Consistency and robustness properties of the S-nonnegative garrote estimator

Peer-reviewed author version

Gijbels, Irène; VERHASSELT, Anneleen & Vrinssen, Inge (2017) Consistency and robustness properties of the S-nonnegative garrote estimator. In: STATISTICS, 51 (4), p. 921-947.

DOI: 10.1080/02331888.2017.1318879
Handle: http://hdl.handle.net/1942/23407

Consistency and robustness properties of the S-nonnegative garrote estimator

I. Gijbels[1], A. Verhasselt[2] and I. Vrinssen[1]

1. KU Leuven, Department of Mathematics and Leuven Statistics Research Center (LStat), Leuven, Belgium.

2. Universiteit Hasselt, Interuniversity Institute for Biostatistics and statistical Bioinformatics, CenStat, Hasselt, Belgium.

Abstract

This paper concerns a robust variable selection method in multiple linear regression: the robust S-nonnegative garrote variable selection method. In this paper the consistency of the method, both in terms of estimation and in terms of variable selection, is established. Moreover, the robustness properties of the method are further investigated by providing a lower bound for the breakdown point, and by deriving the influence function. The provided expressions nicely reveal the impact that the choice of an initial estimator has on the robustness properties of the variable selection method. Illustrative examples of influence functions for the S-nonnegative garrote as well as for the original (non-robust) nonnegative garrote variable selection method are provided.

Key words and Phrases: breakdown point, consistency, influence function, nonnegative garrote, ordinary least-squares estimator, outliers, S-estimation, variable selection.

1 Introduction

Ordinary least squares regression is often used to fit a linear model. When many variables are measured, these models become difficult to interpret. To improve the interpretability of such a model, variable selection methods were introduced. A possible approach for variable selection is to add a penalty term on the regression coefficients to the objective function of ordinary least squares regression. For example, the Bridge (Frank and Friedman, 1993; Fu, 1998) and the Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani, 1996) both use an L_q-type of penalty on the regression coefficients, with $q < 1$ and $q = 1$ respectively. The Smoothly Clipped Absolute Deviation (SCAD) penalty is used by Fan and Li (2001). This penalty function g_λ satisfies $g_\lambda(0) = 0$ and its first-order derivative is given by

$$g'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\},$$

for some $a > 2$ and $\theta > 0$, and with $\lambda > 0$ a regularization parameter. Herein $I(A)$ denotes the indicator function, i.e. $I(A) = 1$ if A holds, and 0 if A does not hold. Another approach consists of the nonnegative garrote method proposed by Breiman (1995). Here one firstly computes the ordinary least squares estimator (OLS), and then shrinks or puts some coefficients of the OLS equal to zero.

The main disadvantage of these methods is that they are not robust to outliers. Therefore, robust versions of the LASSO, SCAD and nonnegative garrote method are proposed in the literature. For the LASSO, different robust alternatives, such as the LAD-LASSO (Wang et al., 2007), the WLAD-LASSO (Arslan, 2012) and the Sparse LTS (Alfons et al., 2013), have been developed in the literature. The LAD-LASSO is a penalized least absolute deviation estimator

that is consistent in estimation and variable selection, but it is not robust to outliers that are also outliers in the covariates (i.e. leverage points). Therefore, the WLAD-LASSO is proposed. This method applies the LAD-LASSO to the data set in which each observation is weighted using weights computed with robust distances. Arslan (2012) proved that the WLAD-LASSO is also consistent in estimation and variable selection. The Sparse LTS is a trimmed version of LASSO, that is robust to vertical outliers and leverage points. Its breakdown point is computed in Alfons et al. (2013) and Öllerer et al. (2015) derived its influence function. A robust version of the SCAD that is consistent in estimation and variable selection, is proposed by Wang and Li (2009). Wang et al. (2013) proposed a penalized robust regression estimator based on the exponential squared loss function, where the penalty function can be of any type. They also proved that this method is consistent in estimation and variable selection and they computed its breakdown point and influence function. In a mean shift regression model with normal errors Xiong and Joseph (2013) consider regression with outlier shrinkage. The computational complexity of such an estimator is comparable to that of an LTS estimator.

Since the theoretical properties of the nonnegative garrote method are well studied in the literature (Yuan and Lin, 2007) and are extended to variable selection in additive regression models and varying coefficient models by Antoniadis et al. (2012a,b), Gijbels and Vrinssen (2015) investigated different robust versions of this variable selection method, among others the S-nonnegative garrote method. An extensive simulation study shows that the S-nonnegative garrote method performs quite well, also in comparison with competitors.

In this paper we provide some theoretical properties of the S-nonnegative garrote method. In Section 2 we state the model assumptions and briefly explain the S-nonnegative garrote. Section 3 establishes oracle properties for this method. In Section 4 we prove that the S-nonnegative garrote method is consistent in variable selection and estimation. In Section 5 its breakdown point is established and in Section 6 we derive its influence function. Some illustrations regarding the influence function are provided. The proofs of the theoretical results are deferred to Section 7.

2 Robust nonnegative garrote variable selection procedure

Consider a multiple linear regression model

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i, \tag{1}$$

with $(X_{i1}, \ldots, X_{ip}, Y_i)$, $i = 1, \ldots, n$, independent and identically distributed observations from $(X_1, \ldots, X_p, Y)$, satisfying the model $Y = \sum_{j=1}^{p} X_j\beta_j + \varepsilon$, where Y is the response, $\boldsymbol{\mathcal{X}} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is a vector with the p covariates with $\mathbf{A}^{\mathrm{T}}$ denoting the transpose of a matrix or vector $\mathbf{A}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of unknown regression coefficients and ε is the error term with mean 0 and variance σ^2. We denote $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$, for $i = 1, \ldots, n$, and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{\mathrm{T}}$. An intercept is included in the model by setting all elements in the first column of $\mathbf{X}$ equal to one. We further assume that the model is sparse, i.e. $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\beta}_{\mathcal{N}}^{\mathrm{T}}\right)^{\mathrm{T}}$ with $\mathcal{S}$ denoting the set of indices containing the non-zero ("to be selected") regression coefficients, $\mathcal{S} = \{j : \beta_j \neq 0\}$ and $\mathcal{N}$ denoting the set of indices containing the zero ("not to be selected") regression coefficients, $\mathcal{N} = \{j : \beta_j = 0\}$. Throughout this paper we will also use this notation to partition other vectors and matrices into vectors (or matrices) related to the non-zero and zero regression coefficients, e.g. $\mathbf{X} = (\mathbf{X}_{\mathcal{S}}, \mathbf{X}_{\mathcal{N}})$, where $\mathbf{X}_{\mathcal{S}}$ and $\mathbf{X}_{\mathcal{N}}$ contain the columns of $\mathbf{X}$ related to $\boldsymbol{\beta}_{\mathcal{S}}$ and $\boldsymbol{\beta}_{\mathcal{N}}$ respectively.

As explained in Section 1, the original nonnegative garrote method of Breiman (1995) starts from an initial estimator, for example the ordinary least squares estimator, and then it shrinks

or puts some coefficients β_j of the initial estimator equal to zero using the nonnegative garrote shrinkage factors. More precisely, let $\widehat{\beta}_j^{\text{OLS}}$ denote the initial least squares estimator of the coefficient β_j, the nonnegative garrote shrinkage factors $\widehat{\mathbf{c}} = (\widehat{c}_1, \ldots, \widehat{c}_p)^{\text{T}}$ are found by solving

$$\begin{cases} \widehat{\mathbf{c}} = \underset{\mathbf{c}}{\text{argmin}} \left\{ \dfrac{1}{2n} \displaystyle\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} c_j \widehat{\beta}_j^{\text{OLS}} X_{ij} \right)^2 + \lambda_n \sum_{j=1}^{p} c_j \right\} \\ \text{s.t. } c_j \geq 0 \ (j = 1, \ldots, p), \end{cases} \tag{2}$$

where $\mathbf{c} = (c_1, \ldots, c_p)^{\text{T}}$ and for a given regularization parameter $\lambda_n > 0$. The nonnegative garrote estimator of the coefficient $\beta_j, j = 1, \ldots, p$, is then given by

$$\widehat{\beta}_j^{\text{NNG}} = \widehat{c}_j \widehat{\beta}_j^{\text{OLS}}.$$

In analogy with the original nonnegative garrote procedure, the S-nonnegative garrote method now starts from a robust initial estimator, such as the MM-estimator (Yohai, 1987) or the τ-estimator (Yohai and Zamar, 1988), and it uses the S-nonnegative garrote shrinkage factors to shrink or put some coefficients of this robust initial estimator equal to zero. Denote the initial estimator with $\widehat{\beta}_j^{\text{init}}$, $\widehat{\beta}_j^{\text{init}} X_{ij}$ with Z_{ij}, for $j = 1, \ldots, p$ and $i = 1, \ldots, n$, $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^{\text{T}}$ and $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\text{T}}$. The S-nonnegative garrote shrinkage factors $\widehat{\mathbf{c}} = (\widehat{c}_1, \ldots, \widehat{c}_p)^{\text{T}}$ are found by solving

$$\begin{cases} \widehat{\mathbf{c}} = \underset{\mathbf{c}}{\text{argmin}} \left\{ \widehat{\sigma}\left(\mathbf{r}(\mathbf{c})\right) + \lambda_n \displaystyle\sum_{j=1}^{p} c_j \right\} \\ \text{s.t. } c_j \geq 0 \ (j = 1, \ldots, p), \end{cases} \tag{3}$$

where $\widehat{\sigma}\left(\mathbf{r}(\mathbf{c})\right)$ solves the equation

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{Y_i - \sum_{j=1}^{p} c_j Z_{ij}}{\widehat{\sigma}\left(\mathbf{r}(\mathbf{c})\right)} \right) = b,$$

with $\mathbf{r}(\mathbf{c}) = (r_1, \ldots, r_n)^{\text{T}}$ with $r_i = Y_i - \sum_{j=1}^{p} c_j Z_{ij}$, $i = 1, \ldots, n$, ρ a loss function satisfying Assumption 4.2 in Section 4 and $b = E(\rho(Z))$, with Z standard normally distributed. The S-nonnegative garrote estimator of the coefficient $\beta_j, j = 1, \ldots, p$, is then given by

$$\widehat{\beta}_j^{\text{S-NNG}} = \widehat{c}_j \widehat{\beta}_j^{\text{init}}. \tag{4}$$

When the initial estimator $\widehat{\beta}_j^{\text{init}}$ is equal to zero, we also set the S-nonnegative garrote shrinkage factor $\widehat{c}_j$ equal to zero.

As explained in Gijbels and Vrinssen (2015) this optimization problem can be approximated by a weighted quadratic programming problem that suggests an iterative procedure. Let $\widehat{\mathbf{c}}^0$ be the current value of $\mathbf{c}$ in the iteration procedure. Then, the value of $\mathbf{c}$ in the next iteration step can be found by solving the optimization problem

$$\begin{cases} \widehat{\mathbf{c}} = \underset{\mathbf{c}}{\text{argmin}} \left\{ \frac{1}{2} \mathbf{c}^{\text{T}} \mathbf{Z}^{\text{T}} \mathbf{W}_S(\widehat{\mathbf{c}}^0) \mathbf{Z} \mathbf{c} - \left(\mathbf{Z}^{\text{T}} \mathbf{W}_S(\widehat{\mathbf{c}}^0) \mathbf{Y} - \frac{\lambda_n}{\omega_S(\widehat{\mathbf{c}}^0)} \mathbf{1}_p \right)^{\text{T}} \mathbf{c} \right\} \\ \text{s.t. } c_j \geq 0 \ (j = 1, \ldots, p), \end{cases} \tag{5}$$

where $\mathbf{W}_S(\mathbf{c}) = \text{diag}(W_{S,i}(\mathbf{c})) \in \mathbb{R}^{n \times n}$ with $W_{S,i}(\mathbf{c}) = \dfrac{\rho'(r_i(\mathbf{c})/\widehat{\sigma}(\mathbf{r}(\mathbf{c})))}{r_i(\mathbf{c})/\widehat{\sigma}(\mathbf{r}(\mathbf{c}))}$, $\omega_S(\mathbf{c}) = \dfrac{\widehat{\sigma}(\mathbf{r}(\mathbf{c}))}{\mathbf{r}^{\text{T}}(\mathbf{c}) \mathbf{W}_S(\mathbf{c}) \mathbf{r}(\mathbf{c})}$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\text{T}}$.

Based on Zou (2006), the S-nonnegative garrote method can also be reformulated as

$$\begin{cases} \widehat{\boldsymbol{\beta}}^{\text{S-NNG}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \widehat{\sigma}\left(\mathbf{r}(\boldsymbol{\beta})\right) + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\widehat{\beta}_j^{\text{init}}|} \right\} \\ \text{s.t. } \beta_j \widehat{\beta}_j^{\text{init}} \geq 0 \ (j = 1, \ldots, p), \end{cases} \tag{6}$$

where $\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta}))$ solves

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{Y_i - \sum_{j=1}^{p} \beta_j X_{ij}}{\widehat{\sigma}\left(\mathbf{r}(\boldsymbol{\beta})\right)} \right) = b,$$

with $\mathbf{r}(\boldsymbol{\beta}) = (r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))^{\mathrm{T}}$ and $r_i(\boldsymbol{\beta}) = Y_i - \sum_{j=1}^{p} \beta_j X_{ij}$, $i = 1, \ldots, n$. If the initial estimator $\widehat{\beta}_j^{\text{init}}$ is equal to zero, we also set the S-nonnegative garrote estimator $\widehat{\beta}_j^{\text{S-NNG}}$ equal to zero.

Throughout this paper we will use the following matrix and vector norm. Let $\mathbf{A}$ be a matrix of size $m \times n$. The L_2-norm of $\mathbf{A}$ is defined as $\|\mathbf{A}\|_2 = \max_{\mathbf{x} \neq \mathbf{0}_n} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$, where $\mathbf{x}$ is a non-null vector of dimension $n \times 1$ and $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$ is the usual L_2-norm of a vector $\mathbf{x}$.

3 Oracle properties

In this section we establish the so-called oracle properties of the S-nonnegative garrote, more precisely of the minimizer of (5), for a given (fixed) $\widehat{\mathbf{c}}^0$. Oracle properties are related to looking at the situation that the set of non-zero coefficients is known, i.e. the set $\mathcal{S}$ is known, and one focuses on estimation of these non-zero coefficients. Recall that the (sub)vector of true non-zero regression coefficients is denoted by $\boldsymbol{\beta}_{\mathcal{S}}$ and denote the corresponding S-nonnegative garrote estimator by $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\text{S-NNG}}$. All asymptotic results in this paper are for fixed p number of covariates.

The oracle properties are derived from the fact that there is a close relation between the adaptive Lasso and the nonnegative garrote. Indeed, Zou (2006) shows that (2) is equivalent to solving

$$\begin{cases} \widehat{\boldsymbol{\beta}}^{\text{NNG}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\widehat{\beta}_j^{\text{OLS}}|} \right\} \\ \text{s.t. } \widehat{\beta}_j^{\text{OLS}} \beta_j \geq 0 \ (j = 1, \ldots, p), \end{cases} \tag{7}$$

where $\widehat{c}_j = \frac{\widehat{\beta}_j^{\text{NNG}}}{\widehat{\beta}_j^{\text{OLS}}}$.

