|$. Using Proposition 4 from [13], we have that $\sup_{z\in\mathcal{Z}} |\hat{r}(z) - E[\hat{r}(z)]| = o_p(1)$. Also,

$$\begin{aligned}
E[\hat{r}(z)] &= \frac{1}{h} E\left[\kappa\left(\frac{z - Z}{h}\right) X\right] \\
&= \frac{1}{h}\int_{z_1}^{z_2} \kappa\left(\frac{z - t}{h}\right) E[X|Z = t] g(t) dt \\
&= \frac{1}{h}\int_{z-h}^{z+h} \kappa\left(\frac{t - z}{h}\right) r(t) dt \\
&= \int_{-1}^{1} \kappa(u) r(z + hu) du \\
&= r(z) + O(h^3),
\end{aligned}$$

where the last step uses a Taylor expansion of $r(z + hu)$ around z and the fact that $r(\cdot)$ is three times continuously differentiable. This closes the proof of (25). The same structure can be applied to the proof of (26). First, $\sup_{z\in\mathcal{Z}} |\hat{r}'(z) - E[\hat{r}'(z)]| = o_p(1)$ follows as a special case of Theorem 2 in [12]. Then,

$$\begin{aligned}
E[\hat{r}'(z)] &= \frac{1}{h}\int_{-1}^{1} \kappa'(u) r(z + hu) du \\
&= \int_{-1}^{1} \kappa(u) r'(z + hu) du \\
&= r'(z) + O(h^3),
\end{aligned}$$

where the second equality uses integration by parts and the fact that $\kappa(-1) = \kappa(1) = 0$. The last equality is obtained using a Taylor expansion of $r'(z+hu)$ around z and the fact that $r(\cdot)$ is four times continuously differentiable. This finishes the proof of (26). The proof of (27) follows that of (21) and uses again Theorem 2 from [12], applied to the second derivative of $\hat{r}(z)$. □

Proposition A.6. *Assume (RC1)–(RC4). Then,*

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{Y_i\hat{D}(X_i,Z_i) - Y_iD(X_i,Z_i) - E[Y\hat{D}(X,Z)|\mathcal{X}_n] + E[YD(X,Z)]\Big\}$$
$$= o_p(n^{-1/2}).$$

Proof. As in the proof of Proposition A.2, we use Lemma 19.24 from [17], now with

$$\begin{aligned}\hat{f}_n &: (x,z,y) \mapsto y\hat{D}(x,z) = y[xd_n^1(z) + d_n^2(z)] + yD(x,z)\\ f_0 &: (x,z,y) \mapsto yD(x,z),\end{aligned}$$

where $d_n^1(z) := \hat{g}(z) - g(z)$ and $d_n^2(z) := \hat{r}(z) - r(z)$. Define the class

$$\mathcal{F} := \{(x,z,y) \to y[xd^1(z) + d^2(z) + D(x,z)], (d^1,d^2) \in C_1^{1+\delta}(\mathcal{Z})\}.$$

From Lemmas A.1, A.4 and A.5, we have that $P(d_n^1 \in C_1^{1+\delta}(\mathcal{Z})) \to 1$ and $P(d_n^2 \in C_1^{1+\delta}(\mathcal{Z})) \to 1$ as $n \to \infty$. With this in mind, the proof that $\mathcal{F}$ is Donsker is the same as in Proposition A.2. It remains to show that

$$\int\int y^2[xd_n^1(x^\intercal\theta_0) + d_n^2(x^\intercal\theta_0)]^2 dF_{Y,X}(y,x) = o_p(1).$$

This is a direct consequence of Lemmas A.4, A.5, and the fact that $E[Y^2] < \infty$. □

Proposition A.7. *Assume (RC1)–(RC4). Then,*

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} Y_i\hat{D}(X_i,Z_i)\\ &\quad= \frac{1}{n}\sum_{i=1}^{n}\Big\{Y_iD(X_i,Z_i) + \int\int y\eta_1(X_i,x)dF_{Y,X}(y,x)\Big\} + o_p(n^{-1/2}),\end{aligned}$$

with

$$\eta_1(X_i,x) := \frac{x}{h}\kappa\left(\frac{x^\intercal\theta_0 - X_i^\intercal\theta_0}{h}\right) - \frac{X_i}{h}\kappa\left(\frac{x^\intercal\theta_0 - X_i^\intercal\theta_0}{h}\right) - D(x, x^\intercal\theta_0).$$

Also, $E[\eta_1(X_i,x)] = O(h^3)$.

Proof. Using Proposition A.6, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Y_i\hat{D}(X_i,Z_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i D(X_i,Z_i) + E\big[Y[\hat{D}(X,Z)-D(X,Z)]|\mathcal{X}_n\big] + o_p(n^{-1/2}).\end{aligned}$$

The expectation above can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\int\int y\eta_1(X_i,x)dF_{Y,X}(y,x).$$

Hence, $E[\eta_1(X_i,x)] = x^\intercal E[\hat{g}(x^\intercal\theta_0)] - E[\hat{r}(x^\intercal\theta_0)] - D(x,x^\intercal\theta_0)$. By Taylor expansions similar to that operated in the proof of Lemma A.5, we have $E[\hat{g}(z)] = g(z)+O(h^3)$ and $E[\hat{r}(z)] = r(z)+O(h^3)$. This implies that $E[\eta_1(X_i,x)] = O(h^3)$ and closes the proof. □

Denote $\hat{D}^2(x,x^\intercal\theta_0) \coloneqq \nabla^2_{\theta_0}\hat{F}_\theta(x^\intercal\theta)$ and $D^2(x,x^\intercal\theta_0) \coloneqq \nabla^2_{\theta_0}F_\theta(x^\intercal\theta)$.

Proposition A.8. *Assume (RC1)–(RC4). If $E[Y^2]<\infty$, then it holds that*

$$\frac{1}{n}\sum_{i=1}^{n} Y_i\hat{D}^2(X_i,Z_i) - E[YD^2(X,Z)] = o_p(1). \tag{28}$$

Proof. The left-hand side of (28) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_iD^2(X_i,Z_i) - E[YD^2(X,Z)]\} + \frac{1}{n}\sum_{i=1}^{n} Y_i[\hat{D}^2(X_i,Z_i) - D^2(X_i,Z_i)].$$

The first term is $o_p(1)$ by the weak law of large numbers since $E[|YD^2(X,Z)|]<\infty$. We now focus on the second term. By Markov's inequality, it is sufficient to show that

$$E\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i[\hat{D}^2(X_i,Z_i) - D^2(X_i,Z_i)]\right|\right] \to 0,$$

as $n\to\infty$. Since $E[Y^2]<\infty$, it is enough to show that

$$\sup_{x\in\mathcal{X}} |\hat{D}^2(x,x^\intercal\theta_0) - D^2(x,x^\intercal\theta_0)| = o_p(1).$$

Notice that

$$\hat{D}^2(x,x^\intercal\theta_0) = \frac{1}{nh^2}\sum_{i=1}^{n}\kappa'\left(\frac{x^\intercal\theta_0 - X_i^\intercal\theta_0}{h}\right)[x-X_i][x-X_i]^\intercal.$$

The $[k,l]$ element of this matrix can be decomposed into $x^k x^l \hat{g}'(x^\intercal\theta_0) - x^k \hat{r}_l'(x^\intercal\theta_0) - x^l \hat{r}_k'(x^\intercal\theta_0) + \hat{m}_{kl}'(x^\intercal\theta_0)$, where

$$\hat{m}_{kl}(z) := \frac{1}{nh}\sum_{i=1}^{n}\kappa\left(\frac{z-Z_i}{h}\right)X_i^k X_i^l.$$

Similarly, we can decompose the $[k,l]$ element of $D^2(x, x^\intercal\theta_0)$ into $x^k x^l g'(x^\intercal\theta_0) - x^k r_l'(x^\intercal\theta_0) - x^l r_k'(x^\intercal\theta_0) + m_{kl}'(x^\intercal\theta_0)$, where

$$m_{kl}(z) := g(z)E[X^k X^l | Z = z].$$

Using Lemmas A.4 and A.5, we are left with showing $\sup_{z\in\mathcal{Z}}|\hat{m}_{kl}'(z) - m_{kl}'(z)| = o_p(1)$ for any k,l, which is proven in the same way as (26). This closes the proof. □

Proof of Theorem 2.1. Let $\alpha_n := 1/\sqrt{n} + a_n$ and $J_n(\theta) := G_n(\theta) - \sum_{k=1}^{p} p_\lambda(|\theta_k|)$. Following [11], we show that for any given $\epsilon > 0$, there exists a constant C such that

$$P\left(\sup_{\|u\|=1, u^\intercal\theta_0=0} J_n\Big((1 - C^2\alpha_n^2)^{1/2}\theta_0 + C\alpha_n u\Big) < J_n(\theta_0)\right) \geq 1 - \epsilon.$$

Let $\theta^* := (1 - C^2\alpha_n^2)^{1/2}\theta_0 + C\alpha_n u$. Since $p_\lambda(|\theta_{0,k}|) = 0$ for $k \in \{s+1, \ldots, d\}$, it holds

$$J_n(\theta^*) - J_n(\theta_0) \leq G_n(\theta^*) - G_n(\theta_0) - \sum_{k=1}^{s}\Big[p_\lambda(|\theta_k^*|) - p_\lambda(|\theta_{0,k}|)\Big].$$

