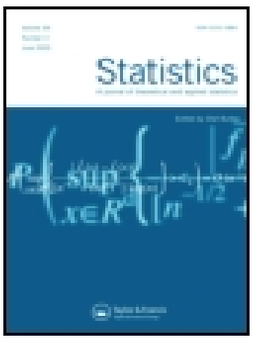

Taylor & Francis
Taylor & Francis Group

Statistics

A Journal of Theoretical and Applied Statistics

ISSN: 0233-1888 (Print) 1029-4910 (Online) Journal homepage: https://www.tandfonline.com/loi/gsta20

Robust estimation and variable selection in heteroscedastic linear regression

I. Gijbels & I. Vrinssen

To cite this article: I. Gijbels & I. Vrinssen (2019): Robust estimation and variable selection in heteroscedastic linear regression, Statistics, DOI: 10.1080/02331888.2019.1579215

To link to this article: https://doi.org/10.1080/02331888.2019.1579215

Published online: 18 Feb 2019.

Submit your article to this journal

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=gsta20

STATISTICS
https://doi.org/10.1080/02331888.2019.1579215

Taylor & Francis
Taylor & Francis Group

Check for updates

Robust estimation and variable selection in heteroscedastic linear regression

I. Gijbels and I. Vrinssen

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

ABSTRACT

The paper concerns robust estimation and variable selection in heteroscedastic linear regression models. After a brief review of existing methods for estimation in such models, a robust S-estimation approach is discussed. For all methods concise descriptions of algorithms are provided. Little is available upon robust variable selection methods for heteroscedastic linear models. The paper gives essential contributions in the area of simultaneous robust estimation and variable selection, relying on basics of the nonnegative garrote method which has been proven to have very good practical as well as theoretical properties in the homoscedastic linear model context. Several numerical examples, simulations and analysis of real data, demonstrate the performances and practical use of the discussed methods. Moreover, we provide expressions for the influence functions of the estimators of the mean and the error variance parameters. Influence functions are plotted in a simple setting providing insights in the sensitivity of the estimators for a single outlying observation.

ARTICLE HISTORY
Received 30 July 2017
Accepted 28 January 2019

KEYWORDS
Heteroscedasticity; influence function; nonnegative garrote; S-estimation; variable selection

1. Introduction

The interest in this paper is in a response variable Y, that is possibly influenced by p covariates, denoted by $(X_1, \ldots, X_p)$, via a linear regression model. The covariates are also allowed to have an impact on the error variance, and hence a heteroscedastic linear model is considered in which observations satisfy

$$Y_i = \sum_{j=0}^{p} X_{ij}\beta_j + \sigma_i \varepsilon_i = \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta} + \sigma_i \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where Y_i are response observations, $\mathbf{X}_i = (1, X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ are vectors with observations on the p covariates, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of unknown regression coefficients, $0 < \sigma_i < \infty$, and ε_i are independent and identically distributed error terms with $\mathrm{E}(\varepsilon_i \mid \mathbf{X}_i) = 0$ and $\mathrm{Var}(\varepsilon_i \mid \mathbf{X}_i) = 1$. With $\mathbf{A}^{\mathrm{T}}$ we denote the transpose of a vector or matrix $\mathbf{A}$. The fixed first component one in $\mathbf{X}_i$ is included for incorporating the unknown intercept parameter β_0 in the model.

CONTACT I. Gijbels irene.gijbels@wis.kuleuven.be Department of Mathematics and Leuven Statistics Research Center (LStat), KU Leuven, Leuven, Belgium

Supplemental data for this article can be accessed here. https://doi.org/10.1080/02331888.2019.1579215

© 2019 Informa UK Limited, trading as Taylor & Francis Group

Maximum likelihood estimation of regression coefficients in heteroscedastic linear regression models, such as model (1), has been discussed in the literature in, for example, Jobson and Fuller [1]. Such an approach requires specification of the error distribution, and its efficiency heavily depends on this correct specification. A commonly-assumed error distribution is a normal distribution. In case of a normal error the maximum likelihood method leads to an objective function which coincides with that of the ordinary least squares method. Robustness requirements against departures of the specified error distribution (such as normality) led to the development of robust estimation methods in heteroscedastic models, such as the M-estimation approach of Carroll and Ruppert [2], the weighted generalized M-estimation methods of Bianco et al. [3] and Bianco and Boente [4], or the forward search method of Atkinson et al. [5].