Since weights $W_{S,i}(\widehat{\mathbf{c}}^0)$ are introduced in (5), this S-nonnegative garrote optimization problem is related to the adaptive Lasso optimization problem in heteroscedastic models (Wagener and Dette (2013)). Wagener and Dette (2013) consider a heteroscedastic linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \Sigma(\boldsymbol{\beta})\tilde{\boldsymbol{\varepsilon}}, \tag{8}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, and where the errors $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n)^{\mathrm{T}}$ satisfy $\mathrm{E}(\tilde{\varepsilon}_i) = 0$, $\mathrm{Var}(\tilde{\varepsilon}_i) = 1$ and $\Sigma(\boldsymbol{\beta})$ =diag$(\sigma(\mathbf{X}_1, \boldsymbol{\beta}), \ldots, \sigma(\mathbf{X}_n, \boldsymbol{\beta}))$, revealing the heteroscedasticity. The S-NNG optimization problem (5) is equivalent to an unweighted adaptive Lasso optimization problem (see the terminology used in Wagener and Dette (2013)) with extra constraint in a heteroscedastic

regression model:

$$\begin{cases} \widehat{\boldsymbol{\beta}}^{\text{S-NNG}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \dfrac{1}{2n} \displaystyle\sum_{i=1}^{n} \left(\dfrac{Y_i - \displaystyle\sum_{j=1}^{p} \beta_j X_{ij}}{W_{S,i}^{-1/2}(\widehat{\mathbf{c}}^0)} \right)^2 + \dfrac{\lambda_n}{\omega_S(\widehat{\mathbf{c}}^0)} \displaystyle\sum_{j=1}^{p} \dfrac{|\beta_j|}{|\widehat{\beta}_j^{\text{init}}|} \right\}, \\ \text{s.t. } \widehat{\beta}_j^{\text{init}} \beta_j \geq 0 \ (j = 1, \ldots, p). \end{cases} \tag{9}$$

Define $\bar{\mathbf{X}} = (\bar{\mathbf{X}}_{\mathcal{S}}, \bar{\mathbf{X}}_{\mathcal{N}}) = \mathbf{W}_S^{1/2}(\widehat{\mathbf{c}}^0)\mathbf{X}$, $\bar{\mathbf{Y}} = \mathbf{W}_S^{1/2}(\widehat{\mathbf{c}}^0)\mathbf{Y}$ and denote by $\bar{X}_{ij}$ the (i,j)-th element of the matrix $\bar{\mathbf{X}}$. With these notations, it can be seen that our model context leads to $\bar{\mathbf{Y}} = \bar{\mathbf{X}}\boldsymbol{\beta} + \mathbf{W}_S^{1/2}(\widehat{\mathbf{c}}^0)\sigma\tilde{\boldsymbol{\varepsilon}}$ which is of the form (8). Furthermore, let $\mathbf{C}_{\mathcal{SS}} = \frac{1}{n}\bar{\mathbf{X}}_{\mathcal{S}}^{\mathrm{T}}\bar{\mathbf{X}}_{\mathcal{S}}$.

In order to prove the oracle properties of the S-NNG, we need the following assumptions.

Assumption 3.1.

1. $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \bar{X}_{ij}^2 = 1$, *for all* $j = 1, \ldots, p$.

2. *There exists a constant* $B > 0$ *such that the initial estimator* $\widehat{\boldsymbol{\beta}}^{init}$ *satisfies*

$$\lim_{n\to\infty} P(B \min_{j\in\mathcal{S}} |\widehat{\beta}_j^{init}| < \min_{j\in\mathcal{S}} |\beta_j|) = 0,$$

3. *There exists a sequence* $r_n \to \infty$ *such that the initial estimator satisfies*

$$\lim_{n\to\infty} P(\max_{j\in\mathcal{N}} |\widehat{\beta}_j^{init}| \geq \frac{1}{r_n}) = 0.$$

4. *The sequences* λ_n *and* r_n *satisfy*

   - $\frac{\lambda_n}{\omega_S(\widehat{\mathbf{c}}^0)\sqrt{n}} \to$ *constant* $\in \mathbb{R}$, *as* $n \to \infty$,
   - $\frac{\ln n \ \sqrt{n}\,\omega_S(\widehat{\mathbf{c}}^0)}{\lambda_n r_n} \to 0$, *as* $n \to \infty$.

5. *There exists constants* κ_1 *and* κ_2, *such that*

$$0 < \kappa_1 \leq \lambda_{\min}(\mathbf{C}_{\mathcal{SS}}) \leq \lambda_{\max}(\mathbf{C}_{\mathcal{SS}}) \leq \kappa_2 < \infty,$$

   where $\lambda_{\min}(\mathbf{C}_{\mathcal{SS}})$ *and* $\lambda_{\max}(\mathbf{C}_{\mathcal{SS}})$ *are the smallest and largest eigenvalue of* $\mathbf{C}_{\mathcal{SS}}$ *respectively.*

6. *There exists a constant* $\bar{\sigma}$ *such that* $0 < W_{S,i}^{-1}(\widehat{\mathbf{c}}^0) \leq \bar{\sigma} < \infty$, *for all* $i = 1, \ldots, n$.

7. $\frac{\lambda_n}{\omega_S(\widehat{\mathbf{c}}^0)\sqrt{n}} \to 0$, *as* $n \to \infty$.

8. $\frac{1}{n}\max_{1\leq i\leq n} \|\bar{\mathbf{X}}_{\mathcal{S},i}\|_2^2 = \frac{1}{n}\max_{1\leq i\leq n} \left(W_{S,i}(\widehat{\mathbf{c}}^0)\|\mathbf{X}_{\mathcal{S},i}\|_2^2\right) \to 0$, *where* $\bar{\mathbf{X}}_{\mathcal{S},i}^{\mathrm{T}}$ *is the* i*-th row of* $\bar{\mathbf{X}}_{\mathcal{S}}$, *as* $n \to \infty$.

The next theorem states the sign consistency of the S-nonnegative garrote estimator, in the sense that

$$\lim_{n\to\infty} P(\widehat{\boldsymbol{\beta}}^{\text{S-NNG}} =_s \boldsymbol{\beta}) = 1,$$

where $\widehat{\boldsymbol{\beta}}^{\text{S-NNG}} =_s \boldsymbol{\beta}$ means that each component of $\widehat{\boldsymbol{\beta}}^{\text{S-NNG}}$ has the same sign as the corresponding component of $\boldsymbol{\beta}$. Since the sign of 0 is defined as 0, sign consistency implies variable selection consistent (in the sense of Theorem 4.1.2).

Theorem 3.1. *Under Assumptions 3.1.1-3.1.6,* **the S-nonnegative garrote estimator $\widehat{\boldsymbol{\beta}}^{\text{S-NNG}}$ obtained from (4) using the minimizer of (5),** *is sign consistent for $\boldsymbol{\beta}$.*

Proof. This theorem immediately follows from Theorem 4.1 of Wagener and Dette (2013) and the equivalence between the nonnegative garrote and the adaptive Lasso with extra sign constraint in (9). Wagener and Dette (2013) give the sign consistency of the adaptive Lasso in a heteroscedastic model.

Furthermore we have to verify that the extra sign constraint in (9) is satisfied, with probability tending to 1. The proof of the latter is similar to the proof of Corollary 2 in Zou (2006). □

The asymptotic normality oracle property (Theorem 3.2) follows from Theorem 4.2 in Wagener and Dette (2013) for the adaptive Lasso in a heteroscedastic model. It establishes the asymptotic normality result of the estimated parameters, restricted to the true non-zero ones (i.e. in a restricted parameter space).

Theorem 3.2. *Let Assumption 3.1 hold. Then for all $\bar{\boldsymbol{\alpha}}_n \in \mathbb{R}^{|\mathcal{S}|}$ (where $|\mathcal{S}|$ is the size of $\mathcal{S}$) with $\|\bar{\boldsymbol{\alpha}}_n\|_2 = 1$, the following holds*

$$\frac{\sqrt{n}}{s_n}\bar{\boldsymbol{\alpha}}_n^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{S\text{-}NNG} - \boldsymbol{\beta}_{\mathcal{S}}) \xrightarrow{D} N(0,1), \qquad as\, n \to \infty,$$

where $s_n^2 = \frac{1}{n}\sigma\bar{\boldsymbol{\alpha}}_n^{\mathrm{T}} C_{\mathcal{S}\mathcal{S}}^{-1}\bar{\mathbf{X}}_{\mathcal{S}}^{\mathrm{T}}\mathbf{W}_S^{1/2}(\widehat{\mathbf{c}}^0)\bar{\mathbf{X}}_{\mathcal{S}}\mathbf{C}_{\mathcal{S}\mathcal{S}}^{-1}\bar{\boldsymbol{\alpha}}_n$.

4 Consistency

In this section we establish that the S-nonnegative garrote estimator is consistent in estimation and variable selection. The latter means that the estimator tends to estimate a true-zero as a zero. **The results in this section complement these of Section 3, where the essence (in particular in Theorem 3.2) is that the set of non-zero coefficients $\mathcal{S}$ is known (the oracle situation). In reality however the set $\mathcal{S}$ is not known. Note that Theorem 3.1 provides the variable selection consistency of the S-nonnegative garrote estimator, but under the restricted setting of a given $\widehat{\mathbf{c}}^0$ (see the assumptions). An obvious good choice for this initial vector would be a vector containing one's (respectively zero's) at positions of non-zero (respectively zero) true coefficients, a knowledge that is available when the set of true (non-zero) coefficients is known.**

In this section we also obtain the variable selection consistency of the S-nonnegative garrote estimator, but under a different (and more realistic) setting. In Section 3 the emphasis was on establishing an asymptotic normality result for the oracle estimator, whereas in this section the main goal is to establish the estimation consistency of the S-nonnegative garrote estimator, including its rate of convergence. As such the results in Sections 3 and 4 are complementary.

We need the following assumptions on the data:

Assumption 4.1.

1. *The matrix $\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}$ is invertible.*
2. *There exists $M > 0$ such that $|X_{ij}| < M$ for all $j = 1, \ldots, p$, $i = 1, \ldots, n$.*
3. *$\varepsilon_i = O(1)$, almost surely (with probability one), as $n \to \infty$, for all $i = 1, \ldots, n$.*
4. *$P(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta} = 0,\ \forall i = 1, \ldots, n) < 0.5$ for all $\boldsymbol{\theta} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$, where $\mathbf{0}_p$ denotes the null vector of dimension p.*

5. $P(\alpha Y_i + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta} = 0,\ \forall i = 1, \ldots, n) < 0.5$ *for all* $\alpha \in \mathbb{R}$ *and* $\boldsymbol{\theta} \in \mathbb{R}^p$ *for which* $|\alpha| + \|\boldsymbol{\theta}\|_2 \neq 0$.

Assumption 4.1.4 implies that the random variables $\mathbf{X}_i$ may not be too concentrated on any subspace of $\mathbb{R}^p$. Assumption 4.1.5 is needed to avoid solutions with $\widehat{\sigma}(\mathbf{r}(\mathbf{c})) = 0$ (see Maronna and Yohai (1981)). For the loss function we need the following assumptions:

Assumption 4.2. *Let* $\rho : \mathbb{R} \to \mathbb{R}$ *be a real function satisfying the following assumptions:*

1. ρ *is symmetric, continuously differentiable and* $\rho(0) = 0$.
2. *There exists* $d > 0$ *such that* ρ *is strictly increasing on* $[0, d]$ *and constant on* $[d, \infty)$ *and such that* $0 < \rho(d) = a < +\infty$.
3. *Denoting* $\psi(u) = \rho'(u)$, *then* $\psi(u)/u$ *is nonincreasing for* $u > 0$ *and* $0 \leq \psi(u)/u \leq 1$.

A loss function that satisfies these assumptions is for example Tukey's biweight loss function

$$\rho_d(x) = \begin{cases} \frac{d^2}{6}\left(1 - \left(1 - \left(\frac{x}{d}\right)^2\right)^3\right) & \text{if } |x| \leq d, \\ \frac{d^2}{6} & \text{if } |x| > d. \end{cases} \tag{10}$$

Theorem 4.1. *Suppose Assumptions 4.1 and 4.2 hold and assume that the initial estimator* $\widehat{\boldsymbol{\beta}}^{\text{init}}$ *is strongly consistent with rate* κ_n, *i.e.*

$$\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\text{init}}\|_2 = O(\kappa_n), \qquad \text{as } n \to \infty,$$

with probability 1, *for some* $\kappa_n \to 0$. *If* λ_n *tends to 0 in a fashion such that* $\kappa_n = o(\lambda_n)$ *and* $n\lambda_n = O(1)$, *then there exists a minimizer of* (5), $\widehat{\boldsymbol{\beta}}^{\text{S-NNG}}$, *such that*

1. $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\text{S-NNG}}\|_2 = O(\lambda_n)$, *as* $n \to \infty$, *with probability* 1.
2. $P\left(\widehat{\beta}_j^{\text{S-NNG}} \neq 0\right) \to 0$, *as* $n \to \infty$, *for any* $j \in \mathcal{N}$.

This theorem states that the S-nonnegative garrote estimator is strongly consistent with rate λ_n. The S-nonnegative garrote estimator has a lower rate of convergence than the initial estimator, but it has the advantage that it estimates the true-zero coefficients as zero. Suppose that the initial estimator is strongly consistent with convergence rate $\kappa_n = n^{-\eta}$, with $\eta > 0$, and that the regularization parameter λ_n is of order $n^{-\xi}$, with $0 < \xi < \min(\eta, 1)$. Then, the S-nonnegative garrote estimator is strongly consistent with rate $n^{-\xi}$. A possible initial estimator for the S-nonnegative garrote method is the ordinary least squares estimator. Chatterjee and Lahiri (2011) proved that, if $E|\varepsilon|^{\zeta} < \infty$ for $1 < \zeta < 2$, the ordinary least squares estimator is strongly consistent with convergence rate $n^{-(\zeta-1)/\zeta}$. Hence, taking values for the regularization parameter λ_n of order $n^{-\xi}$ with $0 < \xi < (\zeta - 1)/\zeta$ results in a strongly consistent S-nonnegative garrote estimator. But using as initial estimator the ordinary least squares estimator would not lead to a robust procedure. A robust method is obtained by using for example as initial estimator the least median absolute estimator, proposed in Ip et al. (2003) for which a uniform strong consistency result was established. As shown in the latter paper this estimator is strongly consistent with rate $O(n^{-1/4}\sqrt{\ln n})$. Taking λ_n of order $n^{-\xi}$ with $0 < \xi < 1/4$ this leads to a strongly consistent and robust S-nonnegative garrote estimator.

Note that the conditions for the variable selection consistency in the second item of Theorem 4.1, are different from these under which variable selection consistency was obtained in Section 3. Indeed in that section, assumptions are formulated in the situation that $\widehat{\mathbf{c}}^0$ and the set $\mathcal{S}$ are given. Consequently, for example, an assumption on the invertibility of $\frac{1}{n}\bar{\mathbf{X}}_{\mathcal{S}}^{\mathrm{T}}\bar{\mathbf{X}}_{\mathcal{S}}$ is needed in Section 3 (see Assumption **3.1.5**), whereas an assumption on the invertibility of $\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}$ is needed in the current section (see Assumption **4.1.1**). Another example is the boundedness imposed on $\bar{X}_{ij}$ (see Assumption **3.1.1**) in Section 3, opposed to the boundedness assumption on X_{ij} (in Assumption **4.1.2**) in the current section.

5 Breakdown point

Let $P_i = \left(\mathbf{X}_i^{\mathrm{T}}, Y_i\right)$ and $\mathbf{P}_n = \{P_1, \ldots, P_n\}$. Assume that $\mathbf{P}_n$ is obtained by adding m arbitrary data points (outliers) to the original uncontaminated sample of size $n - m$. Without loss of generality, assume that the m outliers are the first m observed points, i.e. denox te the outliers by $\mathbf{P}_m = \{P_1, \ldots, P_m\}$ and the original sample by $\mathbf{P}_{n-m} = \{P_{m+1}, \ldots, P_n\}$. The fraction of outliers in $\mathbf{P}_n$ is $\frac{m}{n}$. Let $\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\beta}}(\mathbf{P}_n)$ denote a regression estimator based on the sample $\mathbf{P}_n$. The finite sample breakdown point (Donoho and Huber, 1983) of an estimator is defined as

$$\mathrm{BP}(\widehat{\boldsymbol{\beta}}_n, \mathbf{P}_{n-m}) = \min\left\{\frac{m}{n} : \sup_{\mathbf{P}_m} \|\widehat{\boldsymbol{\beta}}(\mathbf{P}_n) - \widehat{\boldsymbol{\beta}}(\mathbf{P}_{n-m})\|_2 = \infty\right\}.$$

Let

$$a_{nm} = \frac{a_*}{n-m} = \frac{1}{n-m} \max_{\boldsymbol{\theta} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}} \#\{i : m+1 \le i \le n \text{ and } \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta} = 0\},$$

with a_* the maximum number of $\mathbf{X}_i$, $i = m+1, \ldots, n$, lying on the same subspace. If there are two covariates, a_* is the maximum number of $\mathbf{X}_i$'s lying on the same line. When the carriers $\mathbf{X}_i$ are in general position, which means that no more than p of the carriers $\mathbf{X}_i$ lie on a hyperplane of $\mathbb{R}^p$, then $a_* = p$. For example, in the case of two covariates, the carriers are in general position if there are no more than 2 $\mathbf{X}_i$'s on the same line and thus $a_* = 2$.

In the sequel we will use a fixed value for λ_n, namely take $\lambda_n = \lambda$. A lower bound for the breakdown point of the S-nonnegative garrote estimator is provided in Theorem 5.1.