After Taylor expansions of $G_n(\theta^*)$ around θ_0 and of $p_\lambda(|\theta_k^*|)$ around $|\theta_{0,k}|$, we have

$$J_n(\theta^*) - J_n(\theta_0) \leq I_1 + I_2 + I_3,$$

with

$$\begin{aligned}
I_1 &= \big[\nabla G_n(\theta_0)\big]^\intercal[\theta^* - \theta_0] \\
I_2 &= \frac{1}{2}[\theta^* - \theta_0]^\intercal \nabla^2 G_n(\theta_0)[\theta^* - \theta_0][1 + o_p(1)] \\
I_3 &= -\sum_{k=1}^{s}\Big[p_\lambda'(|\theta_{0,k}|)\text{sign}(\theta_{0,k})[\theta_k^* - \theta_{0,k}] + \frac{1}{2}p_\lambda''(|\theta_{0,k}|)[\theta_k^* - \theta_{0,k}]^2[1 + o(1)]\Big].
\end{aligned}$$

We first focus on I_1. Using Proposition A.7, the weak law of large numbers and $E[|YD^2(X,Z)|] < \infty$, we have that $\nabla G_n(\theta_0) = O_p(n^{-1/2})$. Using the definition of θ^* and a Taylor expansion of $(1 - C^2\alpha_n^2)^{\frac{1}{2}}$ around 1, we conclude that $I_1 = O_p(C\alpha_n/\sqrt{n})$. Let us now consider I_2. Using Proposition A.8, $\nabla^2 G_n(\theta_0) = O_p(1)$. Hence, $I_2 = O_p(C^2\alpha_n^2)$. Finally, I_3 is bounded above by

$$s\alpha_n a_n + s\alpha_n^2 C^2 \max\{|p_\lambda''(|\theta_{0,k}|)| : \theta_{0,k} \neq 0\}.$$

Hence, choosing a sufficiently large C, I_2 is the dominating term. The fact that it is asymptotically negative stems from (RC7). □

Proof of Theorem 2.2. Take any θ^A such that $\|\theta^A - \theta_0^A\| = O_p(1/\sqrt{n})$ and $\|\theta^A\| = 1$, and take any constant C. We show that

$$J_n((\theta^{A\intercal}, 0^\intercal)^\intercal) = \max_{\|\theta^I\| \leq C/\sqrt{n}} J_n((\theta^{A\intercal}, \theta^{I\intercal})^\intercal).$$

It is sufficient to show that for some small $\epsilon_n = C/\sqrt{n}$ and for $k = s+1, \dots, p$, we have

$$\begin{aligned} \nabla_k J_n(\theta) &< 0 && \text{for } 0 < \theta_k < \epsilon_n \\ &> 0 && \text{for } -\epsilon_n < \theta_k < 0. \end{aligned}$$

We have

$$\begin{aligned} \nabla_k J_n(\theta) &= \nabla_k G_n(\theta) - p'_\lambda(|\theta_k|)\text{sign}(\theta_k) \\ &= O_p(n^{-1/2}) - p'_\lambda(|\theta_k|)\text{sign}(\theta_k), \end{aligned}$$

using a Taylor expansion of $\nabla_k G_n(\theta)$ around θ_0 as well as Propositions A.7 and A.8. Hence,

$$\nabla_k J_n(\theta) = \lambda\Big[-\frac{p'_\lambda(|\theta_k|)}{\lambda}\text{sign}(\theta_k) + O_p\Big(\frac{1}{\sqrt{n}\lambda}\Big)\Big].$$

Recall that $\liminf_{n\to\infty} \liminf_{x\to 0^+} p'_\lambda(x)/\lambda > 0$. Hence, asymptotically, the sign of $\nabla_k J_n(\theta)$ is fully determined by the sign of θ_k. This closes the first part of this proof. Now we move on with the proof of the asymptotic normality. Since $\theta_{0,s} > 0$, it is easy to show that there exists $(\hat\theta_1, \dots, \hat\theta_{s-1})^\intercal$ that is a $\sqrt{n}$-consistent local solution of the following maximization programme

$$\max_{(\theta_1,\dots,\theta_{s-1})} J_n\left(\left(\theta_1, \dots, \theta_{s-1}, \sqrt{1 - \sum_{k=1}^{s-1} \theta_k^2}, 0^\intercal\right)^\intercal\right).$$

Let $\hat\theta_s = \sqrt{1 - \sum_{k=1}^{s-1} \hat\theta_k^2}$, $\hat\theta^A = (\hat\theta_1, \dots, \hat\theta_s)$ and $\hat\theta = (\hat\theta^{A\intercal}, 0^\intercal)^\intercal$. For all $k = 1, \dots, s-1$, $\hat\theta_k$ must satisfy

$$\frac{\partial}{\partial\hat\theta_k} J_n\left(\left(\hat\theta_1, \dots, \hat\theta_{s-1}, \sqrt{1 - \sum_{k=1}^{s-1} \hat\theta_k^2}, 0^\intercal\right)^\intercal\right) = 0.$$

We develop this derivative, starting with the part related to the non-penalised objective function. We have

$$\begin{aligned} &\frac{\partial}{\partial\hat\theta_k} G_n\left(\left(\hat\theta_1, \dots, \hat\theta_{s-1}, \sqrt{1 - \sum_{k=1}^{s-1} \hat\theta_k^2}, 0^\intercal\right)^\intercal\right) \\ &= \frac{1}{n}\sum_{i=1}^n Y_i \Big[\frac{1}{n}\sum_{j=1}^n \kappa\Big(\frac{X_i^\intercal\hat\theta - X_j^\intercal\hat\theta}{h}\Big)\frac{[X_i^k - X_j^k]}{h}\Big] \end{aligned}$$

$$-\frac{\hat{\theta}_k}{\hat{\theta}_s}\frac{1}{n}\sum_{i=1}^{n} Y_i\Bigg[\frac{1}{n}\sum_{j=1}^{n}\kappa\bigg(\frac{X_i^\intercal\hat{\theta}-X_j^\intercal\hat{\theta}}{h}\bigg)\frac{[X_i^s-X_j^s]}{h}\Bigg].$$

Using a Taylor expansion on the function $(\hat{\theta}_1,\dots,\hat{\theta}_{s-1})^\intercal \mapsto \kappa(\frac{X_i^\intercal\hat{\theta}-X_j^\intercal\hat{\theta}}{h})$ around $(\theta_{0,1},\dots,\theta_{0,s-1})^\intercal$ and the previous notations, the partial derivative above can be written as

$$\begin{aligned}&\nabla_k G_n(\theta_0)-\frac{\hat{\theta}_k}{\hat{\theta}_s}\nabla_s G_n(\theta_0)+\sum_{l=1}^{s-1}\nabla^2_{kl}G_n(\theta_0)[\hat{\theta}_l-\theta_{0,l}]\\ &\quad-\sum_{l=1}^{s-1}\frac{\theta_{0,l}}{\theta_{0,s}}\nabla^2_{ks}G_n(\theta_0)[\hat{\theta}_l-\theta_{0,l}]-\frac{\hat{\theta}_k}{\hat{\theta}_s}\sum_{l=1}^{s-1}\nabla^2_{ls}G_n(\theta_0)[\hat{\theta}_l-\theta_{0,l}]\\ &\quad+\frac{\hat{\theta}_k}{\hat{\theta}_s}\sum_{l=1}^{s-1}\frac{\theta_{0,l}}{\theta_{0,s}}\nabla^2_{ss}G_n(\theta_0)[\hat{\theta}_l-\theta_{0,l}]+o_p(1).\end{aligned}$$

We focus now on the part related to the penalty. Using the consistency of $\hat{\theta}_k$, we have

$$\begin{aligned}&\frac{\partial}{\partial\hat{\theta}_k}\left\{\sum_{k=1}^{s-1}p_\lambda(|\hat{\theta}_k|)+p_\lambda\big(\sqrt{1-\textstyle\sum_{k=1}^{s-1}\hat{\theta}_k^2}\big)\right\}\\ &\qquad=p'_\lambda(|\hat{\theta}_k|)\mathrm{sign}(\hat{\theta}_k)\,\mathbb{1}\{\hat{\theta}_k\neq 0\}-p'_\lambda(\hat{\theta}_s)\frac{\hat{\theta}_k}{\hat{\theta}_s}+o_p(1).\end{aligned}$$

Using again the consistency of $\hat{\theta}_k$, Taylor expansions of $p'_\lambda(|\hat{\theta}_k|)$ around $\theta_{0,k}$ and of $p'_\lambda(\hat{\theta}_s)$ around $\theta_{0,s}$, the partial derivative above writes as

$$\begin{aligned}&\left[p'_\lambda(|\theta_{0,k}|)\mathrm{sign}(\theta_{0,k})-p'_\lambda(\theta_{0,s})\frac{\hat{\theta}_k}{\hat{\theta}_s}+p''_\lambda(|\theta_{0,k}|)[\hat{\theta}_k-\theta_{0,k}]\right]\mathbb{1}\{\hat{\theta}_k\neq 0\}\\ &\quad+\frac{\hat{\theta}_k}{\hat{\theta}_s}\sum_{l=1}^{s-1}\left[p''_\lambda(\theta_{0,s})\frac{\theta_{0,l}}{\theta_{0,s}}\right](\hat{\theta}_l-\theta_{0,l})+o_p(1).\end{aligned}$$

Finally, we can get rid of the indicator since $\mathbb{1}\{\hat{\theta}_k\neq 0\}$ converges in probability to one. Notice that in the case of the SCAD, the penalty is only piecewise differentiable. However, this is not an issue since $\lambda\to 0$ as $n\to\infty$. Hence, for a sufficiently large n, the value of $\theta_{0,k}$ will not lie at a discontinuity point. In what follows, we adopt the following notations. For a vector v, we denote by $\tilde{v}$ the vector obtained by taking the first $s-1$ elements of v. For a matrix M, we denote by $\tilde{M}$ the matrix formed by taking the first $s-1$ rows and columns of M. Also, we denote by $\tilde{M}_{.,s}$ the vector formed by taking the first $s-1$ rows and fixing the s-th column of M. Define

$\Sigma_1 \coloneqq E[Y\tilde{D}^2(X,Z)]$.
$\Sigma_2 \coloneqq E[Y\tilde{D}^2_{.,s}(X,Z)]\frac{\tilde{\theta}_0^\intercal}{\theta_{0,s}}$.