In robust estimation the aim is to have methods that are not (or less) sensitive to, on the one hand, *vertical outliers*, i.e., outliers in the error terms, and, on the other hand, to *leverage points*, i.e., observations that are outliers in the predictor space (i.e the space of the covariates). The robustness to leverage points leads to the so-called robust M-estimation methods. See, for example, Serneels et al. [6] for a brief discussion in case of homoscedastic linear regression.

Another possible approach in robust estimation procedures is to consider an S-type estimation method. See Rousseeuw and Yohai [7] and Yohai [8], among others. For an overview of robust regression methods for homoscedastic linear regression see Maronna et al. [9].

The interest in this paper is two-fold. On the one hand we would like to estimate the effect of the covariates on the mean response, as well as on the heteroscedasticity, but we would also like to select, *simultaneously*, the covariates that effectively influence these two parts. So the aim is robust estimation *and* variable selection in heteroscedastic linear models.

A common way to describe the heteroscedasticity σ_i in (1) is via functions such as, for example, $\sigma_i = \sigma|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}|^{\lambda}$ or $\sigma_i = \sigma(1 + |\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}|)^{\lambda}$ [10], $\sigma_i = \sigma\exp\{\lambda\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\}$ [11], or $\sigma_i = \sigma(1 + \lambda(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})^2)^{1/2}$ [1]. Bianco et al. [3] and Bianco and Boente [4] study robust generalized M-based estimators under a general framework where $\sigma_i = \sigma G(\mathbf{X}_i, \lambda, \boldsymbol{\beta})$, for a given function G.

In this paper we assume that the error variances equal

$$\sigma_i^2 = \sigma^2 h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)$$

and hence depend on the covariates $\mathbf{X}_{i-} = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ via a vector of coefficients $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^{\mathrm{T}}$ through a predefined strictly positive, one-to-one and twice differentiable function h for which $h(0) = 1$ (to ensure identifiability of the model). We want the function h to be one-to-one to avoid computational problems. Examples of an appropriate function h are $h(x) = \exp(x)$ or $h(x) = \sqrt{1 + x^2}$. We denote the vector of response observations with $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, the design matrix with $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{\mathrm{T}}$, the design matrix without intercept column (the column of one's) with $\mathbf{X}_- = (\mathbf{X}_{1-}, \ldots, \mathbf{X}_{n-})^{\mathrm{T}}$, and the vector that collects all unknown coefficients with $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \sigma)^{\mathrm{T}}$. Note that $\boldsymbol{\theta}$ is a column vector of dimension $2p+2$.

We want to select and estimate the relevant coefficients in both the mean function $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ and the variance function $h(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma})$, as well as to estimate the parameter σ^2, and this all in

one action. There are ample of variable selection methods, and the choice among them is influenced by many aspects, from theoretical aspects (consistency, . . .) to computational aspects (algorithms and implementation procedures that are available or can be developed). In this paper we opt for using as a basis the nonnegative garrote method (which is related to an adaptive lasso method), since robust versions of this method have been developed for homoscedastic linear models, revealing nice theoretical and practical properties. See Antoniadis et al. [12,13] and Gijbels et al. [14], among others. Among the studied methods is a heteroscedastic S-nonnegative garrote method, which shows very good estimation and variable selection performance. See Sections 3.2 and 4. With appropriate weighting the influence function of the heteroscedastic S-nonnegative garrote method is bounded for regression outliers. See Section 5.