Theorem 5.1. *Suppose that Assumptions 4.1.4–4.1.5 and 4.2 are satisfied, that* $b/a = 0.5$, $a_{nm} < 0.5$ *and* $0 < \lambda < +\infty$. *Then, if* $\widehat{\boldsymbol{\beta}}^{\text{init}}$ *is the initial estimator of* $\boldsymbol{\beta}$, *we have*

$$\mathrm{BP}(\widehat{\boldsymbol{\beta}}^{\text{S-NNG}}, \mathbf{P}_{n-m}) \ge \min\left\{\mathrm{BP}(\widehat{\boldsymbol{\beta}}^{\text{init}}, \mathbf{P}_{n-m}),\ \frac{1-2a_{nm}}{2-2a_{nm}}\right\}.$$

Regarding the assumption on the loss function in Theorem 5.1 consider, for example, Tukey's biweight loss function in (10) for which $a = d^2/6$. A common choice for robust scale estimation is $d = 1.547$ (see for example Maronna et al. (2006)) which, with the constraint that $b/a = 0.5$, leads to $b = 0.1994341$.

From Theorem 5.1 we get that, if for example the ordinary least squares estimator (which has breakdown point $1/n$) is used as initial estimator, then the lower bound of the breakdown point of the S-nonnegative garrote estimator is at most $1/n$. This means that this S-nonnegative garrote estimator may not be robust to outliers. But, if we use a more robust estimator, such as the S-estimator which has a breakdown point of $\left(\left[\frac{n}{2}\right] - p + 2\right)/n$, then the breakdown point of this S-nonnegative garrote estimator is asymptotically at least 50% when the carriers are in general position.

6 Influence function

In this section we derive the influence function of the S-nonnegative garrote estimator, for any given $\lambda > 0$. Denote the cumulative distribution function of $(\boldsymbol{\mathcal{X}}^{\mathrm{T}}, Y)$ by F. In order to obtain the influence function we first introduce the functional form of the S-nonnegative garrote estimator. Like for the sample level (see optimization problem (3)), the functional form of the S-nonnegative garrote estimator $\left(\left(\boldsymbol{\beta}^{\text{S-NNG}}(F)\right)^{\mathrm{T}}, \sigma(F)\right) = \left(\beta_1^{\text{S-NNG}}(F), \ldots, \beta_p^{\text{S-NNG}}(F), \sigma(F)\right)$ is obtained by shrinking or putting some components of the functional form of the initial estimator equal to zero by using the functional form of the S-nonnegative garrote shrinkage factors. Denote

the functional form of the initial estimator with $\boldsymbol{\beta}^{\text{init}}(F) = \left(\beta_1^{\text{init}}(F), \ldots, \beta_p^{\text{init}}(F)\right)^{\mathrm{T}}$ and let $\boldsymbol{\mathcal{Z}}(F) = (Z_1(F), \ldots, Z_p(F))^{\mathrm{T}}$ where $Z_j(F) = \beta_j^{\text{init}}(F)X_j$ for $j = 1, \ldots, p$. The functional form of the S-nonnegative garrote shrinkage factors $\mathbf{c}^{\text{S-NNG}}(F) = \left(c_1^{\text{S-NNG}}(F), \ldots, c_p^{\text{S-NNG}}(F)\right)^{\mathrm{T}}$ can be found by minimizing

$$\begin{aligned} & S + \lambda \sum_{j=1}^{p} c_j \\ & \text{s.t. } c_j \geq 0 \ (j = 1, \ldots, p), \end{aligned} \tag{11}$$

for $(\mathbf{c}, S) \in (\mathbb{R}_+^p \times \mathbb{R}_+ \setminus \{0\})$, where S solves

$$\int \rho\left(\frac{Y - \boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\mathbf{c}}{S}\right) \mathrm{d}F(\boldsymbol{\mathcal{X}}^{\mathrm{T}}, Y) = b. \tag{12}$$

The functional form of the S-nonnegative garrote estimator for the coefficients β_j, $j = 1, \ldots, p$, is then given by

$$\beta_j^{\text{S-NNG}}(F) = c_j^{\text{S-NNG}}(F)\beta_j^{\text{init}}(F).$$

To simplify the notation, we will use $\mathrm{d}F$ for $\mathrm{d}F(\boldsymbol{\mathcal{X}}^{\mathrm{T}}, Y)$ in the sequel. Further, we assume that the functional of the S-nonnegative garrote shrinkage factors and the S-nonnegative garrote estimator is continuous in F. If F is the empirical distribution function corresponding to the sample $\mathbf{P}_n$, this optimization problem is equivalent to problem (3).

The influence function of a functional T at a distribution F measures the effect on T of an infinitesimal contamination at a single point. If we denote the point mass at $P_0 = (\mathbf{X}_0^{\mathrm{T}}, Y_0)$ with $\mathbf{X}_0 = (X_{01}, \ldots, X_{0p})^{\mathrm{T}}$ by δ_{P_0} and consider the contaminated distribution $F_{\epsilon,P_0} = (1-\epsilon)F + \epsilon\delta_{P_0}$ with $0 < \epsilon < 1$, then the influence function is given by

$$\mathrm{IF}(P_0, T, F) = \lim_{\epsilon \to 0} \frac{T(F_{\epsilon,P_0}) - T(F)}{\epsilon} = \frac{\partial}{\partial \epsilon}\, T(F_{\epsilon,P_0})|_{\epsilon=0}\,.$$

The expression for the influence function of the S-nonnegative garrote estimator is derived in the next theorem.

Theorem 6.1. *Let ρ be a twice differentiable function and let $\lambda \geq 0$. The influence function of the S-nonnegative garrote regression functional at a point $P_0 = (\mathbf{X}_0^{\mathrm{T}}, Y_0)$ with $\mathbf{X}_0 = (X_{01}, \ldots, X_{0p})^{\mathrm{T}}$ is given by $\mathrm{IF}(P_0, \boldsymbol{\beta}^{\text{S-NNG}}, F) = \left(\mathrm{IF}_1(P_0, \boldsymbol{\beta}^{\text{S-NNG}}, F), \ldots, \mathrm{IF}_p(P_0, \boldsymbol{\beta}^{\text{S-NNG}}, F)\right)^{\mathrm{T}}$ with*

$$\mathrm{IF}_j(P_0, \boldsymbol{\beta}^{\text{S-NNG}}, F) = \begin{cases} 0 & \text{if } \beta_j^{\text{S-NNG}}(F) = 0, \\ \Pi_j \Bigg[-\dfrac{\mathbf{b}_F}{\mu_1^2}\left(\rho\left(\dfrac{r_0}{\sigma(F)}\right) - b\right) + \dfrac{1}{\mu_1}\psi\left(\dfrac{r_0}{\sigma(F)}\right)\mathbf{X}_0 \\ \quad + \dfrac{\nu_2 + \mu_1}{\mu_1^3}\left(\rho\left(\dfrac{r_0}{\sigma(F)}\right) - b\right)\mathbf{a}_F - \dfrac{1}{\mu_1^2}\psi\left(\dfrac{r_0}{\sigma(F)}\right)\dfrac{r_0}{\sigma(F)}\mathbf{a}_F \\ \quad + \lambda \mathrm{diag}\left(\boldsymbol{\beta}^{\text{init}}(F)\right)^{-2} \mathrm{IF}(P_0, \boldsymbol{\beta}^{\text{init}}, F) \Bigg] & \text{otherwise,} \end{cases}$$

for $j = 1, \ldots, p$, where Π_j denotes the jth row of

$$\sigma(F)\mu_1^3 \left[\mu_1^2 \mathbf{A} - \mu_1 \mathbf{b}_F \mathbf{a}_F^{\mathrm{T}} - \mu_1 \mathbf{a}_F \mathbf{b}_F^{\mathrm{T}} + \nu_2 \mathbf{a}_F \mathbf{a}_F^{\mathrm{T}}\right]^{-1},$$

$$
\begin{aligned}
r_0 &= Y_0 - \mathbf{X}_0^{\mathrm{T}} \boldsymbol{\beta}^{\text{S-NNG}}(F), & \mathbf{A} &= \int \psi'(u_F)\, \boldsymbol{\mathcal{X}} \boldsymbol{\mathcal{X}}^{\mathrm{T}} \,\mathrm{d}F, \\
\mathbf{a}_F &= \int \psi(u_F) \boldsymbol{\mathcal{X}} \,\mathrm{d}F, & \mathbf{b}_F &= \int \psi'(u_F)\, u_F \boldsymbol{\mathcal{X}} \,\mathrm{d}F, \\
\mu_1 &= \int \psi(u_F) u_F \,\mathrm{d}F, & \nu_2 &= \int \psi'(u_F) u_F^2 \,\mathrm{d}F, \\
u_F &= (Y - \boldsymbol{\mathcal{X}}^{\mathrm{T}} \boldsymbol{\beta}^{\text{S-NNG}}(F))/\sigma(F), & &
\end{aligned} \tag{13}
$$

with ψ' the first derivative of ψ, and where $\operatorname{diag}\left(\left(\boldsymbol{\beta}^{\text{init}}(F)\right)^{-2}\right)$ *is a diagonal matrix where the jth diagonal element is* $\left(\beta_j^{\text{init}}(F)\right)^{-2}$. *The influence function of the scale is given by*

$$
\operatorname{IF}(P_0, \sigma, F) = \frac{1}{\mu_1} \left\{ \sigma(F) \left(\rho\left(\frac{r_0}{\sigma(F)} \right) - b \right) - \mathbf{a}_F^{\mathrm{T}} \operatorname{IF}(P_0, \boldsymbol{\beta}^{\text{S-NNG}}, F) \right\}.
$$

Theorem 6.1 reveals that the boundedness of the influence function $\operatorname{IF}_j(P_0, \boldsymbol{\beta}^{\text{S-NNG}}, F)$ is guaranteed if all quantities involved such as $\mathbf{a}_F$, $\mathbf{b}_F$, $\mathbf{X}_0$, and so on, are finite, as well as the influence function of the initial estimator $\operatorname{IF}(P_0, \boldsymbol{\beta}^{\text{init}}, F)$ is bounded. If the latter does not hold, then the influence function of the S-nonnegative garrote estimator will be unbounded as well. See also the examples further in this section.

For completeness, we also provide the expression of the influence function of the original nonnegative garrote estimator in Theorem 6.2. But we first introduce the functional form of the nonnegative garrote estimator.

To obtain the functional form of the nonnegative garrote estimator, denoted by $\left(\boldsymbol{\beta}^{\text{NNG}}(F)\right)^{\mathrm{T}} = \left(\beta_1^{\text{NNG}}(F), \ldots, \beta_p^{\text{NNG}}(F)\right)^{\mathrm{T}}$, we start with the functional form of the initial estimator and we then shrink or put some components of the functional form of the initial estimator equal to zero by using the functional form of the nonnegative garrote shrinkage factors. This functional form of the nonnegative garrote shrinkage factors $\mathbf{c}^{\text{NNG}}(F) = \left(c_1^{\text{NNG}}(F), \ldots, c_p^{\text{NNG}}(F)\right)^{\mathrm{T}}$ can be found by minimizing

$$
\begin{aligned}
&\frac{1}{2} \int (Y - \boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\mathbf{c})^2 \,\mathrm{d}F + \lambda \sum_{j=1}^{p} c_j \\
&\text{s.t. } c_j \geq 0 \ (j = 1, \ldots, p),
\end{aligned} \tag{14}
$$

for $\mathbf{c} \in \mathbb{R}_+^p$. The functional form of the nonnegative garrote estimator for the coefficients β_j, $j = 1, \ldots, p$, is then given by

$$
\beta_j^{\text{NNG}}(F) = c_j^{\text{NNG}}(F) \beta_j^{\text{init}}(F).
$$

Theorem 6.2. *The influence function of the original nonnegative garrote regression functional for $\lambda \geq 0$ is given by* $\operatorname{IF}(P_0, \boldsymbol{\beta}^{\text{NNG}}, F) = \left(\operatorname{IF}_1(P_0, \boldsymbol{\beta}^{\text{NNG}}, F), \ldots, \operatorname{IF}_p(P_0, \boldsymbol{\beta}^{\text{NNG}}, F)\right)^{\mathrm{T}}$ *with*

$$
\operatorname{IF}_j(P_0, \boldsymbol{\beta}^{\text{NNG}}, F) = \begin{cases} 0 & \text{if } \beta_j^{\text{NNG}}(F) = 0, \\ \Pi_j \left[\left(Y_0 - \mathbf{X}_0^{\mathrm{T}} \boldsymbol{\beta}^{\text{NNG}}(F)\right) \mathbf{X}_0 - \lambda \left(\boldsymbol{\beta}^{\text{init}}(F)\right)^{-1} \right. & \\ \quad \left. + \lambda \operatorname{diag}\left(\boldsymbol{\beta}^{\text{init}}(F)\right)^{-2} \operatorname{IF}(P_0, \boldsymbol{\beta}^{\text{init}}, F) \right] & \\ & \text{otherwise,} \end{cases}
$$

for $j = 1, \ldots, p$, where Π_j denotes the jth row of $\left(\int \boldsymbol{\mathcal{X}} \boldsymbol{\mathcal{X}}^{\mathrm{T}} \,\mathrm{d}F\right)^{-1}$ *and* $\left(\boldsymbol{\beta}^{\text{init}}(F)\right)^{-1}$ *is a column vector of length p where the jth element is* $\left(\beta_j^{\text{init}}(F)\right)^{-1}$.

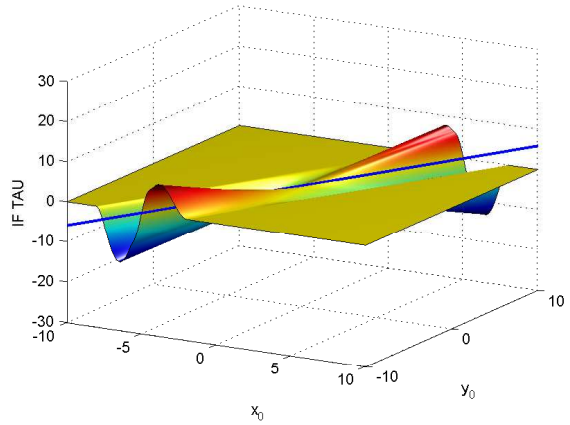
The proof is along the same lines as the proof of Theorem 6.1 and is therefore omitted here.

We will now compare the influence functions of different initial estimators, the original nonnegative garrote estimator and the S-nonnegative garrote estimator. These influence functions are plotted for simple linear regression $Y = X\beta_0 + \varepsilon$ with $\beta_0 = 2$ and where X and ε are independent and both standard normal distributed. Different values for the regularization parameter λ are used, i.e. $\lambda = 0.1$ and $\lambda = 0.5$.

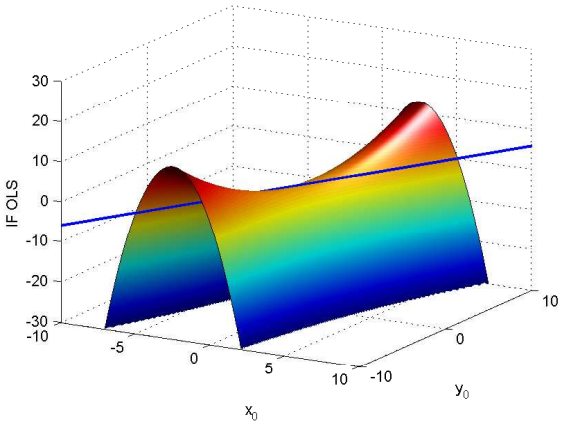
The influence functions of the τ- and OLS-estimator together with those of the S-nonnegative garrote estimator for different values of the regularization parameter λ and different initial estimators are plotted in Figure 1. The blue lines in these plots correspond to the true regression line. It can be seen from Figures 1(a), 1(c) and 1(e) that the shape of the influence functions of the τ-estimator and of the S-nonnegative garrote with the τ-estimator as initial estimator are quite similar. This just illustrates that the main characteristics of the influence function $\mathrm{IF}_j(P_0, \boldsymbol{\beta}^{\text{S-NNG}}, F)$ of the S-nonnegative garrote estimator are determined by these of the influence function of the initial estimator $\mathrm{IF}(P_0, \boldsymbol{\beta}^{\text{init}}, F)$, together with, among others, the boundedness (or not) of $\mathbf{X}_0$. Since the influence functions of, for example, the τ-estimator and the S-estimator are unbounded (see for example Yohai and Zamar (1988) and Yohai and Zamar (1997)) this property is inherited by the influence function of the S-nonnegative garrote estimator with the τ-estimator as initial estimator. Estimators with an unbounded influence function can still have a high breakdown point, and a more detailed study of the maximum bias properties of estimators explains their 'robustness' properties. We refer the readers to Yohai and Zamar (1988), Yohai and Zamar (1997) and Maronna et al. (2006), among others, for background information on these issues.