$\Sigma_3 := \frac{\vartheta_0 \tilde{\theta}_0^\intercal}{\theta_{0,s}^2} E[Y D_{ss}^2(X, Z)]$.
Σ_4 is a $(s-1) \times (s-1)$ diagonal matrix where the diagonal is given by $[p_\lambda''(|\theta_{10}|)\text{sign}(\theta_{10}), \ldots, p_\lambda''(|\theta_{s-1,0}|)\text{sign}(\theta_{s-1,0})]^\intercal$.
$\Sigma_5 := \frac{\vartheta_0 \tilde{\theta}_0^\intercal}{\theta_{0,s}^2} p_\lambda''(\theta_{0,s})$.
$\Omega_1 := E[\tilde{\xi}_i \tilde{\xi}_i^\intercal]$.
$\Omega_2 := \frac{\vartheta_0 \tilde{\theta}_0^\intercal}{\theta_{0,s}^2} E[\xi_{is}^2]$.
$\Omega_3 := E[\tilde{\xi}_i \xi_{is}] \frac{\vartheta_0^\intercal}{\theta_{0,s}}$

where

$$\xi_i := Y_i D(X_i, Z_i) + \int \int y \eta_1(X_i, x) dF_{Y,X}(y, x).$$

Also, let

$$\Sigma := -\Sigma_1 + \Sigma_2 + \Sigma_2^\intercal - \Sigma_3 + \Sigma_4 - \Sigma_5 \tag{29}$$
$$\Omega := \Omega_1 + \Omega_2 - \Omega_3 - \Omega_3^\intercal. \tag{30}$$

In order to avoid heavier notations, we define $\vartheta_0 = \vartheta_0$ and $\hat{\vartheta} = \hat{\tilde{\theta}}$. Recall that b is defined in Equation (9). Then, we have

$$\begin{aligned} \sqrt{n}\Sigma[\hat{\vartheta} - \vartheta_0 + \Sigma^{-1}b] &= \sqrt{n}\tilde{\nabla}G_n(\theta_0) - \frac{\vartheta_0}{\theta_{0,s}}\sqrt{n}\nabla_s G_n(\theta_0) + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\tilde{\xi}_i - \frac{\vartheta_0}{\theta_{0,s}} \xi_{is} \right) + o_p(1), \end{aligned}$$

where the last equality uses Proposition A.7. Notice that $E[Y_i D(X_i, Z_i)] = 0$ because the function $\theta \mapsto E[Y F_\theta(X^\intercal \theta)]$ is maximized in θ_0. Hence, using Proposition A.7 and the central limit theorem, we have

$$\sqrt{n}\Sigma[\hat{\vartheta} - \vartheta_0 + \Sigma^{-1}b] \xrightarrow{d} N(0, \Omega),$$

which concludes the proof. □

Proposition A.9. *Assume (RC1)–(RC7). Then,*

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \Big[& \mathbb{1}\{X_i^\intercal \theta_n \leq x^\intercal \theta_n\} - \mathbb{1}\{X_i^\intercal \theta_0 \leq x^\intercal \theta_0\} - P(X^\intercal \theta_n \leq x^\intercal \theta_n | \mathcal{X}_n) \\ & + P(X^\intercal \theta_0 \leq x^\intercal \theta_0) \Big] = o_p(n^{-1/2}) \end{aligned}$$

uniformly in $x \in \mathcal{X}$ *and in* $\theta_n \xrightarrow{p} \theta_0$*, with* $\theta_{n,s} > 0$.

Proof. In this proof, we will make use of Corollary 2.3.12 in [18]. We first show that $f_n(X_i) := \mathbb{1}\{X_i^\intercal \theta_n \leq x^\intercal \theta_n\} - \mathbb{1}\{X_i^\intercal \theta_0 \leq x^\intercal \theta_0\} - P(X_i^\intercal \theta_n \leq x^\intercal \theta_n | \mathcal{X}_n) +$

$P(X_i^\intercal\theta_0 \leq x^\intercal\theta_0)$ belongs to a Donsker class. Define $\theta_0^C = (\frac{\theta_{0,1}}{\theta_{0,s}}, \dots, \frac{\theta_{0,s-1}}{\theta_{0,s}}, \frac{\theta_{0,s+1}}{\theta_{0,s}}, \dots, \frac{\theta_{0,p}}{\theta_{0,s}})^\intercal$ and $\theta_n^C = (\frac{\theta_{n,1}}{\theta_{n,s}}, \dots, \frac{\theta_{n,s-1}}{\theta_{n,s}}, \frac{\theta_{n,s+1}}{\theta_{n,s}}, \dots, \frac{\theta_{n,p}}{\theta_{n,s}})^\intercal$, where $\theta_n \xrightarrow{p} \theta_0$. Also, let $x^c = (x^1, \dots, x^{s-1}, x^{s+1}, \dots, x^p)^\intercal$. Notice that the first term of $f(X_i)$ can be rewritten as $\mathbb{1}\{X_i^s \leq x^\intercal\theta_n^C - X_i^{C\intercal}\theta_n^C\}$. Define the class

$$\mathcal{F} \coloneqq \Bigg\{ x \mapsto \big[\, \mathbb{1}\{x^s \leq t - x^{C\intercal}\theta^C\} - \mathbb{1}\{x^s \leq u - x^{C\intercal}\theta_0^C\} - P(X^s \leq t - X^{C\intercal}\theta^C) \\ + P(X^s \leq u - x^{C\intercal}\theta_0^C)\big], t \in \mathbb{R}, u \in \mathbb{R}, \theta^C \in \Theta^C \Bigg\},$$

where θ^C is a vector of size $(p-1)$ defined on a compact set Θ^C. We prove that $\mathcal{F}$ is Donsker using the same reasoning as in Proposition A.2. We focus on the class $\mathcal{F}_1 \coloneqq \{x \mapsto \mathbb{1}\{x^s \leq t - (x^C)_\intercal\theta^C\}, \theta^C \in \Theta^C, t \in \mathbb{R}\}$. The other terms can be dealt with in a similar way. Embed θ^C into a hypercube $[\theta_1^l, \theta_1^u] \times \cdots \times [\theta_{p-1}^l, \theta_{p-1}^u]$ of dimension $(p-1)$. For all j, partition $[\theta_j^l, \theta_j^u]$ into $O(\epsilon^{-2})$ intervals of length $O(\epsilon^2)$. Hence, we partitioned θ^C into $O(\epsilon^{-2(p-1)})$ hypercubes. Denote by R_i one such hypercube. For each non-empty R_i, let

$$\Gamma_i^l(X^C) \coloneqq \min_{\theta^C \in (R_i \cap \Theta^C)} X^{C\intercal}\theta^C$$
$$\Gamma_i^u(X^C) \coloneqq \max_{\theta^C \in (R_i \cap \Theta^C)} X^{C\intercal}\theta^C.$$

Now, notice that, for all t, there exists an i such that

$$\mathbb{1}\{X^s \leq t - \Gamma_i^u(X^C)\} \leq \mathbb{1}\{X^s \leq t - (X^C)^\intercal\theta^C\} \leq \mathbb{1}\{X^s \leq t - \Gamma_i^l(X^C)\}.$$

Define $P_i^u(t) \coloneqq P(X^s \leq t - \Gamma_i^u(X^C))$ and partition the line into segments t_{ik}^u, with $k = 1, \dots, O(\epsilon^{-2})$, and with P_i^u-probability less than a fraction of ϵ^2. Similarly, define $P_i^l(t) \coloneqq P(X^s \leq t - \Gamma_i^l(X^C))$ and partition the line into segments t_{ik}^l, with $k = 1, \dots, O(\epsilon^{-2})$, and with P_i^l-probability less than a fraction of ϵ^2. Denote by $t_{ik_1}^l$ the largest $t_{ik}^l \leq t$ and denote by $t_{ik_2}^l$ the smallest $t_{ik}^u \geq t$. Brackets for $\mathbb{1}\{X^s \leq t - (X^C)^\intercal\theta^C\}$ are then given by $[\mathbb{1}\{t_{ik_1}^l - \Gamma_i^u(X^C)\}, \mathbb{1}\{t_{ik_2}^u - \Gamma_i^l(X^C)\}]$. Let us now compute their size. Denoting $\|\cdot\|_{P_X}$ the $L2(P_X)$-norm, we have

$$\begin{aligned}
&\|\, \mathbb{1}\{X^s \leq t_{ik_2}^u - \Gamma_i^l(X^C)\} - \mathbb{1}\{X^s \leq t_{ik_1}^l - \Gamma_i^u(X^C)\}\|_{P_X}^2 \\
&\quad = P(X^s \leq t_{ik_2}^u - \Gamma_i^l(X^C)) - P(X^s \leq t_{ik_1}^l - \Gamma_i^u(X^C)) \\
&\quad = P_i^l(t) - P_i^u(t) + O(\epsilon^2) \\
&\quad = \int \Bigg[F_{X^s|X^C}(t - \Gamma_i^l(x^C)|x^C) - F_{X^s|X^C}(t - \Gamma_i^u(x^C)|x^C)\big] dF_{X^C}(x^C) \Bigg] + O(\epsilon^2) \\
&\quad = \int \Bigg[\sup_{x^s, x^C} f_{X^1|X^C}(x^s|x^C)\epsilon^2 \Bigg] dF_{X^C}(x^C) + O(\epsilon^2) = O(\epsilon^2),
\end{aligned}$$

where $f_{X^s|X^C}$ and $F_{X^s|X^C}$ denote the conditional density and the conditional CDF of X^s with respect to X^C, and F_{X^C} is the CDF of X^C. Hence, the brackets are of size $O(\epsilon)$. In conclusion, the bracketing number associated to $\mathcal{F}_1$ is