The paper is further organized as follows. In Section 2 we discuss main estimation approaches in the considered heteroscedastic linear model. A new contribution here consists of the robust S-estimation approach described in Section 2.2.3. In Section 3.2 we then discuss several robust variable selection methods, with as common building block nonnegative garrote selection. For all methods discussed in Sections 2 and 3.2 algorithms are described in a concise manner. The performances of the different (robust) simultaneous estimation and selection methods are investigated in a simulation study in Section 4. A further illustration of the use of the developed methods is provided via the analyse of a real data example in Section 6. For each of the discussed (robust) estimation/selection methods we establish the influence functions in Section 5. These clearly reveal the impact of choices of initial estimators, and of the proposed weighting scheme that ensures robustness against vertical outliers and/or leverage points. A brief discussion section concludes the paper.

2. Estimation in heteroscedastic linear regression models

In this section we first review several methods that can handle heteroscedasticity. These methods can then be used as initial estimators for the (robust) heteroscedastic nonnegative garrote methods, in Section 3.2. We first review the maximum likelihood method which is a non-robust method. In Section 2.2.1 we discuss maximum trimmed likelihood estimation, which is together with heteroscedastic M-estimation (see Section 2.2.2) and heteroscedastic S-estimation (Section 2.2.3) part of the discussed robust estimation methods.

2.1. Non-robust maximum likelihood estimation

The maximum likelihood method, studied by Jobson and Fuller [1], is often used to estimate the coefficients in a heteroscedastic regression model. Suppose that the conditional density of ε_i given $\mathbf{X}_i$ is entirely known and denote it with f. It is easily seen that, under model (1), the conditional density of Y_i given $\mathbf{X}_i$, evaluated at the point y, is

$$\frac{1}{\sigma\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)}} f\left(\frac{y-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)}}\right).$$

Assume further that the marginal density of $(X_1, \ldots, X_p)$ is non-informative, meaning that it not depends on the unknown parameter $\boldsymbol{\theta}$. Based on the random sample

$(\mathbf{X}_1^{\mathrm{T}}, Y_1), \ldots, (\mathbf{X}_n^{\mathrm{T}}, Y_n)$, the informative part of the likelihood function for $\boldsymbol{\theta}$ then reduces to

$$L(\boldsymbol{\theta}) \equiv \prod_{i=1}^{n} L_i(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)}} f\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)}}\right).$$

The maximum likelihood (ML) estimator is the vector of size $2p+2$ that maximizes the logarithm of the likelihood function. Denoting $\ell(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \log L_i(\boldsymbol{\theta}) \equiv \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta})$, the maximum likelihood estimator for $\boldsymbol{\theta}$ is defined as

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left\{ \log\left(f\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)}}\right)\right) - \log(\sigma) - \frac{1}{2}\log\left(h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)\right)\right\},\end{aligned}$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}^{\mathrm{T}}, \hat{\sigma})^{\mathrm{T}}$, with $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^{\mathrm{T}}$ and $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)^{\mathrm{T}}$.

If we assume, just as for ordinary least-squares inference, that the error terms are standard normally distributed, then the ML-estimator equals

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left\{ -\frac{1}{2}\frac{\left(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2}{\sigma^2 h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)} - \log(\sigma) - \frac{1}{2}\log\left(h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)\right) - \frac{1}{2}\log(2\pi)\right\},$$

and the objective function coincides with the one encountered in ordinary least-squares.

To obtain the solution of the maximum likelihood method, its first order conditions are needed. These first order conditions for respectively $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and σ are given by

$$\begin{aligned}\mathbf{S}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}) &= \left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \sum_{i=1}^{n} \psi\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}}\right)\frac{\mathbf{X}_i}{\hat{\sigma}\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}} \\ &= \frac{1}{\hat{\sigma}^2}\mathbf{X}^{\mathrm{T}}\mathbf{G}(\hat{\boldsymbol{\gamma}})(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}_{p+1}, \qquad (2) \\ \mathbf{S}_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\theta}}) &= \left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \sum_{i=1}^{n}\left[\rho\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}}\right) - b\right]\frac{h'\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}\mathbf{X}_{i-} \\ &= \frac{1}{2\hat{\sigma}^2}\mathbf{X}_{-}^{\mathrm{T}}\mathbf{R}(\hat{\boldsymbol{\beta}})\mathbf{G}^2(\hat{\boldsymbol{\gamma}})\mathbf{F}(\hat{\boldsymbol{\gamma}})\mathbf{1}_n - \frac{1}{2}\mathbf{X}_{-}^{\mathrm{T}}\mathbf{G}(\hat{\boldsymbol{\gamma}})\mathbf{F}(\hat{\boldsymbol{\gamma}})\mathbf{1}_n = \mathbf{0}_p, \qquad (3) \\ \mathbf{S}_{\sigma}(\hat{\boldsymbol{\theta}}) &= \left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \frac{2}{\hat{\sigma}}\sum_{i=1}^{n}\left[\rho\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}}\right) - b\right] = 0, \qquad (4)\end{aligned}$$