There are however also some noticeable (quantitative) differences in the influence functions of the τ-estimator and of the S-nonnegative garrote with the τ-estimator as initial estimator, in particular for points (x_0, y_0) that are away from the regression line, but still close to the regression line. See Figures 1(a), 1(c) and 1(e): the area where the influence function is zero is more extended for the S-nonnegative garrote estimator, and this area is larger for smaller values of λ. At the same time for points (x_0, y_0) away but even closer to the regression line, the values of the influence function for the S-nonnegative garrote estimator tend to be larger. In other words, the S-nonnegative garrote estimator is less influenced by those observations (x_0, y_0) with larger residuals and more influenced by those with smaller residuals (see the behaviour away but close to the straight line in Figures 1(a) and 1(e)). Figures 1(b), 1(d) and 1(f) show that the influence function of the OLS-estimator and of the S-nonnegative garrote estimator with the OLS-estimator as initial estimator are even unbounded for points (x_0, y_0) far away from the regression line (there are no flat zero-valued parts).

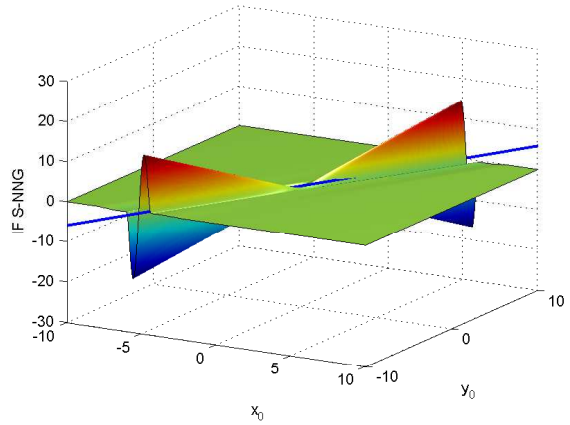
Figure 2 presents again the influence functions of the τ- and OLS-estimator, but now together with those of the original nonnegative garrote estimator for different initial estimators and for different values of the regularization parameter. It can be seen that the influence function of the original nonnegative garrote estimator is even unbounded for points (x_0, y_0) far away from the regression line, when the OLS-estimator *as well as* the τ-estimator are used as initial estimator. In conclusion, the S-nonnegative garrote estimator with as initial estimator a τ-estimator is 'robust' to regression outliers, i.e. points that are far away from the regression line (see e.g. Figures 1(c) and 1(e)) whereas the original nonnegative garrote estimator with as initial estimator a τ-estimator is not robust to such outliers (see for example Figures 2(c) and 2(e)).



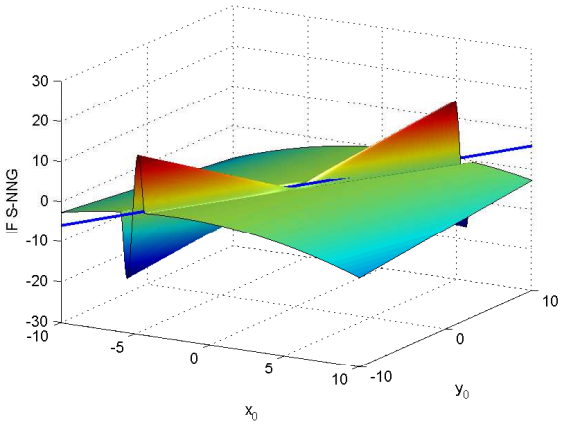
(a) τ-estimator.



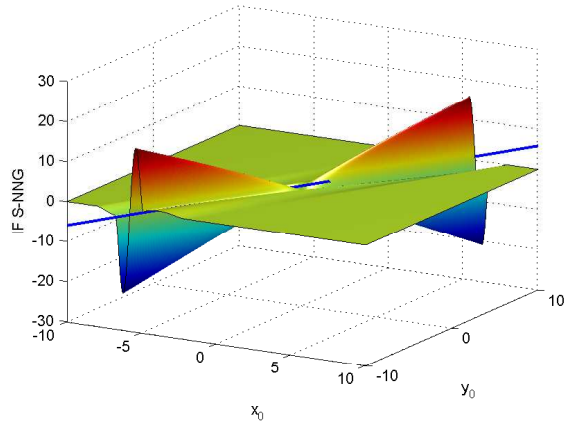
(b) OLS-estimator.



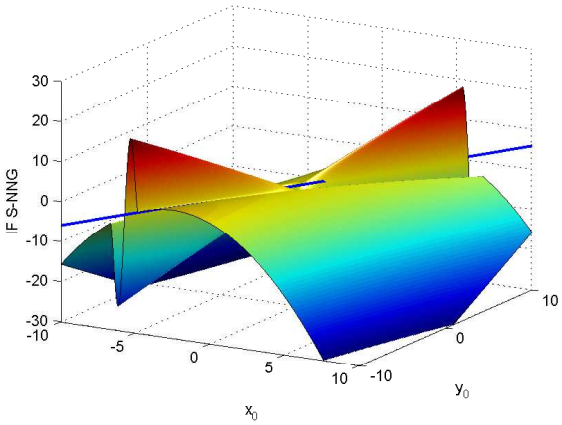
(c) S-NNG estimator with τ-estimator as initial estimator and $\lambda = 0.1$.



(d) S-NNG estimator with OLS-estimator as initial estimator and $\lambda = 0.1$.

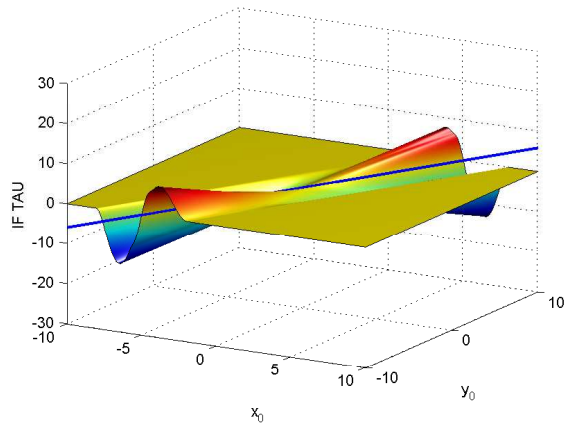


(e) S-NNG estimator with τ-estimator as initial estimator and $\lambda = 0.5$.

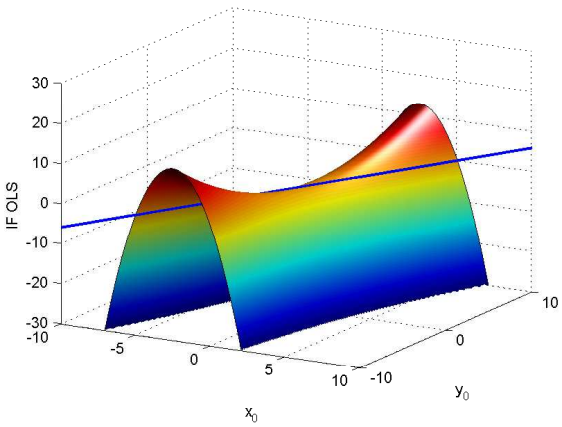


(f) S-NNG estimator with OLS-estimator as initial estimator and $\lambda = 0.5$.

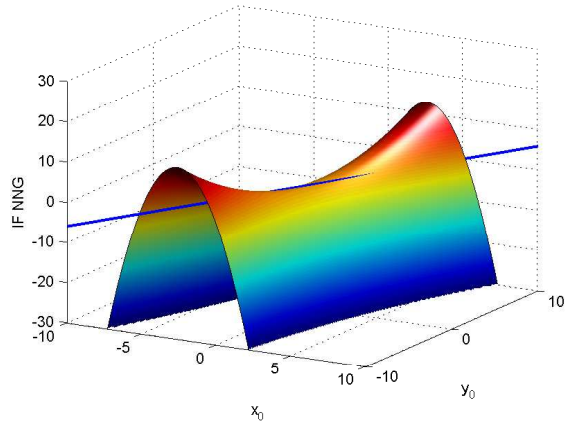
Figure 1: Influence functions for the τ-estimator, the OLS-estimator and the S-NNG estimator.



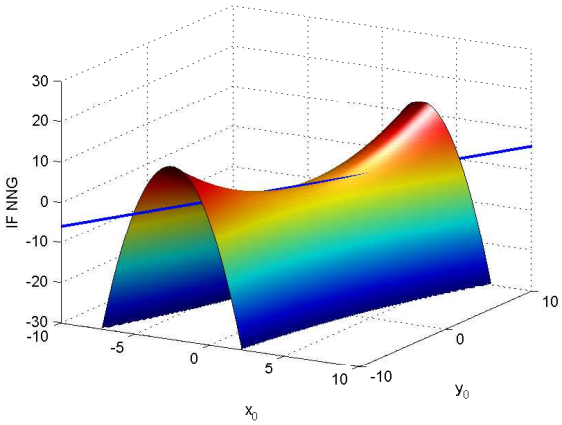
(a) τ-estimator.



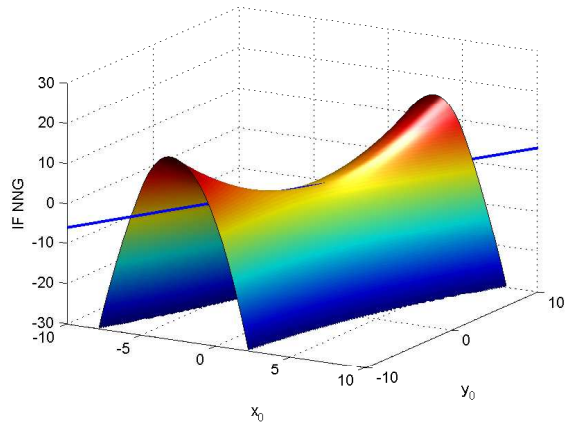
(b) OLS-estimator.



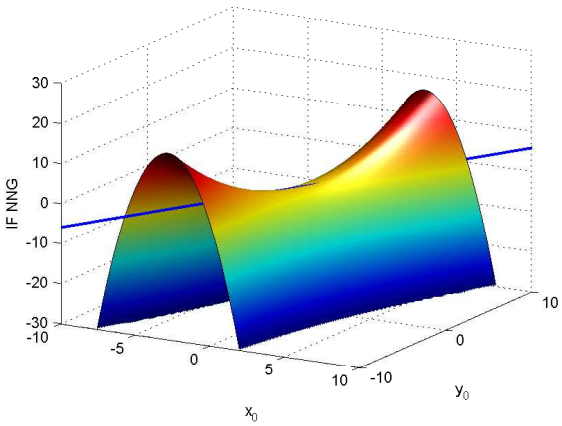
(c) NNG estimator with τ-estimator as initial estimator and $\lambda = 0.1$.



(d) NNG estimator with OLS-estimator as initial estimator and $\lambda = 0.1$.



(e) NNG estimator with τ-estimator as initial estimator and $\lambda = 0.5$.



(f) NNG estimator with OLS-estimator as initial estimator and $\lambda = 0.5$.

Figure 2: Influence functions for the τ-estimator, the OLS-estimator and the original NNG estimator.

7 Proofs of the theorems

7.1 Proof of Theorem 4.1

Before proving Theorem 4.1, we prove the following lemma.

Lemma 7.1. *Suppose Assumptions 4.1.4, 4.1.5 and 4.2 hold and that* $\widehat{\mathbf{c}}^0$ *is the current value of* $\mathbf{c}$ *in the iterative procedure* (5)*, then*

$$0 < \frac{1}{n\omega_S(\widehat{\mathbf{c}}^0)} < +\infty,$$

where $\omega_S(\mathbf{c}) = \dfrac{\widehat{\sigma}(\mathbf{r}(\mathbf{c}))}{\sum_{i=1}^n W_{S,i}(\mathbf{c}) r_i^2(\mathbf{c})}$.

Proof of Lemma 7.1. By Lemma 2.2 in Maronna and Yohai (1981) there exist constants $A_1, A_2 \in (0, +\infty)$ such that $A_1 \leq \widehat{\sigma}(\mathbf{r}(\widehat{\mathbf{c}}^0)) \leq A_2$. Further, we have, for $i = 1, \ldots, n$, that $0 \leq W_{S,i}(\widehat{\mathbf{c}}^0) \leq 1$ because of Assumption 4.2.3, and if $|r_i(\widehat{\mathbf{c}}^0)| > d$, then $W_{S,i}(\widehat{\mathbf{c}}^0) = 0$. Hence,

$$0 < \frac{1}{n\omega_S(\widehat{\mathbf{c}}^0)} = \frac{\sum_{i=1}^n W_{S,i}(\widehat{\mathbf{c}}^0) r_i^2(\widehat{\mathbf{c}}^0)}{n\widehat{\sigma}(\mathbf{r}(\widehat{\mathbf{c}}^0))} \leq \frac{\sum_{i=1}^n W_{S,i}(\widehat{\mathbf{c}}^0) d^2}{n\widehat{\sigma}(\mathbf{r}(\widehat{\mathbf{c}}^0))} \leq \frac{d^2}{\widehat{\sigma}(\mathbf{r}(\widehat{\mathbf{c}}^0))} < +\infty.$$

□

The proof of Theorem 4.1 is based on arguments similar to those used in the proof of Theorem 1 in Yuan and Lin (2007).

Proof of Theorem 4.1. Let

$$\begin{aligned} \Lambda_{\mathcal{NS}} &= \{j : c_j = 0, \beta_j \neq 0\} & \Lambda_{\mathcal{NN}} &= \{j : c_j = 0, \beta_j = 0\} \\ \Lambda_{\mathcal{SS}} &= \{j : c_j > 0, \beta_j \neq 0\} & \Lambda_{\mathcal{SN}} &= \{j : c_j > 0, \beta_j = 0\}\ , \end{aligned}$$

and $p_{ij} = \#(\Lambda_{ij})$, for $i, j = \mathcal{S}, \mathcal{N}$. Also define the events $\mathcal{A} = \{p_{\mathcal{SN}} > 0\}$ and $\mathcal{B} = \{p_{\mathcal{NS}} = 0\}$.

We first prove the second part of the theorem by proving that

$$P(\mathcal{A}) \to 0, \qquad \text{as } n \to \infty, \tag{15}$$

by contradiction type of arguments, and by then showing that

$$P(\mathcal{B} | \mathcal{A}^c) \to 1, \qquad \text{as } n \to \infty. \tag{16}$$

The convergence rate of the S-nonnegative garrote estimator of the first part of the theorem is obtained as a final result in the proof of statement (16).

In this proof, the following notations are introduced: for a vector $\mathbf{c} \in \mathbb{R}^p$, denote $\mathbf{c}_{ij} = \mathbf{c}_{\Lambda_{ij}}$, for $i, j = \mathcal{S}, \mathcal{N}$, $\mathbf{c}_{i\ell} = \left(\mathbf{c}_{i\mathcal{S}}^{\mathrm{T}}, \mathbf{c}_{i\mathcal{N}}^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\mathbf{c}_{\ell j} = \left(\mathbf{c}_{\mathcal{S}j}^{\mathrm{T}}, \mathbf{c}_{\mathcal{N}j}^{\mathrm{T}}\right)^{\mathrm{T}}$ for $i, j = \mathcal{S}, \mathcal{N}$. For all other matrices and vectors, the same type of notation is used. Further, let $\mathbf{1}_p$ be a vector of length p with all elements equal to one, $\mathbf{I}_p$ a diagonal matrix of size $p \times p$ with the diagonal elements equal to one and Δ a diagonal matrix with on the diagonal the elements of $\boldsymbol{\beta}$. By way of illustration, suppose that $p = 10$ and that the set of indices $\mathcal{S}$ contains 4 elements. For 3 of these elements, we have that $c_j > 0$. For 1 element of the set of indices $\mathcal{N}$, we also have that $c_j > 0$. For this example we now have that $p_{\mathcal{SS}} = 3$, $p_{\mathcal{NS}} = 1$, $p_{\mathcal{SN}} = 1$ and $p_{\mathcal{NN}} = 5$. The vector $\mathbf{1}_{p_{\mathcal{NS}}}$ is thus a vector containing one element and the matrix $\Delta_{\mathcal{NS}}$ is a $(1 \times 1)-$matrix containing the value of β_j for which $c_j = 0$.