$O(\epsilon^{-2p})$. Using Theorem 19.5 in [17], it follows that $\mathcal{F}_1$ is Donsker. By straightforward computations, $\mathcal{F}$ is also Donsker. Let $V[\cdot]$ denote the variance, we have

$$\begin{aligned} V[f_n(X)|\mathcal{X}_n] &= V[\mathbb{1}\{X^\intercal\theta_n \le x^\intercal\theta_n\} - \mathbb{1}\{X^\intercal\theta_0 \le x^\intercal\theta_0\}|\mathcal{X}_n] \\ &\le E[(\mathbb{1}\{X^\intercal\theta_n \le x^\intercal\theta_n\} - \mathbb{1}\{X^\intercal\theta_0 \le x^\intercal\theta_0\})^2|\mathcal{X}_n] \\ &\le E[(\mathbb{1}\{X^\intercal\theta_n \le x^\intercal\theta_n\} - \mathbb{1}\{X^\intercal\theta_0 \le x^\intercal\theta_0\})|\mathcal{X}_n]^{2/3} \\ &= o_p(1), \end{aligned}$$

where the last line uses the continuous mapping theorem and the fact that $\theta_n \xrightarrow{p} \theta_0$. Since $\mathcal{F}$ is Donsker, it follows from Corollary 2.3.12 in [18] that

$$\lim_{\alpha\downarrow 0}\limsup_{n\to\infty} P\left(\sup_{f\in\mathcal{F},V[f]<\alpha} n^{-1/2}\left|\sum_{i=1}^n f(X_i)\right| > \overline{\epsilon}\right) = 0,$$

for each $\overline{\epsilon} > 0$. We obtain the desired result by restricting the above probability to elements in $\mathcal{F}$ satisfying $\theta^C = \theta_n^C$. □

Proof of Theorem 2.3. We can rewrite

$$\widehat{\text{Gi}}_{Y,X} = \frac{1}{\overline{Y}}\frac{2}{n}\sum_{i=1}^n Y_i\left[\frac{1}{n}\sum_{j=1}^n \mathbb{1}\{X_j^\intercal\hat{\theta} \le X_i^\intercal\hat{\theta}\}\right] - \frac{n+1}{n}$$

where $\hat{\theta}$ is the estimator of Theorem 2.2. Using Proposition A.9 and the same notations as in Theorem 2.2, we can write

$$\frac{2}{\sqrt{n}}\sum_{i=1}^n Y_i\left[\frac{1}{n}\sum_{j=1}^n \mathbb{1}\{X_j^\intercal\hat{\theta} \le X_i^\intercal\hat{\theta}\}\right] = A_{n1} + A_{n2} + o_p(1),$$

with

$$\begin{aligned} A_{n1} &= \frac{2}{\sqrt{n}}\sum_{i=1}^n Y_i\left[\frac{1}{n}\sum_{j=1}^n \mathbb{1}\{X_j^\intercal\theta_0 \le X_i^\intercal\theta_0\}\right] \\ A_{n2} &= \frac{2}{n}\sum_{i=1}^n Y_i \int \sqrt{n}\Bigg[\mathbb{1}\left\{x^s \le [\tilde{X}_i - \tilde{x}]^\intercal\frac{\hat{\vartheta}}{\hat{\theta}_s} + [X_i^I - x^I]^\intercal\frac{\hat{\theta}^I}{\hat{\theta}_s} + X_i^s\right\} \\ &\qquad - \mathbb{1}\left\{x^s \le [\tilde{X}_i - \tilde{x}]^\intercal\frac{\vartheta_0}{\theta_{0,s}} + X_i^s\right\}\Bigg] dF_X(\tilde{x}, x^s), \end{aligned}$$

where we recall that $\tilde{X}_i$ is the vector obtained by taking the first $s-1$ elements of X_i. Let us first focus on A_{n2}. We can rewrite the integral displayed inside the sum as

$$\begin{aligned} \int \sqrt{n}\Bigg[&F_{X^s}\left([\tilde{X}_i - \tilde{x}]^\intercal\frac{\hat{\vartheta}}{\hat{\theta}_s} + [X_i^I - x^I]^\intercal\frac{\hat{\theta}^I}{\hat{\theta}_s} + X_i^s\right) \\ &- F_{X^s}\left([\tilde{X}_i - \tilde{x}]^\intercal\frac{\vartheta_0}{\theta_{0,s}} + X_i^s\right)\Bigg] dF_{\tilde{X}}(\tilde{x}), \end{aligned}$$

where F_{X^s} is the CDF of X^s and $F_{\tilde{X}}$ is the CDF of $\tilde{X}$. Using Theorem 2.1, a Taylor expansion on the piece inside the integral yields

$$f_{X^s}\left([\tilde{X}_i - \tilde{x}]^\intercal \frac{\vartheta_0}{\theta_{0,s}} + X_i^s\right)\left([\tilde{X}_i - \tilde{x}]^\intercal \sqrt{n}\left[\frac{\hat{\vartheta}}{\hat{\theta}_s} - \frac{\vartheta_0}{\theta_{0,s}}\right] + [X_i^I - x^I]^\intercal \sqrt{n}\frac{\hat{\theta}^I}{\hat{\theta}_s}\right) \\ + o_p(1).$$

By Theorem 2.1 and the property of sparsity proven in Theorem 2.2, the part related to $\hat{\theta}^I$ is negligible. Using the same development as in Theorem 2.2, we have

$$\begin{aligned}\sqrt{n}\left[\frac{\hat{\vartheta}}{\hat{\theta}_s} - \frac{\vartheta_0}{\theta_{0,s}}\right] &= \frac{\sqrt{n}[\hat{\vartheta} - \vartheta_0]}{\theta_{0,s}} - \frac{\hat{\vartheta}}{\hat{\theta}_s}\frac{\sqrt{n}[\hat{\theta}_s - \theta_{0,s}]}{\theta_{0,s}} \\ &= \frac{1}{\theta_{0,s}^3}\frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{[\theta_{0s}^2 + \vartheta_0\vartheta_0^\intercal]\tilde{\xi}_i - \frac{\vartheta_0}{\theta_{0,s}}\xi_{is}\right\} + o_p(1).\end{aligned}$$

Hence,

$$A_{n2} = \frac{1}{\theta_{0,s}^3}\frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{E[\rho_i^\intercal]\left([\theta_{0s}^2 + \vartheta_0\vartheta_0^\intercal]\tilde{\xi}_i - \frac{\vartheta_0}{\theta_{0,s}}\xi_{is}\right)\right\} + o_p(1),$$

where

$$\rho_i = 2Y_i \int f_{X^s}\left([\tilde{X}_i - \tilde{x}]^\intercal \frac{\vartheta_0}{\theta_{0,s}} + X_i^s\right)[\tilde{X}_i - \tilde{x}] dF_{\tilde{X}}(\tilde{x}).$$

We now focus on A_{n1}. We can write

$$A_{n1} = \sqrt{n}U_{n1} + o_p(1),$$

where

$$U_{n1} = \frac{2}{n(n-1)}\sum_{i<j} Y_i\, \mathbb{1}\{X_j^\intercal\theta_0 \le X_i^\intercal\theta_0\} + Y_j\, \mathbb{1}\{X_i^\intercal\theta_0 \le X_j^\intercal\theta_0\}$$

is a U-statistic. Let $m_H(t) \coloneqq E[Y_i\, \mathbb{1}\{X_i^\intercal\theta_0 \ge t\}]$. Using Theorem 12.3 in [17] and the fact that $E[Y_i^2] < \infty$, we have

$$\sqrt{n}[U_{n1} - 2G(\theta_0)] = \frac{2}{n}\sum_{i=1}^n \left[Y_i F_{\theta_0}(X_i^\intercal\theta_0) + m_H(X_i^\intercal\theta_0) - 2G(\theta_0)\right] + o_p(1),$$

and has mean zero. Using the last developments and the fact that $E[Y_i] > 0$, we can write

$$\sqrt{n}[\widehat{\mathrm{Gi}}_{Y,X} - \mathrm{Gi}_{Y,X}] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \zeta_i + o_p(1),$$

where

$$\zeta_i := \frac{1}{E[Y_i]}\Bigg(2Y_i F_{\theta_0}(X_i^{\intercal}\theta_0) + 2m_H(X_i^{\intercal}\theta_0) - 4G(\theta_0) \\ + \frac{E[\rho_i^{\intercal}]}{\theta_{0,s}^3}\bigg([\theta_{0,s}^2 + \vartheta_0\vartheta_0^{\intercal}]\tilde{\xi}_i - \frac{\vartheta_0}{\theta_{0,s}}\xi_{is}\bigg)\Bigg). \tag{31}$$