with $\rho(u) = u^2/2$, $\psi(u) = \rho'(u) = u$ and $b = E(\rho(Z)) = 1/2$, with Z a standard normal distributed random variable. Further $\mathbf{1}_n$ is a column vector of length n with all elements

Table 1. Overview of notations used in Sections 2 and 3.

Notations Section 2	Adjustments or additional notations Section 3
$\boldsymbol{\theta} = (\boldsymbol{\beta}^\mathsf{T}, \boldsymbol{\gamma}^\mathsf{T}, \sigma)^\mathsf{T}$	$\boldsymbol{\alpha} = (\mathbf{c}^\mathsf{T}, \mathbf{d}^\mathsf{T}, \sigma)^\mathsf{T}$ $Z_{ij} = \hat{\beta}_j^{\text{init}} X_{ij}$ $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^\mathsf{T}$ where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\mathsf{T}$ $Q_{ij} = \hat{\gamma}_j^{\text{init}} X_{ij}$ $\mathbf{Q} = (\mathbf{Q}_1, \ldots, \mathbf{Q}_n)^\mathsf{T}$ where $\mathbf{Q}_i = (Q_{i1}, \ldots, Q_{ip})^\mathsf{T}$
	replace $\mathbf{X}_i^\mathsf{T}\boldsymbol{\beta}$ by $\hat{\beta}_0^{\text{init}} + \mathbf{Z}_i^\mathsf{T}\mathbf{c}$
$\mathbf{G}(\boldsymbol{\gamma}) = \text{diag}(G_i)$ with $G_i = 1/h(\mathbf{X}_{i-}^\mathsf{T}\boldsymbol{\gamma})$	replace $\mathbf{X}_{i-}^\mathsf{T}\boldsymbol{\gamma}$ by $\mathbf{Q}_i^\mathsf{T}\mathbf{d}$ $\Longrightarrow \mathbf{G}(\mathbf{d})$
$\mathbf{F}(\boldsymbol{\gamma}) = \text{diag}(F_i)$ with $F_i = h'(\mathbf{X}_{i-}^\mathsf{T}\boldsymbol{\gamma})$	$\Longrightarrow \mathbf{F}(\mathbf{d})$
$\mathbf{R}(\boldsymbol{\beta}) = \text{diag}(r_i^2)$ with $r_i = Y_i - \mathbf{X}_i^\mathsf{T}\boldsymbol{\beta}$	replace $Y_i - \mathbf{X}_i^\mathsf{T}\boldsymbol{\beta}$ with $Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^\mathsf{T}\mathbf{c}$ $\Longrightarrow \mathbf{R}(\mathbf{c})$
$\mathbf{V}(\boldsymbol{\gamma}) = \text{diag}(V_i)$ with $V_i = h''(\mathbf{X}_{i-}^\mathsf{T}\boldsymbol{\gamma})$	$\Longrightarrow \mathbf{V}(\mathbf{d})$
$\mathbf{W}_1(\boldsymbol{\theta}) = \text{diag}(W_{1i})$ with $W_{1i} = w_1\left(\frac{Y_i - \mathbf{X}_i^\mathsf{T}\boldsymbol{\beta}}{\sigma\sqrt{h(\mathbf{X}_{i-}^\mathsf{T}\boldsymbol{\gamma})}}\right)$	$\Longrightarrow \mathbf{W}_1(\boldsymbol{\alpha})$
$w_1(u) = \frac{\psi(u)}{u}$	
$\mathbf{A}(\boldsymbol{\theta}) = \text{diag}(A_i)$ with $A_i = \rho\left(\frac{Y_i - \mathbf{X}_i^\mathsf{T}\boldsymbol{\beta}}{\sigma\sqrt{h(\mathbf{X}_{i-}^\mathsf{T}\boldsymbol{\gamma})}}\right) - b$	$\Longrightarrow \mathbf{A}(\boldsymbol{\alpha})$
$w_2(u) = \frac{\rho(u)}{u^2}$	
$\mathbf{W}(\mathbf{X}_-) = \text{diag}(W_i)$ with W_i as in (9)	
$\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (r_{1*}, \ldots, r_{n*})^\mathsf{T}$ with $r_{i*} = \frac{Y_i - \mathbf{X}_i^\mathsf{T}\boldsymbol{\beta}}{\sqrt{h(\mathbf{X}_{i-}^\mathsf{T}\boldsymbol{\gamma})}}$	$\Longrightarrow \mathbf{r}_*(\mathbf{c}, \mathbf{d})$
$\mathbf{W}_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \text{diag}(W_{1i})$ with $W_{1i} = w_1\left(\frac{Y_i - \mathbf{X}_i^\mathsf{T}\beta}{\hat{\sigma}(\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma}))\sqrt{(h(\mathbf{X}_{i-}^\mathsf{T}\boldsymbol{\gamma})}}\right)$	$\mathbf{W}_1(\mathbf{c}, \mathbf{d}) = \text{diag}(W_{1i})$ $W_{1i} = w_1\left(\frac{Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^\mathsf{T}\mathbf{c}}{\hat{\sigma}(\mathbf{r}_*(\mathbf{c}, \mathbf{d}))\sqrt{h(\mathbf{Q}_i^\mathsf{T}\mathbf{d})}}\right)$
$\omega(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{\hat{\sigma}(\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma}))}{\mathbf{r}_*^\mathsf{T}(\boldsymbol{\beta}, \boldsymbol{\gamma})\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1(\boldsymbol{\beta}, \boldsymbol{\gamma})\mathbf{G}(\boldsymbol{\gamma})\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma})}$	$\omega(\mathbf{c}, \mathbf{d}) = \frac{\hat{\sigma}(\mathbf{r}_*(\mathbf{c}, \mathbf{d}))}{\mathbf{r}_*^\mathsf{T}(\mathbf{c}, \mathbf{d})\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1(\mathbf{c}, \mathbf{d})\mathbf{G}(\mathbf{d})\mathbf{r}_*(\mathbf{c}, \mathbf{d})}$