1. Proof of statement (15). Let $\widehat{\mathbf{c}}_{\mathcal{S}\ell}$ be the unconstrained minimizer of

$$\frac{1}{2}\mathbf{c}^{\mathrm{T}}\mathbf{Z}_{\mathcal{S}\ell}^{\mathrm{T}}\mathbf{W}_S(\widehat{\mathbf{c}}^0)\mathbf{Z}_{\mathcal{S}\ell}\mathbf{c} - \left(\mathbf{Z}_{\mathcal{S}\ell}^{\mathrm{T}}\mathbf{W}_S(\widehat{\mathbf{c}}^0)\mathbf{Y} - \frac{\lambda_n}{\omega_S(\widehat{\mathbf{c}}^0)}\mathbf{1}_{p_{\mathcal{S}\ell}}\right)^{\mathrm{T}}\mathbf{c}$$

with $\mathbf{c} \in \mathbb{R}^{p_{\mathcal{S}\ell}}$. Hence, $\widehat{\mathbf{c}}_{\mathcal{S}\ell}$ is given by

$$\begin{aligned}\begin{pmatrix}\widehat{\mathbf{c}}_{\mathcal{S}\mathcal{S}}\\ \widehat{\mathbf{c}}_{\mathcal{S}\mathcal{N}}\end{pmatrix} &= \begin{pmatrix}\mathbf{Z}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\mathbf{W}_S(\widehat{\mathbf{c}}^0)\mathbf{Z}_{\mathcal{S}\mathcal{S}} & \mathbf{Z}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\mathbf{W}_S(\widehat{\mathbf{c}}^0)\mathbf{Z}_{\mathcal{S}\mathcal{N}}\\ \mathbf{Z}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\mathbf{W}_S(\widehat{\mathbf{c}}^0)\mathbf{Z}_{\mathcal{S}\mathcal{S}} & \mathbf{Z}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\mathbf{W}_S(\widehat{\mathbf{c}}^0)\mathbf{Z}_{\mathcal{S}\mathcal{N}}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{Z}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\mathbf{W}_S(\widehat{\mathbf{c}}^0)\mathbf{Y} - \frac{\lambda_n}{\omega_S(\widehat{\mathbf{c}}^0)}\mathbf{1}_{p_{\mathcal{S}\mathcal{S}}}\\ \mathbf{Z}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\mathbf{W}_S(\widehat{\mathbf{c}}^0)\mathbf{Y} - \frac{\lambda_n}{\omega_S(\widehat{\mathbf{c}}^0)}\mathbf{1}_{p_{\mathcal{S}\mathcal{N}}}\end{pmatrix}\\ &= \begin{pmatrix}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}/n & \widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}/n\\ \widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}/n & \widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}/n\end{pmatrix}^{-1}\begin{pmatrix}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\widetilde{\mathbf{Y}}/n - \frac{\lambda_n}{\omega_n}\mathbf{1}_{p_{\mathcal{S}\mathcal{S}}}\\ \widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\widetilde{\mathbf{Y}}/n - \frac{\lambda_n}{\omega_n}\mathbf{1}_{p_{\mathcal{S}\mathcal{N}}}\end{pmatrix}\end{aligned}$$

where $\widetilde{\mathbf{Z}}_{\mathcal{S}j} = \mathbf{W}_S^{1/2}(\widehat{\mathbf{c}}^0)\mathbf{Z}_{\mathcal{S}j}$, for $j = \mathcal{S},\mathcal{N}$, $\widetilde{\mathbf{Y}} = \mathbf{W}_S^{1/2}(\widehat{\mathbf{c}}^0)\mathbf{Y}$ and $\omega_n = n\omega_S(\widehat{\mathbf{c}}^0)$. Denote

$$\mathbf{A} = \begin{pmatrix}\mathbf{A}_{\mathcal{S}\mathcal{S}} & \mathbf{A}_{\mathcal{S}\mathcal{N}}\\ \mathbf{A}_{\mathcal{N}\mathcal{S}} & \mathbf{A}_{\mathcal{N}\mathcal{N}}\end{pmatrix},$$

with

$$\begin{aligned}&\mathbf{A}_{ij} = \widetilde{\mathbf{Z}}_{\mathcal{S}i}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}j}/n \qquad \text{for } i,j = \mathcal{S},\mathcal{N},\\ &\mathbf{B} = \mathbf{A}_{\mathcal{N}\mathcal{N}} - \mathbf{A}_{\mathcal{N}\mathcal{S}}\mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{A}_{\mathcal{S}\mathcal{N}} = \frac{1}{n}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\left\{\mathbf{I}_{p_{\mathcal{S}\mathcal{S}}} - \widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\left(\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\right)^{-1}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\right\}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}\end{aligned}$$

and because

$$\mathbf{I}_{p_{\mathcal{S}\mathcal{S}}} - \widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\left(\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\right)^{-1}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}$$

is a projection matrix, and therefore positive definite, we have that $\mathbf{B}$ is a positive semidefinite matrix. By using these notations we can obtain the inverse of the matrix $\mathbf{A}$,

$$\mathbf{A}^{-1} = \begin{pmatrix}\mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1} + \mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{A}_{\mathcal{S}\mathcal{N}}\mathbf{B}^{-1}\mathbf{A}_{\mathcal{N}\mathcal{S}}\mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1} & -\mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{A}_{\mathcal{S}\mathcal{N}}\mathbf{B}^{-1}\\ -\mathbf{B}^{-1}\mathbf{A}_{\mathcal{N}\mathcal{S}}\mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1} & \mathbf{B}^{-1}\end{pmatrix}.$$

Therefore we have that,

$$\widehat{\mathbf{c}}_{\mathcal{S}\mathcal{N}} = -\mathbf{B}^{-1}\mathbf{A}_{\mathcal{N}\mathcal{S}}\mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1}\left(\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\widetilde{\mathbf{Y}}/n - \frac{\lambda_n}{\omega_n}\mathbf{1}_{p_{\mathcal{S}\mathcal{S}}}\right) + \mathbf{B}^{-1}\left(\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\widetilde{\mathbf{Y}}/n - \frac{\lambda_n}{\omega_n}\mathbf{1}_{p_{\mathcal{S}\mathcal{N}}}\right) = \mathbf{B}^{-1}\mathbf{w}$$

with

$$\begin{aligned}\mathbf{w} &= \widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\widetilde{\mathbf{Y}}/n - \frac{\lambda_n}{\omega_n}\mathbf{1}_{p_{\mathcal{S}\mathcal{N}}} - \mathbf{A}_{\mathcal{N}\mathcal{S}}\mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\widetilde{\mathbf{Y}}/n + \frac{\lambda_n}{\omega_n}\mathbf{A}_{\mathcal{N}\mathcal{S}}\mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{1}_{p_{\mathcal{S}\mathcal{S}}}.\\ &= \widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\widetilde{\mathbf{Y}}/n - \widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\left(\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\right)^{-1}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\widetilde{\mathbf{Y}}/n - \frac{\lambda_n}{\omega_n}\mathbf{1}_{p_{\mathcal{S}\mathcal{N}}} + \frac{\lambda_n}{\omega_n}\mathbf{A}_{\mathcal{N}\mathcal{S}}\mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{1}_{p_{\mathcal{S}\mathcal{S}}}\\ &= \widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\left(\mathbf{I}_{p_{\mathcal{S}\mathcal{S}}} - \widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\left(\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\right)^{-1}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\right)\widetilde{\mathbf{Y}}/n - \frac{\lambda_n}{\omega_n}\mathbf{1}_{p_{\mathcal{S}\mathcal{N}}} + \frac{\lambda_n}{\omega_n}\mathbf{A}_{\mathcal{N}\mathcal{S}}\mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{1}_{p_{\mathcal{S}\mathcal{S}}}.\end{aligned}$$

We first look for an upper bound for

$$\mathbf{A}_{\mathcal{N}\mathcal{S}}\mathbf{A}_{\mathcal{S}\mathcal{S}}^{-1} = \left(\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\right)\left(\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\right)^{-1} = \left(\frac{1}{n}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{N}}^{\mathrm{T}}\frac{1}{n}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\right)\left(\frac{1}{n}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}^{\mathrm{T}}\frac{1}{n}\widetilde{\mathbf{Z}}_{\mathcal{S}\mathcal{S}}\right)^{-1}.$$

Since $\widehat{\boldsymbol{\beta}}^{\text{init}}$ is a strongly consistent estimator with rate κ_n and by using Assumption 4.1.2, we find for $(\mathbf{Z}_{\mathcal{SN}})_j$, the jth column of $\mathbf{Z}_{\mathcal{SN}}$, that

$$\frac{1}{n}\|(\mathbf{Z}_{\mathcal{SN}})_j - \mathbf{0}\|_2^2 = O(\kappa_n^2),$$

with probability 1 and thus is $\frac{1}{\sqrt{n}}\|(\mathbf{Z}_{\mathcal{SN}})_j\|_2 = O(\kappa_n)$ with probability 1.
In addition, we have for the jth column of $\mathbf{Z}_{\mathcal{SS}}$ that

$$\begin{aligned}\frac{1}{\sqrt{n}}\|(\mathbf{Z}_{\mathcal{SS}})_j\|_2 &\leq \frac{1}{\sqrt{n}}\|(\mathbf{Z}_{\mathcal{SS}})_j - (\mathbf{X}_{\mathcal{SS}})_j\beta_j\|_2 + \frac{1}{\sqrt{n}}\|(\mathbf{X}_{\mathcal{SS}})_j\beta_j\|_2 \\ &= O(\kappa_n) + \frac{1}{\sqrt{n}}\|(\mathbf{X}_{\mathcal{SS}})_j\beta_j\|_2 < +\infty,\end{aligned}$$

since $\widehat{\boldsymbol{\beta}}^{\text{init}}$ is strongly consistent with rate κ_n and $(\mathbf{X}_{\mathcal{SS}})_j$ is uniformly bounded.
Since $\|\mathbf{W}_S(\widehat{\mathbf{c}}^0)\|_2 \leq 1$, we have that,

$$\begin{aligned}\frac{1}{n}\|\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{SS}}\|_2 &\leq \frac{1}{n}\|\mathbf{Z}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{Z}_{\mathcal{SS}}\|_2 \\ &\leq \frac{1}{n}\|\mathbf{Z}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{Z}_{\mathcal{SS}} - \Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2 + \frac{1}{n}\|\Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2 \\ &= \frac{1}{n}\|\mathbf{Z}_{\mathcal{SS}}^{\mathrm{T}}(\mathbf{Z}_{\mathcal{SS}} - \mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}) + (\mathbf{Z}_{\mathcal{SS}}^{\mathrm{T}} - \Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}})\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2 + \frac{1}{n}\|\Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2 \\ &\leq \frac{1}{\sqrt{n}}\|\mathbf{Z}_{\mathcal{SS}}\|_2\frac{1}{\sqrt{n}}\|\mathbf{Z}_{\mathcal{SS}} - \mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2 + \frac{1}{\sqrt{n}}\|\mathbf{Z}_{\mathcal{SS}}^{\mathrm{T}} - \Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\|_2\frac{1}{\sqrt{n}}\|\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2 \\ &+ \frac{1}{n}\|\Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2 \\ &= \left(O(\kappa_n) + \frac{1}{\sqrt{n}}\|\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2\right)O(\kappa_n) + O(\kappa_n)\frac{1}{\sqrt{n}}\|\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2 + \frac{1}{n}\|\Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2 \\ &= O(\kappa_n^2) + O(\kappa_n) + \frac{1}{n}\|\Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2 \\ &= O(\kappa_n) + \frac{1}{n}\|\Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2,\end{aligned}$$

with probability 1. Therefore we obtain that,

$$\begin{aligned}\left(\frac{1}{n}\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{SS}}\right)^{-1} &= \left(\frac{1}{n}\Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\right)^{-1}\left(\mathbf{I}_{p_{\mathcal{SS}}} + O(\kappa_n)\left(\frac{1}{n}\Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\right)^{-1}\right) \\ &= \left(\frac{1}{n}\Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\right)^{-1}(\mathbf{I}_{p_{\mathcal{SS}}} + O(\kappa_n)\mathbf{I}_{p_{\mathcal{SS}}})\end{aligned}$$

and

$$\|(\frac{1}{n}\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{SS}})^{-1}\|_2 = \|(\frac{1}{n}\Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}})^{-1}\|_2(1 + O(\kappa_n)),$$

with probability 1. An upper bound for $\mathbf{A}_{\mathcal{NS}}\mathbf{A}_{\mathcal{SS}}^{-1}$ is now given by

$$\begin{aligned}\|\mathbf{A}_{\mathcal{NS}}\mathbf{A}_{\mathcal{SS}}^{-1}\|_2 &\leq \|\mathbf{A}_{\mathcal{NS}}\|_2\|\mathbf{A}_{\mathcal{SS}}^{-1}\|_2 \\ &= \|\frac{1}{n}\widetilde{\mathbf{Z}}_{\mathcal{SN}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{SS}}\|_2\|(\frac{1}{n}\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{SS}})^{-1}\|_2 \\ &\leq \frac{1}{\sqrt{n}}\|\mathbf{Z}_{\mathcal{SN}}\|_2\frac{1}{\sqrt{n}}\|\mathbf{Z}_{\mathcal{SS}}\|_2\|(\frac{1}{n}\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathcal{SS}})^{-1}\|_2 \\ &= O(\kappa_n)\left(O(\kappa_n) + \frac{1}{\sqrt{n}}\|\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}}\|_2\right)\|(\frac{1}{n}\Delta_{\mathcal{SS}}\mathbf{X}_{\mathcal{SS}}^{\mathrm{T}}\mathbf{X}_{\mathcal{SS}}\Delta_{\mathcal{SS}})^{-1}\|_2(1 + O(\kappa_n)) \\ &= O(\kappa_n),\end{aligned}$$

with probability 1. Hence,

$$\mathbf{w} = \widetilde{\mathbf{Z}}_{\mathcal{SN}}^{\mathrm{T}} \left(\mathbf{I}_{p_{\mathcal{SS}}} - \widetilde{\mathbf{Z}}_{\mathcal{SS}} (\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}} \widetilde{\mathbf{Z}}_{\mathcal{SS}})^{-1} \widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}} \right) \widetilde{\mathbf{Y}}/n - \frac{\lambda_n}{\omega_n}(1 + O(\kappa_n)) \mathbf{1}_{p_{\mathcal{SN}}}.$$

Because $\mathbf{I}_{p_{\mathcal{SS}}} \widetilde{\mathbf{Z}}_{\mathcal{SS}} (\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}} \widetilde{\mathbf{Z}}_{\mathcal{SS}})^{-1} \widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}}$ is a projection matrix, we have that

$$\| \mathbf{I}_{p_{\mathcal{SS}}} - \widetilde{\mathbf{Z}}_{\mathcal{SS}} (\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}} \widetilde{\mathbf{Z}}_{\mathcal{SS}})^{-1} \widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}} \|_2 \leq 1$$

and therefore,

$$\begin{aligned} \| (\mathbf{I}_{p_{\mathcal{SS}}} - \widetilde{\mathbf{Z}}_{\mathcal{SS}} (\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}} \widetilde{\mathbf{Z}}_{\mathcal{SS}})^{-1} \widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}}) \widetilde{\mathbf{Y}} \|_2 &\leq \| \mathbf{I}_{p_{\mathcal{SS}}} - \widetilde{\mathbf{Z}}_{\mathcal{SS}} (\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}} \widetilde{\mathbf{Z}}_{\mathcal{SS}})^{-1} \widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}} \|_2 \| \widetilde{\mathbf{Y}} \|_2 \\ &\leq \| \mathbf{Y} \|_2 = O(\sqrt{n}), \end{aligned}$$

with probability 1, because of Assumption 4.1.2 and 4.1.3 and by using Model (8). Since

$$\begin{aligned} \| \widetilde{\mathbf{Z}}_{\mathcal{SN}}^{\mathrm{T}} (\mathbf{I}_{p_{\mathcal{SS}}} - \widetilde{\mathbf{Z}}_{\mathcal{SS}} (\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}} \widetilde{\mathbf{Z}}_{\mathcal{SS}})^{-1} \widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}}) \widetilde{\mathbf{Y}} \|_2 &\leq \| \widetilde{\mathbf{Z}}_{\mathcal{SN}} \|_2 \| (\mathbf{I}_{p_{\mathcal{SS}}} - \widetilde{\mathbf{Z}}_{\mathcal{SS}} (\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}} \widetilde{\mathbf{Z}}_{\mathcal{SS}})^{-1} \widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}}) \widetilde{\mathbf{Y}} \|_2 \\ &\leq \| \mathbf{Z}_{\mathcal{SN}} \|_2 \| (\mathbf{I}_{p_{\mathcal{SS}}} - \widetilde{\mathbf{Z}}_{\mathcal{SS}} (\widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}} \widetilde{\mathbf{Z}}_{\mathcal{SS}})^{-1} \widetilde{\mathbf{Z}}_{\mathcal{SS}}^{\mathrm{T}}) \widetilde{\mathbf{Y}} \|_2 \\ &= O(\sqrt{n}\kappa_n) O(\sqrt{n}) = O(n\kappa_n), \end{aligned}$$

with probability 1, and because of Lemma 7.1, we find that

$$\mathbf{w} = O(\kappa_n) \mathbf{1}_{p_{\mathcal{SN}}} - \frac{\lambda_n}{\omega_n}(1 + O(\kappa_n)) \mathbf{1}_{p_{\mathcal{SN}}} = -\frac{\lambda_n}{\omega_n}(1 + o(1)) \mathbf{1}_{p_{\mathcal{SN}}} = -O(\lambda_n) \mathbf{1}_{p_{\mathcal{SN}}}.$$

The contradiction is now obtained from the fact that we have that for any $j \in \Lambda_{\mathcal{SN}}$ $\widehat{c}_j > 0$ and therefore $\mathbf{w}^{\mathrm{T}} \widehat{\mathbf{c}}_{\mathcal{SN}} < 0$. But we also have that $\widehat{\mathbf{c}}_{\mathcal{SN}} = \mathbf{B}^{-1}\mathbf{w}$ and therefore $\mathbf{w}^{\mathrm{T}} \mathbf{B}^{-1} \mathbf{w}$ is positive, because $\mathbf{B}^{-1}$ is a positive definite matrix. Consequently, $P(\mathcal{A}) \to 0$ for $n \to \infty$.