Hence, $\sqrt{n}[\widehat{\mathrm{Gi}}_{Y,X} - \mathrm{Gi}_{Y,X}] \xrightarrow{d} N(0, \sigma_\zeta^2)$, with $\sigma_\zeta^2 := V[\zeta_i]$. □

Appendix B: Properties of the SCAD-FABS algorithm

Proof of Lemma 3.1. Let $k \in \mathcal{A}^t$ be updated via a forward step. Then, it both holds that

$$L(\theta^t - \mathrm{sign}(\theta_l^t)\mathbf{1}_l\epsilon) - L(\theta^t) - \epsilon p'_{\lambda^t}(|\theta_l^t|) \geq 0 \qquad \forall l \in \mathcal{A}^t \tag{32}$$

$$L(\theta^{t+1}) < L(\theta^t). \tag{33}$$

(32) holds because $t \to t+1$ is not a backward step and (33) comes from the fact that a forward step always decreases the loss. Now, assume per contra that $\theta_k^{t+1} = \theta_k^t - \mathrm{sign}(\theta_k^t)\epsilon$. Equation (33) implies $L(\theta^t - \mathrm{sign}(\theta_k^t)\mathbf{1}_k\epsilon) - L(\theta^t) < 0$, which contradicts (32). □

Lemma B.1. *Consider that coefficient k is updated via a forward step.*

(a) If $\lambda^{t+1} = \lambda^t$ and $|\theta_k^t| \leq \lambda^t$, then $\lambda^t \leq \lambda_A^{t+1}$.
(b) If $\lambda^{t+1} = \lambda^t$ and $\lambda^t < |\theta_k^t| \leq a\lambda^t$, then $\lambda^t \leq \lambda_B^{t+1}$.

Proof. Recall that λ^t is updated according to (13). Also, λ_A^{t+1} and λ_B^{t+1} are determined by (15) and (16).

We start with the proof of part (a). Let $\lambda^{t+1} = \lambda^t$ and $|\theta_k^t| \leq \lambda^t$. We may face two scenarios

$$\lambda_A^{t+1} \geq \lambda_B^{t+1} \tag{34}$$

$$\lambda_B^{t+1} > \lambda_A^{t+1}. \tag{35}$$

In situation (34), $\lambda^t \leq \lambda_A^{t+1}$ holds trivially because $\lambda^t = \lambda^{t+1}$. Assume now that (35) holds. Using the definition of λ_A^{t+1} and λ_B^{t+1}, we have $\lambda_B^{t+1} < |\theta_k^t|$. Hence, we must have $\lambda_B^{t+1} < |\theta_k^t| \leq \lambda^t$, which contradicts $\lambda^t \leq \lambda_B^{t+1}$. Hence, (35) cannot happen, which closes the proof of part (a).

We move to the proof of part (b). Let $\lambda^{t+1} = \lambda^t$ and $\lambda^t < |\theta_k^t| \leq a\lambda^t$. Once again, we may face either (34) or (35). If (35) holds, then $\lambda^t \leq \lambda_B^{t+1}$ arises trivially from $\lambda^{t+1} = \lambda^t$. Assume now (34). Using the definition of λ_A^{t+1} and λ_B^{t+1}, we now have $\lambda_B^{t+1} \geq |\theta_k^t|$. Hence, we must have $\lambda^t < |\theta_k^t| \leq \lambda_B^{t+1}$, which closes the proof of part (b). □

Lemma B.2. *Let $k \in \{1, \ldots, p\}$ be updated via a forward step and $\lambda > 0$. Then, there exists some $c_k^t \in [0,1]$ such that*

$$Q(\theta^{t+1}, \lambda) - Q(\theta^t, \lambda) = L(\theta^{t+1}) - L(\theta^t) + p'_\lambda(|\theta_k^t|)\epsilon - \frac{c_k^t\epsilon^2}{2(a-1)}.$$

Proof. Let $\lambda > 0$ and consider that $k \in \{1, \dots, p\}$ is updated via a forward step. By Lemma 3.1, we have $\theta^{t+1} = \theta^t + \text{sign}(\theta_k^t)\mathbf{1}_k\epsilon$. We face the following situations:

$$|\theta_k^{t+1}| \leq \lambda \tag{36}$$

$$|\theta_k^t| \leq \lambda < |\theta_k^{t+1}| \leq a\lambda \tag{37}$$

$$\lambda < |\theta_k^t| < |\theta_k^{t+1}| \leq a\lambda \tag{38}$$

$$\lambda < |\theta_k^t| \leq a\lambda < |\theta_k^{t+1}| \tag{39}$$

$$|\theta_k^t| > a\lambda. \tag{40}$$

In situation (36), there is no approximation error since

$$p_\lambda(|\theta_k^{t+1}|) - p_\lambda(|\theta_k^t|) = \lambda\epsilon = p'_\lambda(|\theta_k^t|)\epsilon.$$

Let us move to situation (37). Then, we have

$$p_\lambda(|\theta_k^{t+1}|) - p_\lambda(|\theta_k^t|) = p'_\lambda(|\theta_k^t|)\epsilon - c_{k,1}^2 \frac{\epsilon^2}{2(a-1)},$$

where $c_{k,1} := \frac{|\theta_k^{t+1}| - \lambda}{\epsilon} \in [0, 1]$. Consider now situation (38). Then, we have

$$p_\lambda(|\theta_k^{t+1}|) - p_\lambda(|\theta_k^t|) = p'_\lambda(|\theta_k^t|)\epsilon - \frac{\epsilon^2}{2(a-1)}.$$

In cases covered by (39), we have

$$p_\lambda(|\theta_k^{t+1}|) - p_\lambda(|\theta_k^t|) = p'_\lambda(|\theta_k^t|)\epsilon - (1 - c_{k,2}^2)\frac{\epsilon^2}{2(a-1)},$$

where $c_{k,2} := \frac{|\theta_k^{t+1}| - a\lambda}{\epsilon} \in [0, 1]$. Obviously, there is no approximation error in (40) since

$$p_\lambda(|\theta_k^{t+1}|) - p_\lambda(|\theta_k^t|) = 0 = p'_\lambda(|\theta_k^t|)\epsilon.$$

□

Lemma B.3. *Let $k \in \mathcal{A}^t$ be updated via a backward step and $\lambda > 0$. Then, there exists some $d_k^t \in [0, 1]$ such that*

$$Q(\theta^{t+1}, \lambda) - Q(\theta^t, \lambda) = L(\theta^{t+1}) - L(\theta^t) - p'_\lambda(|\theta_k^t|)\epsilon - \frac{d_k^t\epsilon^2}{2(a-1)}.$$

This result can be proven by the same reasoning as the one used in the proof of Lemma B.2.

Lemma B.4. *Let $k \in \{1, \dots, p\}$ be updated via a forward step. Then, $\forall l \neq k$ such that $L(\theta^t - sign(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon) - L(\theta^t) < 0$, it holds*

$$L(\theta^{t+1}) - L(\theta^t - sign(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon) + \left[p'_{\lambda^t}(|\theta_k^t|) - p'_{\lambda^t}(|\theta_l^t|)\right]\epsilon \leq \frac{e_k^t - e_l^t}{2}\epsilon^2,$$

where $e_k^t = \nabla_{kk}^2 L(\dot{\theta})$, with $\dot{\theta}$ between θ^t and $\theta^t - \text{sign}(\nabla_m L(\theta^t))\mathbf{1}_k\epsilon$ and ∇^2 denotes the Hessian matrix.

Proof. Consider that $k \in \{1, \dots, p\}$ is updated via a forward step. Using a Taylor expansion, we have

$$L(\theta^{t+1}) - L(\theta^t) = -\nabla_k L(\theta^t)\text{sign}(\nabla_k L(\theta^t))\epsilon + \frac{e_k^t}{2}\epsilon^2.$$

This implies that

$$[|\nabla_k L(\theta^t)| - p'_{\lambda^t}(|\theta_k^t|)]\epsilon = L(\theta^t) - L(\theta^{t+1}) - p'_{\lambda^t}(|\theta_k^t|)\epsilon + \frac{e_k^t}{2}\epsilon^2.$$

Remark that, in a forward step, the left hand side of the previous equation is maximized on the set $\{l \in \{1, \dots, p\} : L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon) - L(\theta^t) < 0\}$. Hence, for such coefficients, it holds that

$$\begin{aligned} L(\theta^t) - L(\theta^t - \text{sign}(\nabla_l L(\theta^t))) - p'_{\lambda^t}(|\theta_k^t|)\epsilon + \frac{e_l^t}{2}\epsilon^2 \\ \leq L(\theta^t) - L(\theta^{t+1}) - p'_{\lambda^t}(|\theta_k^t|)\epsilon + \frac{e_k^t}{2}\epsilon^2, \end{aligned}$$

which leads us to the desired result. □

Proof of Proposition 3.2. Let k be the index of the updated coefficient. We first show that, in a backward step, it holds

$$Q(\theta^{t+1}, \lambda^{t+1}) < Q(\theta^t, \lambda^t) - \frac{d_k^t\epsilon^2}{2(a-1)}.$$

Since we are considering a backward step, it both holds $\lambda^{t+1} = \lambda^t$ and

$$L(\theta^{t+1}) - L(\theta^t) - \epsilon p'_{\lambda^t}(|\theta_k^t|) < 0.$$

Using Lemma B.3, this implies

$$Q(\theta^{t+1}, \lambda^{t+1}) - Q(\theta^t, \lambda^t) < -d_k^t\frac{\epsilon^2}{2(a-1)}.$$

We now prove that, in a forward step, it holds

$$Q(\theta^{t+1}, \lambda^{t+1}) \leq Q(\theta^t, \lambda^t) - \frac{c_k^t\epsilon^2}{2(a-1)}. \tag{41}$$

Assume first that $\lambda^{t+1} = \lambda_A^{t+1}$. Recall that $\lambda_A^{t+1} = L_\epsilon^{t,t+1}$ and $|\theta_k^t| \leq \lambda_A^{t+1}$. Using Lemma B.2, it holds

$$Q(\theta^{t+1}, \lambda_A^{t+1}) - Q(\theta^t, \lambda_A^{t+1}) = -\frac{c_k^t\epsilon^2}{2(a-1)}.$$

Since $\lambda_A^{t+1} \leq \lambda^t$, we also have $Q(\theta^t, \lambda_A^{t+1}) \leq Q(\theta^t, \lambda^t)$. Taken together, the last two results yield (41). Second, consider that $\lambda^{t+1} = \lambda_B^{t+1}$. Recall that $\lambda_B^{t+1} =$

$\frac{1}{a}[(a-1)L_\epsilon^{t,t+1} + |\theta_k^t|]$ and $|\theta_k^t| > \lambda_B^{t+1}$. Using the same reasoning as in the last point yields (41). Third, we assume that $\lambda^{t+1} = \lambda^t$ and $|\theta_k^t| \leq \lambda^t$. Using Lemma B.2, it holds

$$\begin{aligned} Q(\theta^{t+1}, \lambda^t) - Q(\theta^t, \lambda^t) &= L(\theta^{t+1}) - L(\theta^t) + \epsilon\lambda^t - \frac{c_k^t \epsilon^2}{2(a-1)} \\ &\leq L(\theta^{t+1}) - L(\theta^t) + \epsilon\lambda_A^{t+1} - \frac{c_k^t \epsilon^2}{2(a-1)} \\ &= -\frac{c_k^t \epsilon^2}{2(a-1)}, \end{aligned}$$

where the inequality in the second line is due to Lemma B.1. Fourth, let $\lambda^{t+1} = \lambda^t$ and $\lambda^t < |\theta_k^t| \leq a\lambda^t$. Relation (41) is proven similarly to the last situation. Finally, assume $|\theta_k^t| > a\lambda^t$. The result emerges using Lemma B.2 and the fact that a forward step never increases the loss. □

Proposition B.5. *If $\lambda^{t+1} < \lambda^t$ and k is the index of the updated coefficient, then $\forall l \in \mathcal{A}^t$, it holds*

$$\begin{aligned} Q(\theta^t, \lambda^t) &\leq Q(\theta^t + \mathrm{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) + \frac{\epsilon^2}{2}\left[\frac{c_l^t}{a-1} + (e_k^t - e_l^t)\right] \\ Q(\theta^t, \lambda^t) &\leq Q(\theta^t - \mathrm{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) + \frac{d_l^t \epsilon^2}{2(a-1)}. \end{aligned}$$

Proof. Let $\lambda^{t+1} < \lambda^t$ and k be the index of the updated coefficient. Consider any $l \in \mathcal{A}^t$. We first show that

$$Q(\theta^t, \lambda^t) \leq Q(\theta^t + \mathrm{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) + \frac{\epsilon^2}{2}\left[\frac{c_l^t}{a-1} + (e_k^t - e_l^t)\right]. \tag{42}$$

We can safely restrict to the set $\{l \in \mathcal{A}^t : L(\theta^t + \mathrm{sign}(\theta_l^t)\mathbf{1}_l\epsilon) - L(\theta^t) < 0\}$. Indeed, on the complementary set, (42) is obviously true since θ^t leads to lower penalty and loss values. Write $\tilde{\theta} \coloneqq \theta^t + \mathrm{sign}(\theta_l^t)\mathbf{1}_l\epsilon$. We prove the result by showing that it both holds

$$Q(\theta^t, \lambda^t) \leq Q(\theta^{t+1}, \lambda^t) + \frac{c_k^t \epsilon^2}{2(a-1)} \tag{43}$$

$$Q(\theta^{t+1}, \lambda^t) - Q(\tilde{\theta}, \lambda^t) \leq \frac{\epsilon^2}{2}\left(\frac{(c_l^t - c_k^t)}{a-1} + (e_k^t - e_l^t)\right). \tag{44}$$

Let us first focus on (43). Suppose first that $\lambda^{t+1} = \lambda_A^{t+1}$. In that case, $\lambda^t \geq \lambda_A^{t+1} \geq |\theta_k^t|$. Hence

$$\begin{aligned} Q(\theta^t, \lambda^t) &= Q(\theta^{t+1}, \lambda^t) + L(\theta^t) - L(\theta^{t+1}) - \epsilon\lambda^t + \frac{c_k^t \epsilon^2}{2(a-1)} \\ &\leq Q(\theta^{t+1}, \lambda^t) + L(\theta^t) - L(\theta^{t+1}) - \epsilon\lambda_A^{t+1} + \frac{c_k^t \epsilon^2}{2(a-1)} \end{aligned}$$

$$= Q(\theta^{t+1}, \lambda^t) + \frac{c_k^t \epsilon^2}{2(a-1)}.$$

Consider now that $\lambda^{t+1} = \lambda_B^{t+1}$. We have $\lambda^t \geq \lambda_B^{t+1} \geq \lambda_A^{t+1}$ and $\lambda_B^{t+1} < |\theta_k^t| \leq a\lambda_B^{t+1}$. Hence, $|\theta_k^t| \leq a\lambda^t$. But it might either be that $|\theta_k^t| \leq \lambda^t$ or $\lambda^t < |\theta_k^t| \leq a\lambda^t$. In the first situation, relation (43) follows from the exact same reasoning as above. If $\lambda^t < |\theta_k^t| \leq a\lambda^t$, we have

$$\begin{aligned}
Q(\theta^t, \lambda^t) &= Q(\theta^{t+1}, \lambda^t) + L(\theta^t) - L(\theta^{t+1}) - \epsilon\frac{a\lambda^t - |\theta_k^t|}{a-1} + \frac{c_k^t \epsilon^2}{2(a-1)} \\
&\leq Q(\theta^{t+1}, \lambda^t) + L(\theta^t) - L(\theta^{t+1}) - \epsilon\frac{a\lambda_B^{t+1} - |\theta_k^t|}{a-1} + \frac{c_k^t \epsilon^2}{2(a-1)} \\
&= Q(\theta^{t+1}, \lambda^t) + \frac{c_k^t \epsilon^2}{2(a-1)}.
\end{aligned}$$

We now turn to proving (44). We have

$$\begin{aligned}
Q(\tilde{\theta}, \lambda^t) - Q(\theta^t, \lambda^t) &= L(\tilde{\theta}) - L(\theta^t) + p'_{\lambda^t}(|\theta_l^t|)\epsilon - \frac{c_l^t \epsilon^2}{2(a-1)} \\
Q(\theta^{t+1}, \lambda^t) - Q(\theta^t, \lambda^t) &= L(\theta^{t+1}) - L(\theta^t) + p'_{\lambda^t}(|\theta_k^t|)\epsilon - \frac{c_k^t \epsilon^2}{2(a-1)}.
\end{aligned}$$

Together, these two equations imply that

$$\begin{aligned}
&Q(\theta^{t+1}, \lambda^t) - Q(\tilde{\theta}, \lambda^t) \\
&\quad = L(\theta^{t+1}) + p'_{\lambda^t}(|\theta_k^t|)\epsilon - [L(\tilde{\theta}) + p'_{\lambda^t}(|\theta_l^t|)\epsilon] + \frac{\epsilon^2}{2(a-1)}(c_l^t - c_k^t) \\
&\quad \leq \frac{\epsilon^2}{2}\left(\frac{(c_l^t - c_k^t)}{a-1} + (e_k^t - e_l^t)\right).
\end{aligned}$$

The inequality in the last equation is justified as follows. First, following the proof of Lemma 3.1, it is easy to show that $\text{sign}(\theta_l^t) = -\text{sign}(\nabla_l L(\theta^t))$. We can then use Lemma B.4 to obtain the desired inequality. Now, we show that

$$Q(\theta^t, \lambda^t) \leq Q(\theta^t - \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) + \frac{d_l^t \epsilon^2}{2(a-1)}.$$

Since $t \mapsto t+1$ is a forward step, it means that no backward step is taken. Hence, for all $l \in \mathcal{A}^t$, we have

$$L(\theta^t - \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon) - L(\theta^t) - \epsilon p'_{\lambda^t}(|\theta_l^t|) \geq 0.$$

Using Lemma B.3 leads to the desired result. □

Proof of Theorem 3.4. Let t be such that $\lambda^t < \lambda^{t+1}$ and consider that k is the index of the updated coefficient. Recall that m is the upper bound of the second order derivatives of $L(\cdot)$.