2. Proof of statement (16). We now show that $P(\mathcal{B} | \mathcal{A}^c) \to 1$ and therefore we assume $p_{\mathcal{SN}} = 0$. Let $\widehat{\mathbf{c}}_{\ell\mathcal{S}}$ be the unconstrained minimizer of

$$\frac{1}{2} \mathbf{c}^{\mathrm{T}} \mathbf{Z}_{\ell\mathcal{S}}^{\mathrm{T}} \mathbf{W}_S(\widehat{\mathbf{c}}^0) \mathbf{Z}_{\ell\mathcal{S}} \mathbf{c} - \left(\mathbf{Z}_{\ell\mathcal{S}}^{\mathrm{T}} \mathbf{W}_S(\widehat{\mathbf{c}}^0) \mathbf{Y} - \frac{\lambda_n}{\omega_S(\widehat{\mathbf{c}}^0)} \mathbf{1}_{p_{\ell\mathcal{S}}} \right)^{\mathrm{T}} \mathbf{c},$$

where $\mathbf{c} \in \mathbb{R}^{p_{\ell\mathcal{S}}}$. Denote again $\mathbf{W}_S^{1/2}(\widehat{\mathbf{c}}^0) \mathbf{Z}_{\ell\mathcal{S}}$ with $\widetilde{\mathbf{Z}}_{\ell\mathcal{S}}$, $\mathbf{W}_S^{1/2}(\widehat{\mathbf{c}}^0) \mathbf{Y}$ with $\widetilde{\mathbf{Y}}$ and $n\omega_S(\widehat{\mathbf{c}}^0)$ with ω_n. Using similar calculations as in the first part of this proof we obtain that

$$\begin{aligned} \widehat{\mathbf{c}}_{\ell\mathcal{S}} &= (\widetilde{\mathbf{Z}}_{\ell\mathcal{S}}^{\mathrm{T}} \widetilde{\mathbf{Z}}_{\ell\mathcal{S}}/n)^{-1} \left(\widetilde{\mathbf{Z}}_{\ell\mathcal{S}}^{\mathrm{T}} \widetilde{\mathbf{Y}}/n - \frac{\lambda_n}{\omega_n} \mathbf{1}_{p_{\ell\mathcal{S}}} \right), \\ \frac{1}{\sqrt{n}} \mathbf{Z}_{\ell\mathcal{S}} &= \frac{1}{\sqrt{n}} \mathbf{X}_{\ell\mathcal{S}} \Delta_{\ell\mathcal{S}} + O(\kappa_n), \\ \frac{1}{n} \widetilde{\mathbf{Z}}_{\ell\mathcal{S}}^{\mathrm{T}} \widetilde{\mathbf{Z}}_{\ell\mathcal{S}} &= \frac{1}{n} \Delta_{\ell\mathcal{S}} \mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}} \mathbf{X}_{\ell\mathcal{S}} \Delta_{\ell\mathcal{S}} + O(\kappa_n), \\ \| \widetilde{\mathbf{Z}}_{\ell\mathcal{S}}^{\mathrm{T}} \widetilde{\mathbf{Y}}/n \|_2 &\leq \| \Delta_{\ell\mathcal{S}} \mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}} \mathbf{Y}/n \|_2 + O(\kappa_n), \end{aligned}$$

with probability 1, and hence

$$\widehat{\mathbf{c}}_{\ell\mathcal{S}} = \left(\frac{1}{n} \Delta_{\ell\mathcal{S}} \mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}} \mathbf{X}_{\ell\mathcal{S}} \Delta_{\ell\mathcal{S}} \right)^{-1} \left(\Delta_{\ell\mathcal{S}} \mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}} \mathbf{Y}/n - \frac{\lambda_n}{\omega_n} \mathbf{1}_{p_{\ell\mathcal{S}}} \right) (1 + O(\kappa_n)).$$

Furthermore, we find that, because $\left\|\left(\frac{1}{n}\Delta_{\ell\mathcal{S}}\mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\ell\mathcal{S}}\Delta_{\ell\mathcal{S}}\right)^{-1}\right\|_2 < \infty$, $\mathbf{Y} = \mathbf{X}_{\ell\mathcal{S}}\Delta_{\ell\mathcal{S}}\mathbf{1}_{p_{\ell\mathcal{S}}} + \boldsymbol{\varepsilon}$ and because of Assumption 4.1.3,

$$\begin{aligned}\left(\frac{1}{n}\Delta_{\ell\mathcal{S}}\mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\ell\mathcal{S}}\Delta_{\ell\mathcal{S}}\right)^{-1}\Delta_{\ell\mathcal{S}}\mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}}\mathbf{Y}/n &= \left(\frac{1}{n}\Delta_{\ell\mathcal{S}}\mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\ell\mathcal{S}}\Delta_{\ell\mathcal{S}}\right)^{-1}\frac{1}{n}\Delta_{\ell\mathcal{S}}\mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\ell\mathcal{S}}\Delta_{\ell\mathcal{S}}\mathbf{1}_{p_{\ell\mathcal{S}}} \\ &\quad + \left(\frac{1}{n}\Delta_{\ell\mathcal{S}}\mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\ell\mathcal{S}}\Delta_{\ell\mathcal{S}}\right)^{-1}\Delta_{\ell\mathcal{S}}\mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}}\boldsymbol{\varepsilon}/n \\ &= \mathbf{1}_{p_{\ell\mathcal{S}}} + O\left(\frac{1}{n}\right)\mathbf{1}_{p_{\ell\mathcal{S}}}.\end{aligned}$$

We now have that

$$\begin{aligned}\widehat{\mathbf{c}}_{\ell\mathcal{S}} &= \left(\mathbf{1}_{p_{\ell\mathcal{S}}} + O\left(\frac{1}{n}\right)\mathbf{1}_{p_{\ell\mathcal{S}}}\right)(1+O(\kappa_n)) - \frac{\lambda_n}{\omega_n}\left(\frac{1}{n}\Delta_{\ell\mathcal{S}}\mathbf{X}_{\ell\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\ell\mathcal{S}}\Delta_{\ell\mathcal{S}}\right)^{-1}\mathbf{1}_{p_{\ell\mathcal{S}}}(1+O(\kappa_n)) \\ &= \mathbf{1}_{p_{\ell\mathcal{S}}} + \mathbf{1}_{p_{\ell\mathcal{S}}}O(\lambda_n),\end{aligned}$$

provided that $n\lambda_n = O(1)$, as $n \to \infty$.

Hence, $\widehat{\beta}_j^{\text{S-NNG}} = \widehat{\beta}_j^{\text{init}}(1+O(\lambda_n))$ for all j such that $\beta_j \neq 0$ and $P(\widehat{\beta}_j^{\text{S-NNG}} = 0) \to 1$ for all $j \in \mathcal{N}$. As a final result we obtain that

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}^{\text{S-NNG}} - \boldsymbol{\beta}\|_2 &\leq \|\widehat{\boldsymbol{\beta}}^{\text{S-NNG}} - \widehat{\boldsymbol{\beta}}^{\text{init}}\|_2 + \|\widehat{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}\|_2 \\ &= O(\lambda_n) + O(\kappa_n) = O(\lambda_n),\end{aligned}$$

with probability 1. □

7.2 Proof of Theorem 5.1

Before proving Theorem 5.1, we will prove the following Lemma. The proofs of this lemma and of Theorem 5.1 are inspired by the proofs of Theorem 2.1 in Yohai (1987), Theorem 3.1 in Yohai and Zamar (1988) and Theorem 2 in Wang et al. (2013).

Lemma 7.2. *Consider the same assumptions as in Theorem 5.1. Then, if* $0 < \lambda < +\infty$ *and for given* $\epsilon < \min\left\{\mathrm{BP}(\widehat{\boldsymbol{\beta}}^{\text{init}}, \mathbf{P}_{n-m}), \frac{1-2a_{nm}}{2-2a_{nm}}\right\}$, *there exists a* K *such that* $\frac{m}{n} \leq \epsilon$ *implies*

$$\inf_{\|\boldsymbol{\beta}\|_2 \geq K}\left\{\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta})) + \lambda\sum_{j=1}^{p}\frac{|\beta_j|}{|\widehat{\beta}_i^{\text{init}}|}\right\} > \widehat{\sigma}(\mathbf{r}(\widehat{\boldsymbol{\beta}}^{\text{init}})) + \lambda p,$$

where $\mathbf{r}(\boldsymbol{\beta}) = (r_1, \ldots, r_n)^{\mathrm{T}}$ *with* $r_i = Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ *and* $\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta}))$ *is the solution of*

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta}))}\right) = b.$$

Proof of Lemma 7.2. By definition of a_{nm}, we have

$$\#\left\{i : m+1 \leq i \leq n \text{ and } |\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}| > 0\right\}/(n-m) \geq 1 - a_{nm}$$

for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Take $a_n^* > a_{nm}$ such that $\epsilon < \frac{1-2a_n^*}{2-2a_n^*}$ too. Therefore, we can find $\delta > 0$ such that

$$\inf_{\|\boldsymbol{\beta}\|_2=1}\#\left\{i : m+1 \leq i \leq n \text{ and } |\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}| > \delta\right\}/(n-m) \geq 1 - a_n^*. \tag{17}$$

Since $1-\epsilon > 1-\frac{1-2a_n^*}{2-2a_n^*} = \frac{1}{2-2a_n^*}$, there exists $0 < a_0 < a = \sup_u \rho(u)$ such that $\frac{m}{n} \le \epsilon$ implies

$$a_0 \frac{n-m}{n} \ge (1-\epsilon)a_0 > \frac{a}{2-2a_n^*}. \tag{18}$$

Since $a = \sup_u \rho(u)$ and ρ is continuous, there exists k_2 such that $\rho(k_2) = a_0$. Furthermore, since $\epsilon < \mathrm{BP}(\widehat{\boldsymbol{\beta}}^{\text{init}}, \mathbf{P}_{n-m})$, there exists k_1 such that $\|\widehat{\boldsymbol{\beta}}^{\text{init}}\|_2 \le k_1$, and let $\widehat{\sigma}(\mathbf{r}(\widehat{\boldsymbol{\beta}}^{\text{init}})) = k_0$.
Now let $K_1 = \left(\max_{m+1\le i\le n} |Y_i| + k_0 k_2\right)/\delta$ and suppose that $\frac{m}{n} \le \epsilon$ and $\|\boldsymbol{\beta}\|_2 \ge K_1$. For any $i = 1, \ldots, n$, we have

$$\begin{aligned} |r_i| &= |Y_i - \mathbf{X}_i\boldsymbol{\beta}| \ge \left||\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}| - |Y_i|\right| \\ &= \left|\|\boldsymbol{\beta}\|_2 \left|\mathbf{X}_i^{\mathrm{T}}\left(\frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}\right)\right| - |Y_i|\right| \ge \left|K_1\left|\mathbf{X}_i^{\mathrm{T}}\left(\frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}\right)\right| - |Y_i|\right|. \end{aligned}$$

Hence we see that, by using (17) and (18), $\frac{m}{n} \le \epsilon$ implies

$$\begin{aligned} \inf_{\|\boldsymbol{\beta}\|_2\ge K_1} \sum_{i=1}^n \rho\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{k_0}\right) &\ge \inf_{\|\boldsymbol{\beta}\|_2=1} \sum_{i=1}^n \rho\left(\frac{K_1|X_i^{\mathrm{T}}\boldsymbol{\beta}| - |Y_i|}{k_0}\right) \\ &\ge \inf_{\|\boldsymbol{\beta}\|_2=1} \sum_{i\in A} \rho\left(\frac{K_1|X_i^{\mathrm{T}}\boldsymbol{\beta}| - |Y_i|}{k_0}\right) \\ &\ge (n-m)(1-a_n^*)\rho(k_2) \\ &= (n-m)(1-a_n^*)a_0 \\ &> n\frac{a}{2} = nb, \end{aligned}$$

where $A = \{i : m+1 \le i \le n \text{ and } |\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}| > \delta \text{ (for } \boldsymbol{\beta} \text{ with } \|\boldsymbol{\beta}\|_2 = 1)\}$. Therefore, for all $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_2 \ge K_1$, we have

$$\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta})) > k_0 = \widehat{\sigma}(\mathbf{r}(\widehat{\boldsymbol{\beta}}^{\text{init}})). \tag{19}$$

Let $K_2 = p^{3/2}k_1$. Take $K = \max(K_1, K_2)$, we have

$$\begin{aligned} \inf_{\|\boldsymbol{\beta}\|_2\ge K}\left\{\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta})) + \lambda\sum_{j=1}^p \frac{|\beta_j|}{|\widehat{\beta}_j^{\text{init}}|}\right\} &\ge \inf_{\|\boldsymbol{\beta}\|_2\ge K}\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta})) + \inf_{\|\boldsymbol{\beta}\|_2\ge K}\lambda\sum_{j=1}^p \frac{|\beta_j|}{|\widehat{\beta}_j^{\text{init}}|} \\ &\ge \inf_{\|\boldsymbol{\beta}\|_2\ge K_1}\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta})) + \inf_{\|\boldsymbol{\beta}\|_2\ge K_2}\lambda\sum_{j=1}^p \frac{|\beta_j|}{|\widehat{\beta}_j^{\text{init}}|}. \end{aligned}$$

Because of (19), we only need to deal with the second term. Since $\|\boldsymbol{\beta}\|_2 \ge K_2$ implies that there exists an element β_j of $\boldsymbol{\beta}$ such that $|\beta_j| \ge K_2/\sqrt{p}$ for some j, we have that

$$\inf_{\|\boldsymbol{\beta}\|_2\ge K_2}\lambda\sum_{j=1}^p \frac{|\beta_j|}{|\widehat{\beta}_j^{\text{init}}|} \ge \lambda\frac{|\beta_j|}{|\widehat{\beta}_j^{\text{init}}|} \ge \lambda\frac{|\beta_j|}{\|\widehat{\boldsymbol{\beta}}^{\text{init}}\|_2} \ge \lambda\frac{|\beta_j|}{k_1} \ge \frac{\lambda K_2}{\sqrt{p}k_1} = \lambda p.$$

□

Proof of Theorem 5.1. Suppose $\epsilon < \min\left\{\mathrm{BP}(\widehat{\boldsymbol{\beta}}^{\text{init}}, \mathbf{P}_{n-m}),\ \frac{1-2a_{nm}}{2-2a_{nm}}\right\}$. For a contaminated sample $\mathbf{P}_n$ and for $\frac{m}{n} \le \epsilon$ we have that, according to Lemma 7.2, if there exists a K such that $\|\widehat{\boldsymbol{\beta}}^{\text{S-NNG}}\|_2 \ge K$,

$$\widehat{\sigma}(\mathbf{r}(\widehat{\boldsymbol{\beta}}^{\text{S-NNG}})) + \lambda\sum_{j=1}^p \frac{|\widehat{\beta}_j^{\text{S-NNG}}|}{|\widehat{\beta}_j^{\text{init}}|} > \widehat{\sigma}(\mathbf{r}(\widehat{\boldsymbol{\beta}}^{\text{init}})) + \lambda p.$$

Since this is a contradiction to the fact that $\widehat{\boldsymbol{\beta}}^{\text{S-NNG}}$ minimizes $\left\{ \widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta})) + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\widehat{\beta}_j^{\text{init}}|} \right\}$ for $\boldsymbol{\beta} \in \mathbb{R}^p$, we have that

$$\text{BP}(\widehat{\boldsymbol{\beta}}^{\text{S-NNG}}, \mathbf{P}_{n-m}) \geq \min \left\{ \text{BP}(\widehat{\boldsymbol{\beta}}^{\text{init}}, \mathbf{P}_{n-m}),\ \frac{1 - a_{nm}}{2 - 2a_{nm}} \right\}.$$

□

7.3 Proof of Theorem 6.1

Proof of Theorem 6.1. By the chainrule, we have that the influence function of $\boldsymbol{\beta}_j^{\text{S-NNG}}(F)$, $j = 1, \ldots, p$, is given by

$$\begin{aligned}
\text{IF}(P_0, \beta_j^{\text{S-NNG}}, F) &= \frac{\partial}{\partial \epsilon}\, \beta_j^{\text{S-NNG}}(F_\epsilon)\big|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon}\, c_j^{\text{S-NNG}}(F_\epsilon)\big|_{\epsilon=0} \beta_j^{\text{init}}(F) + c_j^{\text{S-NNG}}(F) \frac{\partial}{\partial \epsilon}\, \beta_j^{\text{init}}(F_\epsilon)\Big|_{\epsilon=0} \\
&= \text{IF}(P_0, c_j^{\text{S-NNG}}, F) \beta_j^{\text{init}}(F) + c_j^{\text{S-NNG}}(F)\, \text{IF}(P_0, \beta_j^{\text{init}}, F) \qquad (20)
\end{aligned}$$

with $F_\epsilon := (1 - \epsilon)F + \epsilon \delta_{P_0}$.