Suppose first that $l \in \mathcal{A}^t$, we want to show that $|\nabla_l Q(\theta^t, \lambda^t)| \leq m\epsilon$. Remark that $\nabla_l Q(\theta^t, \lambda^t) = \nabla_l L(\theta^t) + p'_{\lambda^t}(|\theta^t_l|)\text{sign}(\theta^t_l)$. Using Taylor expansions of $L(\theta^t + \text{sign}(\theta^t_l)\mathbf{1}_l\epsilon)$ and of $L(\theta^t - \text{sign}(\theta^t_l)\mathbf{1}_l\epsilon)$, both around θ^t, we have

$$L(\theta^t + \text{sign}(\theta^t_l)\mathbf{1}_l\epsilon) - L(\theta^t) = \nabla_l L(\theta^t)\text{sign}(\theta^t_l)\epsilon + \frac{e^t_l}{2}\epsilon^2 \tag{45}$$

$$L(\theta^t - \text{sign}(\theta^t_l)\mathbf{1}_l\epsilon) - L(\theta^t) = -\nabla_l L(\theta^t)\text{sign}(\theta^t_l)\epsilon + \frac{e^t_l}{2}\epsilon^2. \tag{46}$$

Using Lemma B.2, we can rewrite (45) as

$$\begin{aligned} &Q(\theta^t + \text{sign}(\theta^t_l)\mathbf{1}_l\epsilon, \lambda^t) \\ &\quad = Q(\theta^t, \lambda^t) + p'_{\lambda^t}(|\theta^t_l|)\epsilon + \nabla_l L(\theta^t)\text{sign}(\theta^t_l)\epsilon + \frac{e^t_l}{2}\epsilon^2 - \frac{c^t_l\epsilon^2}{2(a-1)}. \end{aligned}$$

Using Proposition B.5, we have

$$Q(\theta^t, \lambda^t) \leq Q(\theta^t + \text{sign}(\theta^t_l)\mathbf{1}_l\epsilon, \lambda^t) + \frac{\epsilon^2 c^t_l}{2(a-1)} + \frac{\epsilon^2}{2}[e^t_k - e^t_l].$$

Putting the last two results together, we obtain that

$$\epsilon p'_{\lambda^t}(|\theta^t_l|) + \nabla_l L(\theta^t)\text{sign}(\theta^t_l)\epsilon + \frac{e^t_k}{2}\epsilon^2 \geq 0.$$

Since $e^t_k \leq m$, we get

$$-\left[\nabla_l L(\theta^t)\text{sign}(\theta^t_l) + p'_{\lambda^t}(|\theta^t_l|)\right] \leq \frac{m\epsilon}{2}. \tag{47}$$

We now focus on (46). Using Lemma B.3, we rewrite it as

$$\begin{aligned} &Q(\theta^t - \text{sign}(\theta^t_l)\mathbf{1}_l\epsilon, \lambda^t) \\ &\quad = Q(\theta^t, \lambda^t) + p'_{\lambda^t}(|\theta^t_l|)\epsilon - \nabla_l L(\theta^t)\text{sign}(\theta^t_l)\epsilon + \frac{e^t_l}{2}\epsilon^2 - \frac{d^t_l\epsilon^2}{2(a-1)}. \end{aligned}$$

Again, using Proposition B.5, we have that

$$Q(\theta^t, \lambda^t) \leq Q(\theta^t - \text{sign}(\theta^t_l)\mathbf{1}_l\epsilon, \lambda^t) + \frac{\epsilon^2 d^t_l}{2(a-1)}.$$

These last two results imply that

$$\nabla_l L(\theta^t)\text{sign}(\theta^t_l) + p'_{\lambda^t}(|\theta^t_l|) \leq \frac{m\epsilon}{2}. \tag{48}$$

Together, (47) and (48) lead to the desired result.

Consider now that $l \notin \mathcal{A}^t$. We want to show that $|\nabla_l L(\theta^t)| \leq p'_{\lambda^t}(|\theta^t_l|) + m\epsilon$. Using a Taylor expansion of $L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon)$ around θ^t, we have

$$L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon) - L(\theta^t) = -\nabla_l L(\theta^t)\text{sign}(\nabla_l L(\theta^t))\epsilon + \frac{e^t_l}{2}\epsilon^2,$$

which we can rewrite as

$$|\nabla_l L(\theta^t)| = \frac{L(\theta^t) - L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon)}{\epsilon} + \frac{e_l^t}{2}\epsilon. \tag{49}$$

In what follows, we restrict to the set $\{l \notin \mathcal{A}^t : L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon) - L(\theta^t) < 0\}$. Notice that in the complementary set, we directly have $|\nabla_l L(\theta^t)| \leq m\epsilon$. Using Lemma B.4, we have

$$\frac{L(\theta^{t+1}) - L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon)}{\epsilon} \leq p'_{\lambda^t}(|\theta_l^t|) - p'_{\lambda^t}(|\theta_k^t|) + \frac{e_k^t - e_l^t}{2}\epsilon. \tag{50}$$

Returning to Equation (49), we can write

$$\begin{aligned} |\nabla_l L(\theta^t)| &= \frac{L(\theta^t) - L(\theta^{t+1})}{\epsilon} + \frac{L(\theta^{t+1}) - L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon)}{\epsilon} + \frac{e_l^t\epsilon}{2} \\ &\leq L_\epsilon^{t,t+1} + p'_{\lambda^t}(|\theta_l^t|) - p'_{\lambda^t}(|\theta_k^t|) + \frac{e_k^t\epsilon}{2}, \end{aligned}$$

where the inequality is implied by (50). Suppose that $|\theta_k^t| \leq \lambda^t$. Then, it holds

$$L_\epsilon^{t,t+1} - p'_{\lambda^t}(|\theta_k^t|) = \lambda_A^{t+1} - \lambda^t \leq 0,$$

where the inequality is due to $\lambda_A^{t+1} = \lambda^{t+1} < \lambda^t$. Consider instead that $\lambda^t < |\theta_k^t| \leq a\lambda^t$. In that case,

$$L_\epsilon^{t,t+1} - p'_{\lambda^t}(|\theta_k^t|) = \frac{a\lambda_B^{t+1} - |\theta_k^t|}{a-1} - \frac{a\lambda^t - |\theta_k^t|}{a-1} \leq 0,$$

where the inequality is due to $\lambda_B^{t+1} = \lambda^{t+1} < \lambda^t$. Finally, the situation covered by $|\theta_k^t| > a\lambda^t$ can be disregarded since k is the index of the updated coefficient and $\lambda^{t+1} < \lambda^t$. In conclusion, we have

$$\begin{aligned} |\nabla_l L(\theta^t)| - p'_{\lambda^t}(|\theta_l^t|) &\leq \frac{e_k^t\epsilon}{2} \\ &\leq \frac{m\epsilon}{2}, \end{aligned}$$

which closes the second part of the proof. □

Appendix C: Supplementary results concerning the Monte-Carlo simulations

In what follows, we provide supplementary results concerning the simulation studies performed in Section 4. Tables 10, 11 and 12 display the results of the comparison between the estimation procedures obtained for the Student distribution, see Tables 3, 4 and 5 for their equivalent in the Gaussian case. Next, Figure 8 provide a qqplot and histogram of the estimated explained Gini coefficient obtained in Section 4.3.

Table 10
Comparison of the estimation procedures – Student distribution with $n = 100$ and $Gi_{Y,X} = 0.15$

	Setup 1				Setup 2			
	$\text{Gi}_{Y,X}$	θ_0	FPR	FNR	$\text{Gi}_{Y,X}$	θ_0	FPR	FNR
PLR-SCAD (BIC)	1.55	0.26	1.93	13.30	1.50	0.33	0.93	16.90
PLR-SCAD (Bootstrap)	1.51	0.26	19.72	4.75	1.70	0.34	4.57	13.25
PLR-LASSO (BIC)	1.62	0.31	4.18	14.55	1.72	0.42	1.24	23.60
PLR-LASSO (Bootstrap)	1.51	0.29	26.73	4.95	1.81	0.43	6.51	16.15
PMSRC (FABS)	1.70	0.31	6.02	13.20	2.12	0.44	0.82	28.50
PMSRC (LP)	2.15	0.41	2.82	25.30	2.76	0.53	0.66	36.65

Table 11
Comparison of the estimation procedure – Student distribution with $n = 100$ and $Gi_{Y,X} \in \{0.05, 0.25\}$.

	Setup 1				Setup 2			
	$\text{Gi}_{Y,X}$	θ_0	FPR	FNR	$\text{Gi}_{Y,X}$	θ_0	FPR	FNR
	Low explained Gini coefficient ($\text{Gi}_{Y X} = 0.05$)							
PLR-SCAD (BIC)	1.65	0.78	25.32	31.30	3.48	1.10	8.19	55.80
PLR-SCAD (Bootstrap)	1.80	0.81	44.53	27.05	4.39	1.09	28.70	41.95
PLR-LASSO (BIC)	1.56	0.75	29.57	27.40	2.72	1.04	7.84	53.55
PLR-LASSO (Bootstrap)	1.80	0.80	45.02	26.25	4.32	1.10	27.44	42.80
PMSRC (FABS)	1.31	0.84	9.28	52.85	1.39	1.10	1.57	76.50
PMSRC (LP)	1.33	0.91	5.20	63.95	1.39	1.12	1.27	80.15
	High explained Gini coefficient ($\text{Gi}_{Y X} = 0.25$)							
PLR-SCAD (BIC)	2.41	0.20	0.23	11.75	2.38	0.22	0.11	13.85
PLR-SCAD (Bootstrap)	2.29	0.17	13.05	2.35	2.26	0.22	3.50	5.35
PLR-LASSO (BIC)	2.52	0.24	1.92	10.95	2.76	0.34	0.55	19.95
PLR-LASSO (Bootstrap)	2.26	0.20	23.90	1.40	2.22	0.29	5.92	5.80
PMSRC (FABS)	2.40	0.19	4.92	4.75	2.76	0.29	0.62	15.60
PMSRC (LP)	2.88	0.29	1.80	16.50	3.45	0.37	0.50	26.30

Table 12
Comparison of the estimation procedure – Student distribution and Setup 2 with $n \in \{50, 200\}$.