In this proof we denote the set of indices containing the non-zero regression coefficients of $\boldsymbol{\beta}^{\text{S-NNG}}(F)$ with $\mathcal{S}_\lambda = \{j :\ \beta_j^{\text{S-NNG}}(F) \neq 0\}$ and the set of indices containing the zero regression coefficients of $\boldsymbol{\beta}^{\text{S-NNG}}(F)$ with $\mathcal{N}_\lambda = \{j :\ \beta_j^{\text{S-NNG}}(F) = 0\}$.

Note that the necessary conditions (Karush-Kuhn-Tucker conditions; see Kuhn and Tucker (1951) or Boyd and Vandenberghe (2004), among others) for minimizing (11) are

$$\begin{aligned}
\frac{\partial S}{\partial c_j} + \lambda - \mu_j &= 0 \qquad \text{for } j = 1, \ldots, p \qquad (21) \\
c_j &\geq 0 \qquad \text{for } j = 1, \ldots, p \\
\mu_j c_j &= 0 \qquad \text{for } j = 1, \ldots, p \\
\mu_j &\geq 0 \qquad \text{for } j = 1, \ldots, p,
\end{aligned}$$

where the μ_j are the KKT (Karush-Kuhn-Tucker) multipliers corresponding to the positivity constraint on the c_j.

First we calculate the influence function of $\beta_j^{\text{S-NNG}}(F)$ for $j \in \mathcal{N}_\lambda$. There are two possible ways to have that $\beta_j^{\text{S-NNG}}(F) = 0$. In the first case we have that $\beta_j^{\text{init}}(F) = 0$. Since we then also set $c_j^{\text{S-NNG}}(F)$ equal to zero, we have, by (20), that $\text{IF}(P_0, \beta_j^{\text{S-NNG}}, F) = 0$. In the second case $\beta_j^{\text{init}}(F) \neq 0$, but by the choice of the KKT multipliers μ_j we have that $c_j^{\text{S-NNG}}(F) = 0$. Since $c_j^{\text{S-NNG}}(F)$ is continuous this implies that $\text{IF}(P_0, c_j^{\text{S-NNG}}, F) = 0$ and by (20) that $\text{IF}(P_0, \boldsymbol{\beta}_j^{\text{S-NNG}}, F) = 0$. Hence, for all $j \in \mathcal{N}_\lambda$ we have that $\text{IF}(P_0, \beta_j^{\text{S-NNG}}, F) = 0$.

To find the expressions for the influence functions of the S-nonnegative garrote shrinkage factors $c_j^{\text{S-NNG}}(F)$, for $j \in \mathcal{S}_\lambda$, and $\sigma(F)$, we first differentiate the estimating equations of $c_j^{\text{S-NNG}}(F)$ and $\sigma(F)$ at the contaminated model with distribution F_ϵ with respect to ϵ and then take the limit of these expressions for ϵ going to zero. The influence functions of the regression coefficients $\beta_j^{\text{S-NNG}}(F)$ for $j \in \mathcal{S}_\lambda$ are then obtained by (20). The estimating equations of the S-nonnegative garrote shrinkage factors at the population level can be derived in a similar way as these at the sample level. Note that the quantity u_F in (13) equals $u_F = (Y - \boldsymbol{\mathcal{Z}}^{\text{T}}(F)\mathbf{c}^{\text{S-NNG}}(F))/\sigma(F)$.

If $j \in \mathcal{S}_\lambda$, then the KKT multiplier μ_j equals zero and (21) reduces to

$$\frac{\partial S}{\partial c_j} + \lambda = 0,$$

where $\dfrac{\partial S}{\partial c_j} = -\displaystyle\int \psi\left(u_F\right) Z_j(F)\,\mathrm{d}F \Big/ \int \psi\left(u_F\right) u_F\,\mathrm{d}F$, which can be obtained by taking the derivative of equation (12) with respect to c_j. See Gijbels and Vrinssen (2015) for more details.

Hence, $c_j^{\text{S-NNG}}(F)$ for $j \in \mathcal{S}_\lambda$ and $\sigma(F)$ can be represented by the following equations:

$$-\frac{\displaystyle\int \psi\left(u_F\right) Z_j(F)\,\mathrm{d}F}{\displaystyle\int \psi\left(u_F\right) u_F\,\mathrm{d}F} + \lambda = 0, \qquad \text{for } j \in \mathcal{S}_\lambda, \tag{22}$$

and

$$\int \rho(u_F)\,\mathrm{d}F = b. \tag{23}$$

We first derive the expression for the influence function of $\sigma(F)$ and then these for $c_j^{\text{S-NNG}}(F)$, $j \in \mathcal{S}_\lambda$. We finalize the proof by combining the two obtained equations.

1. By differentiating equation (23) at the contaminated model with distribution F_ϵ with respect to ϵ, we have, since $\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\mathbf{c}^{\text{S-NNG}}(F) = \boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{\beta}^{\text{S-NNG}}(F)$, that

$$\begin{aligned}
0 =& \frac{\partial}{\partial\epsilon}\left[(1-\epsilon)\int \rho(u_{F_\epsilon})\,\mathrm{d}F + \epsilon\rho\left(\frac{Y_0 - \mathbf{X}_0^{\mathrm{T}}\boldsymbol{\beta}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right)\right] \\
=& -\int \rho(u_{F_\epsilon})\,\mathrm{d}F + \rho\left(\frac{Y_0 - \mathbf{X}_0^{\mathrm{T}}\boldsymbol{\beta}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right) \\
&+ (1-\epsilon)\int \psi(u_{F_\epsilon})\left(-\frac{\boldsymbol{\mathcal{X}}^{\mathrm{T}}}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\boldsymbol{\beta}^{\text{S-NNG}}(F_\epsilon) - \frac{u_{F_\epsilon}}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\sigma(F_\epsilon)\right)\mathrm{d}F \\
&- \epsilon\psi\left(\frac{Y_0 - \mathbf{X}_0^{\mathrm{T}}\boldsymbol{\beta}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right)\frac{\mathbf{X}_0^{\mathrm{T}}}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\boldsymbol{\beta}^{\text{S-NNG}}(F_\epsilon) \\
&- \epsilon\psi\left(\frac{Y_0 - \mathbf{X}_0^{\mathrm{T}}\boldsymbol{\beta}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right)\frac{Y_0 - \mathbf{X}_0^{\mathrm{T}}\boldsymbol{\beta}^{\text{S-NNG}}(F_\epsilon)}{\sigma^2(F_\epsilon)}\frac{\partial}{\partial\epsilon}\sigma(F_\epsilon).
\end{aligned}$$

Letting $\epsilon \to 0$, we obtain

$$0 = -\int \rho(u_F)\,\mathrm{d}F + \rho\left(\frac{r_0}{\sigma(F)}\right) + \int \psi(u_F)\left(-\frac{\boldsymbol{\mathcal{X}}^{\mathrm{T}}}{\sigma(F)}\,\mathrm{IF}(P_0, \boldsymbol{\beta}^{\text{S-NNG}}, F) - \frac{u_F}{\sigma(F)}\,\mathrm{IF}(P_0, \sigma, F)\right)\mathrm{d}F,$$

where $r_0 = Y_0 - \mathbf{X}_0^{\mathrm{T}}\boldsymbol{\beta}^{\text{S-NNG}}(F)$. Using (23), we get

$$\mathrm{IF}(P_0, \sigma, F) = \frac{\sigma(F)\left(\rho\left(\dfrac{r_0}{\sigma(F)}\right) - b\right) - \displaystyle\int \psi(u_F)\boldsymbol{\mathcal{X}}^{\mathrm{T}}\,\mathrm{d}F\ \mathrm{IF}(P_0, \boldsymbol{\beta}^{\text{S-NNG}}, F)}{\displaystyle\int \psi(u_F)u_F\,\mathrm{d}F}. \tag{24}$$

2. At the contaminated model with distribution F_ϵ (22) yields for $j \in \mathcal{S}_\lambda$,

$$0 = \lambda + \frac{N(F_\epsilon)}{D(F_\epsilon)},$$

where

$$\begin{aligned}
N(F_\epsilon) =& -(1-\epsilon)\int \psi\left(u_{F_\epsilon}\right) Z_j(F_\epsilon)\,\mathrm{d}F - \epsilon\psi\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right) Z_{0j}(F_\epsilon),\\
D(F_\epsilon) =& (1-\epsilon)\int \psi\left(u_{F_\epsilon}\right) u_{F_\epsilon}\,\mathrm{d}F + \epsilon\psi\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right)\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)},
\end{aligned}$$

with $\mathbf{Z}_0(F_\epsilon) = \mathrm{diag}(\beta^{\text{init}}(F_\epsilon))\mathbf{X}_0$. Differentiating with respect to ϵ gives

$$\begin{aligned}
0 =& \frac{1}{D(F_\epsilon)}\int \psi\left(u_{F_\epsilon}\right) Z_j(F_\epsilon)\,\mathrm{d}F - \frac{1}{D(F_\epsilon)}\psi\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right) Z_{0j}(F_\epsilon)\\
&+ \frac{1-\epsilon}{D(F_\epsilon)}\int \psi'\left(u_{F_\epsilon}\right) Z_j(F_\epsilon)\left(\frac{\partial}{\partial\epsilon}\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F_\epsilon)\frac{\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)} + \frac{\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F_\epsilon)}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\mathbf{c}^{\text{S-NNG}}(F_\epsilon)\right)\mathrm{d}F\\
&+ \frac{1-\epsilon}{D(F_\epsilon)}\int \psi'\left(u_{F_\epsilon}\right) Z_j(F_\epsilon)\frac{u_{F_\epsilon}}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\sigma(F_\epsilon)\,\mathrm{d}F - \frac{1-\epsilon}{D(F_\epsilon)}\int \psi\left(u_{F_\epsilon}\right)\frac{\partial}{\partial\epsilon}Z_j(F_\epsilon)\,\mathrm{d}F\\
&+ \frac{\epsilon}{D(F_\epsilon)}\psi'\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right) Z_{0j}(F_\epsilon)\left(\frac{\partial}{\partial\epsilon}\mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\frac{\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)} + \frac{\mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\mathbf{c}^{\text{S-NNG}}(F_\epsilon)\right)\\
&+ \frac{\epsilon}{D(F_\epsilon)}\psi'\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right) Z_{0j}(F_\epsilon)\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma^2(F_\epsilon)}\frac{\partial}{\partial\epsilon}\sigma(F_\epsilon)\\
&- \frac{\epsilon}{D(F_\epsilon)}\psi\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right)\frac{\partial}{\partial\epsilon}Z_{0j}(F_\epsilon)\\
&- \frac{N(F_\epsilon)}{D^2(F_\epsilon)}\Bigg\{-\int \psi\left(u_{F_\epsilon}\right) u_{F_\epsilon}\,\mathrm{d}F + \psi\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right)\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\\
&\qquad - (1-\epsilon)\int \psi'\left(u_{F_\epsilon}\right) u_{F_\epsilon}\left(\frac{\partial}{\partial\epsilon}\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F_\epsilon)\frac{\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)} + \frac{\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F_\epsilon)}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\mathbf{c}^{\text{S-NNG}}(F_\epsilon)\right)\mathrm{d}F\\
&\qquad - (1-\epsilon)\int \psi'\left(u_{F_\epsilon}\right) u_{F_\epsilon}\frac{u_{F_\epsilon}}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\sigma(F_\epsilon)\,\mathrm{d}F - (1-\epsilon)\int \psi\left(u_{F_\epsilon}\right)\frac{u_{F_\epsilon}}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\sigma(F_\epsilon)\,\mathrm{d}F\\
&\qquad - (1-\epsilon)\int \psi\left(u_{F_\epsilon}\right)\left(\frac{\partial}{\partial\epsilon}\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F_\epsilon)\frac{\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)} + \frac{\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F_\epsilon)}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\mathbf{c}^{\text{S-NNG}}(F_\epsilon)\right)\mathrm{d}F\\
&\qquad - \epsilon\psi'\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right)\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\frac{\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\\
&\qquad - \epsilon\psi'\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right)\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\frac{\mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\mathbf{c}^{\text{S-NNG}}(F_\epsilon)\\
&\qquad - \epsilon\psi'\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right)\frac{\left(Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)\right)^2}{\sigma^3(F_\epsilon)}\frac{\partial}{\partial\epsilon}\sigma(F_\epsilon)\\
&\qquad - \epsilon\psi\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right)\left(\frac{\partial}{\partial\epsilon}\mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\frac{\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)} + \frac{\mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)}{\sigma(F_\epsilon)}\frac{\partial}{\partial\epsilon}\mathbf{c}^{\text{S-NNG}}(F_\epsilon)\right)\\
&\qquad - \epsilon\psi\left(\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma(F_\epsilon)}\right)\frac{Y_0 - \mathbf{Z}_0^{\mathrm{T}}(F_\epsilon)\mathbf{c}^{\text{S-NNG}}(F_\epsilon)}{\sigma^2(F_\epsilon)}\frac{\partial}{\partial\epsilon}\sigma(F_\epsilon)\Bigg\},
\end{aligned}$$

where $\frac{\partial}{\partial\epsilon}\boldsymbol{\mathcal{Z}}(F_\epsilon) = \operatorname{diag}\left(\frac{\partial}{\partial\epsilon}\boldsymbol{\beta}^{\text{init}}(F_\epsilon)\right)\boldsymbol{\mathcal{X}}$. Now let $\epsilon \to 0$,

$$
\begin{aligned}
0 =& \frac{\int \psi(u_F) Z_j(F)\,\mathrm{d}F}{\int \psi(u_F) u_F\,\mathrm{d}F} - \frac{\psi\left(\frac{r_0}{\sigma(F)}\right) Z_{0j}(F)}{\int \psi(u_F) u_F\,\mathrm{d}F} \\
&+ \frac{\int \psi'(u_F)\, Z_j(F)\left(\operatorname{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}} \mathbf{c}^{\text{S-NNG}}(F) + \boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\operatorname{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F)\right)\mathrm{d}F}{\sigma(F)\int \psi(u_F) u_F\,\mathrm{d}F} \\
&+ \frac{\int \psi'(u_F)\, u_F Z_j(F)\,\mathrm{d}F}{\sigma(F)\int \psi(u_F) u_F\,\mathrm{d}F}\operatorname{IF}(P_0, \sigma, F) - \frac{\int \psi(u_F)\ \operatorname{IF}(P_0, Z_j, F)\,\mathrm{d}F}{\int \psi(u_F) u_F\,\mathrm{d}F} \\
&- \frac{\int \psi(u_F) Z_j(F)\,\mathrm{d}F}{\int \psi(u_F) u_F\,\mathrm{d}F} + \frac{\int \psi(u_F) Z_j(F)\,\mathrm{d}F}{\left(\int \psi(u_F) u_F\,\mathrm{d}F\right)^2}\psi\left(\frac{r_0}{\sigma(F)}\right)\frac{r_0}{\sigma(F)} \\
&- \frac{\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi'(u_F) u_F\left(\operatorname{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}} \mathbf{c}^{\text{S-NNG}}(F) + \boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\operatorname{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F)\right)\mathrm{d}F}{\sigma(F)\left(\int \psi(u_F) u_F\,\mathrm{d}F\right)^2} \\
&- \frac{\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi'(u_F) u_F^2\,\mathrm{d}F}{\sigma(F)\left(\int \psi(u_F) u_F\,\mathrm{d}F\right)^2}\operatorname{IF}(P_0, \sigma, F) - \frac{\int \psi(u_F) Z_j(F)\,\mathrm{d}F}{\sigma(F)\int \psi(u_F) u_F\,\mathrm{d}F}\operatorname{IF}(P_0, \sigma, F) \\
&- \frac{\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi(u_F)\left(\operatorname{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}} \mathbf{c}^{\text{S-NNG}}(F) + \boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\operatorname{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F)\right)\mathrm{d}F}{\sigma(F)\left(\int \psi(u_F) u_F\,\mathrm{d}F\right)^2},
\end{aligned}
\tag{25}
$$

where $\operatorname{IF}(P_0, \boldsymbol{\mathcal{Z}}, F) = \operatorname{diag}(\operatorname{IF}(P_0, \boldsymbol{\beta}^{\text{init}}, F))\boldsymbol{\mathcal{X}}$.

3. The proof is then completed by plugging (24) (in which we replaced $\boldsymbol{\mathcal{X}}^{\mathrm{T}}\,\mathrm{IF}(P_0, \boldsymbol{\beta}^{\text{S-NNG}}, F)$ with $\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}}\mathbf{c}^{\text{S-NNG}}(F) + \boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\,\mathrm{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F)$) into (25). We get

$$
\begin{aligned}
0 = &-\frac{1}{\mu_1}\psi\left(\frac{r_0}{\sigma(F)}\right) Z_{0j}(F) \\
&+\frac{1}{\sigma(F)\mu_1}\int \psi'\left(u_F\right) Z_j(F)\left(\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}}\mathbf{c}^{\text{S-NNG}}(F) + \boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\,\mathrm{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F)\right)\mathrm{d}F \\
&+\frac{1}{\mu_1^2}\int \psi'\left(u_F\right) u_F Z_j(F)\,\mathrm{d}F\left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right) \\
&-\frac{1}{\sigma(F)\mu_1^2}\int \psi'\left(u_F\right) u_F Z_j(F)\,\mathrm{d}F \int \psi(u_F)\,\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}}\mathbf{c}^{\text{S-NNG}}(F)\,\mathrm{d}F \\
&-\frac{1}{\sigma(F)\mu_1^2}\int \psi'\left(u_F\right) u_F Z_j(F)\,\mathrm{d}F \int \psi(u_F)\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\,\mathrm{d}F\,\mathrm{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F) \\
&-\frac{1}{\mu_1}\int \psi(u_F)\,\mathrm{IF}(P_0, Z_j, F)\,\mathrm{d}F + \frac{1}{\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F\psi\left(\frac{r_0}{\sigma(F)}\right)\frac{r_0}{\sigma(F)} \\
&-\frac{1}{\sigma(F)\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi'(u_F) u_F\,\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}}\mathbf{c}^{\text{S-NNG}}(F)\,\mathrm{d}F \\
&-\frac{1}{\sigma(F)\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi'(u_F) u_F\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\,\mathrm{d}F\,\mathrm{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F) \\
&-\frac{\nu_2}{\mu_1^3}\int \psi(u_F) Z_j(F)\,\mathrm{d}F\left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right) \\
&+\frac{\nu_2}{\sigma(F)\mu_1^3}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi(u_F)\,\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}}\mathbf{c}^{\text{S-NNG}}(F)\,\mathrm{d}F \\
&+\frac{\nu_2}{\sigma(F)\mu_1^3}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi(u_F)\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\,\mathrm{d}F\,\mathrm{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F) \\
&-\frac{1}{\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F\left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right) \\
&+\frac{1}{\sigma(F)\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi(u_F)\,\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}}\mathbf{c}^{\text{S-NNG}}(F)\,\mathrm{d}F \\
&+\frac{1}{\sigma(F)\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi(u_F)\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\,\mathrm{d}F\,\mathrm{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F) \\
&-\frac{1}{\sigma(F)\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi(u_F)\,\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}}\mathbf{c}^{\text{S-NNG}}(F)\,\mathrm{d}F \\
&-\frac{1}{\sigma(F)\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi(u_F)\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\,\mathrm{d}F\,\mathrm{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F).
\end{aligned}
$$

Noting that the last four terms cancel out and regrouping all terms in $\mathrm{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F)$, we get

$$
\begin{aligned}
&\Bigg\{+\frac{1}{\sigma(F)\mu_1}\int \psi'\left(u_F\right) Z_j(F)\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\,\mathrm{d}F - \frac{1}{\sigma(F)\mu_1^2}\int \psi'\left(u_F\right) u_F Z_j(F)\,\mathrm{d}F \int \psi(u_F)\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\,\mathrm{d}F \\
&\quad -\frac{1}{\sigma(F)\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi'(u_F) u_F \boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\,\mathrm{d}F \\
&\quad +\frac{\nu_2}{\sigma(F)\mu_1^3}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi(u_F)\boldsymbol{\mathcal{Z}}^{\mathrm{T}}(F)\,\mathrm{d}F\Bigg\}\ \mathrm{IF}(P_0, \mathbf{c}^{\text{S-NNG}}, F) \\
&\quad = \frac{1}{\mu_1}\psi\left(\frac{r_0}{\sigma(F)}\right) Z_{0j}(F) - \frac{1}{\mu_1^2}\int \psi'\left(u_F\right) u_F Z_j(F)\,\mathrm{d}F\left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right) \\
&\qquad - \frac{1}{\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \psi\left(\frac{r_0}{\sigma(F)}\right)\frac{r_0}{\sigma(F)} + \frac{\nu_2}{\mu_1^3}\int \psi(u_F) Z_j(F)\,\mathrm{d}F\left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right) \\
&\qquad + \frac{1}{\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F\left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right) + \frac{1}{\mu_1}\int \psi(u_F)\,\mathrm{IF}(P_0, Z_j, F)\,\mathrm{d}F \\
&\qquad - \frac{1}{\sigma(F)\mu_1}\int \psi'\left(u_F\right) Z_j(F)\,\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}}\mathbf{c}^{\text{S-NNG}}(F)\ \mathrm{d}F \\
&\qquad + \frac{1}{\sigma(F)\mu_1^2}\int \psi'\left(u_F\right) u_F Z_j(F)\,\mathrm{d}F \int \psi(u_F)\,\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}}\mathbf{c}^{\text{S-NNG}}(F)\ \mathrm{d}F \\
&\qquad + \frac{1}{\sigma(F)\mu_1^2}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi'(u_F) u_F\,\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}}\mathbf{c}^{\text{S-NNG}}(F)\ \mathrm{d}F \\
&\qquad - \frac{\nu_2}{\sigma(F)\mu_1^3}\int \psi(u_F) Z_j(F)\,\mathrm{d}F \int \psi(u_F)\,\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}, F)^{\mathrm{T}}\mathbf{c}^{\text{S-NNG}}(F)\ \mathrm{d}F.
\end{aligned}
$$

To obtain the expressions for the influence functions of $c_j^{\text{S-NNG}}(F)$ for $j \in \mathcal{S}_\lambda$, we take all the equations for $c_j^{\text{S-NNG}}(F)$ for $j \in \mathcal{S}_\lambda$ together and use that $\boldsymbol{\mathcal{Z}}(F) = \mathrm{diag}\left(\boldsymbol{\beta}^{\text{init}}(F)\right)\boldsymbol{\mathcal{X}}$. Since $\mathrm{IF}(P_0, \mathbf{c}_{\mathcal{N}_\lambda}^{\text{S-NNG}}, F) = 0$, we find

$$
\begin{aligned}
\mathrm{IF}(P_0,&\mathbf{c}_{\mathcal{S}_\lambda}^{\text{S-NNG}}, F) = \mathrm{diag}\left(\boldsymbol{\beta}_{\mathcal{S}_\lambda}^{\text{init}}(F)\right)^{-1}\Pi_{\mathcal{S}_\lambda}\mathrm{diag}\left(\boldsymbol{\beta}_{\mathcal{S}_\lambda}^{\text{init}}(F)\right)^{-1}\mathrm{diag}\left(\boldsymbol{\beta}_{\mathcal{S}_\lambda}^{\text{init}}(F)\right)\Bigg[\frac{1}{\mu_1}\psi\left(\frac{r_0}{\sigma(F)}\right)\mathbf{X}_{0\mathcal{S}_\lambda} \\
&- \frac{1}{\mu_1^2}\int \psi'\left(u_F\right) u_F \boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}\,\mathrm{d}F\left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right) - \frac{1}{\mu_1^2}\int \psi(u_F)\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}\,\mathrm{d}F\psi\left(\frac{r_0}{\sigma(F)}\right)\frac{r_0}{\sigma(F)} \\
&+ \frac{\nu_2}{\mu_1^3}\int \psi(u_F)\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}\,\mathrm{d}F\left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right) + \frac{1}{\mu_1^2}\int \psi(u_F)\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}\,\mathrm{d}F\left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right) \\
&+ \lambda\mathrm{diag}\left(\boldsymbol{\beta}_{\mathcal{S}_\lambda}^{\text{init}}(F)\right)^{-2}\mathrm{IF}(P_0, \boldsymbol{\beta}_{\mathcal{S}_\lambda}^{\text{init}}, F) - \Pi_{\mathcal{S}_\lambda}^{-1}\mathrm{diag}\left(\mathrm{IF}(P_0, \boldsymbol{\beta}_{\mathcal{S}_\lambda}^{\text{init}}, F)\right)\mathbf{c}_{\mathcal{S}_\lambda}^{\text{S-NNG}}(F)\Bigg],
\end{aligned}
$$

where

$$
\begin{aligned}
\Pi_{\mathcal{S}_\lambda} = \Bigg[&\frac{1}{\sigma(F)\mu_1}\int \psi'\left(u_F\right)\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}^{\mathrm{T}}\,\mathrm{d}F - \frac{1}{\sigma(F)\mu_1^2}\int \psi'\left(u_F\right) u_F\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}\,\mathrm{d}F\int \psi(u_F)\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}^{\mathrm{T}}\,\mathrm{d}F \\
&- \frac{1}{\sigma(F)\mu_1^2}\int \psi(u_F)\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}\,\mathrm{d}F\int \psi'(u_F) u_F\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}^{\mathrm{T}}\,\mathrm{d}F + \frac{\nu_2}{\sigma(F)\mu_1^3}\int \psi(u_F)\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}\,\mathrm{d}F\int \psi(u_F)\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}^{\mathrm{T}}\,\mathrm{d}F\Bigg]^{-1}.
\end{aligned}
$$

Note that the term $\lambda\mathrm{diag}\left(\boldsymbol{\beta}_{\mathcal{S}_\lambda}^{\text{init}}(F)\right)^{-2}\mathrm{IF}(P_0, \boldsymbol{\beta}_{\mathcal{S}_\lambda}^{\text{init}}, F)$ is obtained by using (22) on the term $\frac{1}{\mu_1}\int \psi(u_F)\,\mathrm{IF}(P_0, \boldsymbol{\mathcal{Z}}_{\mathcal{S}_\lambda}, F)\,\mathrm{d}F$.

If we now use equation (20) and the following notations,

$$
\begin{aligned}
\mathbf{A}_{\mathcal{S}_\lambda} &= \int \psi'\left(u_F\right)\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}^{\mathrm{T}}\,\mathrm{d}F, \qquad \mathbf{a}_{F\mathcal{S}_\lambda} = \int \psi(u_F)\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}\,\mathrm{d}F, \\
\mathbf{b}_{F\mathcal{S}_\lambda} &= \int \psi'\left(u_F\right) u_F\boldsymbol{\mathcal{X}}_{\mathcal{S}_\lambda}\,\mathrm{d}F,
\end{aligned}
$$

we obtain

$$\left\{ \frac{\mathbf{A}_{\mathcal{S}_\lambda}}{\sigma(F)\mu_1} - \frac{\mathbf{b}_{F\mathcal{S}_\lambda}\mathbf{a}_{F\mathcal{S}_\lambda}^{\mathrm{T}}}{\sigma(F)\mu_1^2} - \frac{\mathbf{a}_{F\mathcal{S}_\lambda}\mathbf{b}_{F\mathcal{S}_\lambda}^{\mathrm{T}}}{\sigma(F)\mu_1^2} + \frac{\nu_2\mathbf{a}_{F\mathcal{S}_\lambda}\mathbf{a}_{F\mathcal{S}_\lambda}^{\mathrm{T}}}{\sigma(F)\mu_1^3} \right\} \mathrm{IF}(P_0, \boldsymbol{\beta}_{\mathcal{S}_\lambda}^{\text{S-NNG}}, F) =$$
$$\psi\left(\frac{r_0}{\sigma(F)}\right)\frac{\mathbf{X}_{0\mathcal{S}_\lambda}}{\mu_1} - \left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right)\frac{\mathbf{b}_{F\mathcal{S}_\lambda}}{\mu_1^2} - \psi\left(\frac{r_0}{\sigma(F)}\right)\frac{r_0}{\sigma(F)}\frac{\mathbf{a}_{F\mathcal{S}_\lambda}}{\mu_1^2}$$
$$+ \left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right)\frac{\nu_2\mathbf{a}_{F\mathcal{S}_\lambda}}{\mu_1^3} + \left(\rho\left(\frac{r_0}{\sigma(F)}\right) - b\right)\frac{\mathbf{a}_{F\mathcal{S}_\lambda}}{\mu_1^2} + \lambda\mathrm{diag}\left(\left(\boldsymbol{\beta}_{\mathcal{S}_\lambda}^{\text{init}}(F)\right)^{-2}\right)\mathrm{IF}(P_0, \boldsymbol{\beta}_{\mathcal{S}_\lambda}^{\text{init}}, F).$$

□

Acknowledgement. The authors thank the Editor, an Associate Editor and two referees for their valuable comments which led to a considerable improvement of the paper. This research is supported by the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy), project GOA/12/014 of the Research Fund of the KU Leuven, and FWO research grant 1.5.137.13N.

References

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

Antoniadis, A., Gijbels, I., and Verhasselt, A. (2012a). Variable selection in additive models using P-splines. *Technometrics*, 54(4):425–438.

Antoniadis, A., Gijbels, I., and Verhasselt, A. (2012b). Variable selection in varying-coefficient models using P-splines. *Journal of Computational and Graphical Statistics*, 21(3):638–661.

Arslan, O. (2012). Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. *Computational Statistics & Data Analysis*, 56(6):1952–1965.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Chatterjee, A. and Lahiri, S. N. (2011). Strong consistency of Lasso estimators. *Sankhya A: The Indian Journal of Statistics*, 73(1):55–78.

Donoho, D. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 157–184. Wadsworth, Belmont, CA.

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

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Gijbels, I. and Vrinssen, I. (2015). Robust nonnegative garrote variable selection in linear regression. *Computational Statistics & Data Analysis*, 85:1–22.

Ip, W., Yang, Y., Kwan, P., and Kwan, Y. (2003). Strong convergence rate of the least median absolute estimator in linear regression models. *Statistical Papers*, 44:183–201.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. *Proceedings of 2nd Berkeley Symposium*, pages 481–492. Berkeley: University of California Press.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. Theory and methods.

Maronna, R. A. and Yohai, V. J. (1981). Asymptotic behavior of general M-estimates for regression and scale with random carriers. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 58(1):7–20.

Öllerer, V., Croux, C., and Alfons, A. (2015). The influence function of penalized regression estimators. *Statistics*, 49(4):741–765.

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

Wagener, J. and Dette, H. (2013). The adaptive lasso in high-dimensional sparse heteroscedastic models. *Mathematical Methods of Statistics*, 22(2):137–154.

Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business & Economic Statistics*, 25(3):347–355.

Wang, L. and Li, R. (2009). Weighted Wilcoxon-type smoothly clipped absolute deviation method. *Biometrics. Journal of the International Biometric Society*, 65(2):564–571.

Wang, X., Jiang, Y., Huang, M., and Zhang, H. (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 108(502):632–643.

Xiong, S. and Joseph, V. (2013). Regression with outlier shrinkage. *Journal of Statistical Planning and Inference*, 143:1988–2001.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2):642–656.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, 83(402):406–413.

Yohai, V. J. and Zamar, R. H. (1997). Optimal locally robut m-estimates of regression. *Journal of Statistical Planning and Inference*, 64:309–323.

Yuan, M. and Lin, Y. (2007). On the non-negative garrote estimator. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 69(2):143–161.

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