	$n = 50$				$n = 200$			
	$\text{Gi}_{Y,X}$	θ_0	FPR	FNR	$\text{Gi}_{Y,X}$	θ_0	FPR	FNR
PLR-SCAD (BIC)	2.18	0.70	1.79	43.95	1.15	0.20	0.51	6.9
PLR-SCAD (Bootstrap)	2.95	0.73	10.27	31.65	1.21	0.20	1.47	6.0
PLR-LASSO (BIC)	2.54	0.75	1.76	47.35	1.24	0.24	0.84	8.5
PLR-LASSO (Bootstrap)	3.03	0.77	13.31	32.15	1.22	0.26	2.09	7.5
PMSRC (FABS)	4.13	0.81	0.64	59.35	1.30	0.24	0.84	9.30
PMSRC (LP)	4.95	0.89	0.34	67.90	1.59	0.29	0.41	19.10

Appendix D: Supplementary results concerning the real data example

In this section, we provide supplementary results concerning the real data example of Section 5. Figure 9 displays the traceplot obtained with the FABS algorithm. Tables 13 and 14 show the model fits and estimated coefficients obtained when the regularization parameter is chosen via BIC, see Tables 6 and 7 for their equivalent when the bootstrap method is used instead.

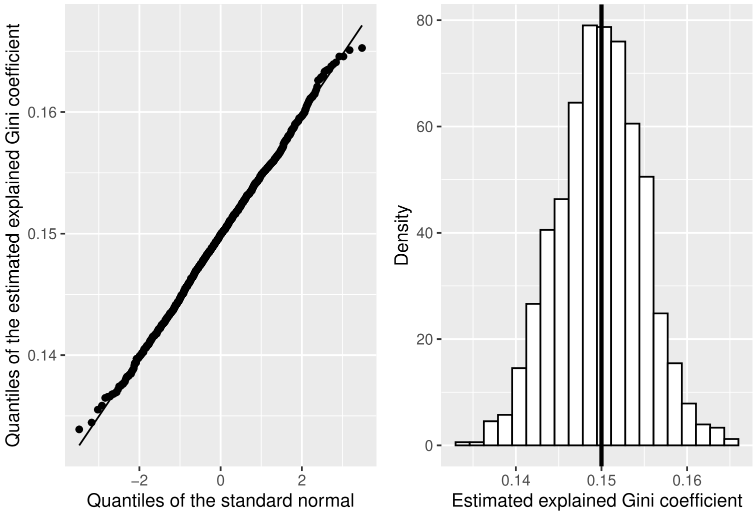

FIG 8. *Asymptotic Normality of* $\widehat{Gi}_{Y,X}$ *obtained with the PLR-SCAD on Setup 1, samples of size* 1000.

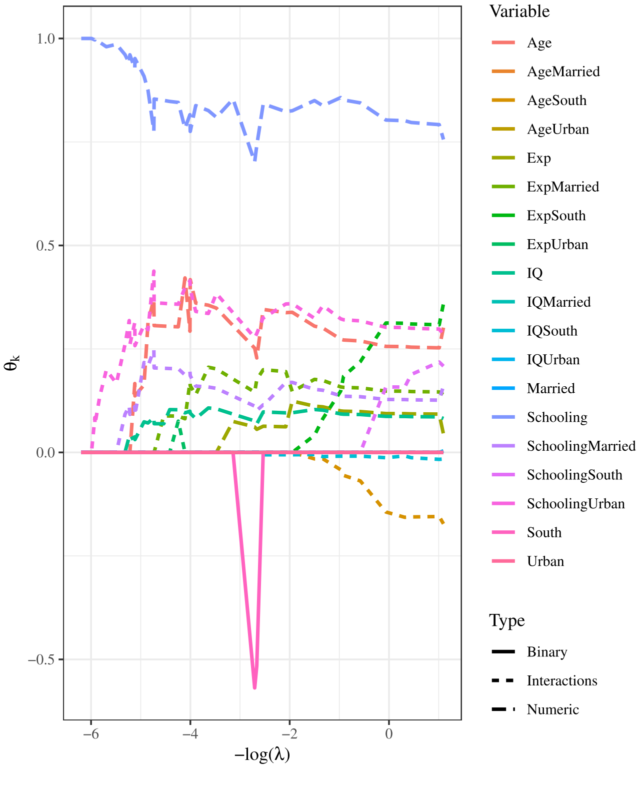

FIG 9. *Trace plot of the FABS on the Griliches data.*

Table 13

Results of the PLR-SCAD and PLR-LASSO on the Griliches data (BIC selection).

		PLR-SCAD		PLR-LASSO
	$(n_{\text{fwd}} = 5)$	$(n_{\text{fwd}} = 20)$	$(n_{\text{fwd}} = 50)$	
$\widehat{\text{Gi}}_{Y,X}$	10.19	10.33	10.31	10.30
95% CI	[8.52, 11.87]	[8.69, 11.97]	[8.66, 11.96]	[8.58, 12.02]
# variables	5	6	6	6
BIC score	−2.305	−2.296	−2.299	−2.299

Table 14

Coefficients estimated by the PLR-SCAD and PLR-LASSO on the Griliches data (BIC selection).

		PLR-SCAD		PLR-LASSO
	$(n_{\text{fwd}} = 5)$	$(n_{\text{fwd}} = 20)$	$(n_{\text{fwd}} = 50)$	
Age	0.257	0.357	0.304	0.305
Schooling	0.930	0.829	0.848	0.850
IQ	0.058	0.081	0.103	0.069
SchoolingMarried	/	0.079	0.203	0.203
SchoolingUrban	/	0.393	0.359	0.360
ExpMarried	/	0.138	0.088	0.088
ExpUrban	0.256	/	/	/
IQMarried	0.028	/	/	/

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

Computational resources have been provided by the supercomputing facilities of the UCLouvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11.

References

[1] Bourguignon, F., Ferreira, F. H. G. and Menéndez, M. (2007). Inequality of Opportunity in Brazil. *Review of Income and Wealth* **53** 585–618.

[2] Brunori, P., Hufe, P. and Mahler, D. G. (2018). The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees Working Paper, World Bank, Washington, DC, available at SSRN: https://ssrn.com/abstract=3127234. https://doi.org/10.1596/1813-9450-8349

[3] Cavanagh, C. and Sherman, R. P. (1998). Rank Estimators for Monotonic Index Models. *Journal of Econometrics* **84** 351–381. MR1630210

[4] Cowell, F. A. and Flachaire, E. (2015). Chapter 6 – Statistical Methods for Distributional Analysis. In *Handbook of Income Distribution*, (A. B. Atkinson and F. Bourguignon, eds.). *Handbook of Income Distribution* **2** 359-465. Elsevier. https://doi.org/10.1016/B978-0-444-59428-0.00007-2

[5] Escanciano, J. C. and Terschuur, J. R. (2023). Debiased Semiparametric U-Statistics: Machine Learning Inference on Inequality of Opportunity.

[6] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association* **96** 1348–1360. https://doi.org/10.1198/016214501753382273. MR1946581

[7] Ferreira, F. H. G. and Gignoux, J. (2011). The Measurement of Inequality of Opportunity: Theory and an Application to Latin America. *Review of Income and Wealth* **57** 622–657. https://doi.org/10.1111/j.1475-4991.2011.00467.x

[8] Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise Coordinate Optimization. *The Annals of Applied Statistics* **1** 302–332. https://doi.org/10.1214/07-AOAS131. MR2415737

[9] Han, A. K. (1987). Non-Parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator. *Journal of Econometrics* **35** 303–316. https://doi.org/10.1016/0304-4076(87)90030-3. MR0903188

[10] Heuchenne, C. and Jacquemain, A. (2022). Inference for Monotone Single-Index Conditional Means: A Lorenz Regression Approach. *Computational Statistics & Data Analysis* **167** 107347. MR4317417

[11] Lin, H. and Peng, H. (2013). Smoothed Rank Correlation of the Linear Transformation Regression Model. *Computational Statistics & Data Analysis* **57** 615–630. https://doi.org/10.1016/j.csda.2012.07.012. MR2981113

[12] Mack, Y. P. and Müller, H.-G. (1989). Derivative Estimation in Nonparametric Regression with Random Predictor Variable. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)* **51** 59–72. MR1065559

[13] Mack, Y. P. and Silverman, B. W. (1982). Weak and Strong Uniform Consistency of Kernel Regression Estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **61** 405–415. https://doi.org/10.1007/BF00539840. MR0679685

[14] Roemer, J. E. and Trannoy, A. (2015). Chapter 4 – Equality of Opportunity. In *Handbook of Income Distribution*, (Atkinson, Antony, Bourguignon and François, eds.). *Handbook of Income Distribution* **2** 217–300. Elsevier.

[15] Shi, X., Huang, Y., Huang, J. and Ma, S. (2018). A Forward and Backward Stagewise Algorithm for Nonconvex Loss Functions with Adaptive Lasso. *Computational Statistics & Data Analysis* **124** 235–251. https://doi.org/10.1016/j.csda.2018.03.006. MR3787624

[16] Silverman, B. W. (1978). Weak and Strong Uniform Consistency of the

Kernel Estimate of a Density and Its Derivatives. *The Annals of Statistics* **6** 177–184. MR0471166

[17] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. MR1652247

[18] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics. Springer Series in Statistics*. Springer New York, NY. https://doi.org/10.1007/978-1-4757-2545-2_15. MR1385671

[19] Yamato, H. (1973). Uniform Convergence of an Estimator of a Distribution Function. *Bulletin of Mathematical Statistics* **15** 69–78. https://doi.org/info:doi/10.5109/13073. MR0329113

[20] Zhang, C.-H. and Zhang, S. S. (2014). Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **76** 217–242. MR3153940