1, $\mathbf{0}_p$ is a column vector of dimension p with all elements 0, and other notations are in Table 1. Table 1 summarizes notations of Sections 2 and 3, since although the various methods discussed are quite different, similar structures can be seen behind them, and this is emphasized by the notations.

Since all three equations (2)–(4) depend on $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}$, we have to solve these score-equations simultaneously using an iterative procedure. Therefore, let $\hat{\boldsymbol{\theta}}^{(k)} = (\hat{\boldsymbol{\beta}}^{(k)^\mathsf{T}}, \hat{\boldsymbol{\gamma}}^{(k)^\mathsf{T}}, \hat{\sigma}^{(k)})^\mathsf{T}$ be the current value of $\boldsymbol{\theta}$ in this iteration procedure. The values of $\boldsymbol{\beta}$ and σ in the next iteration step are easily derived from, respectively, (2) and (4), and are given by

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \left(\mathbf{X}^\mathsf{T}\mathbf{G}\left(\hat{\boldsymbol{\gamma}}^{(k)}\right)\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{G}\left(\hat{\boldsymbol{\gamma}}^{(k)}\right)\mathbf{Y}, \tag{5}$$

$$\hat{\sigma}^{(k+1)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{\left(Y_i - \mathbf{X}_i^\mathsf{T}\hat{\boldsymbol{\beta}}^{(k)}\right)^2}{h\left(\mathbf{X}_{i-}^\mathsf{T}\hat{\boldsymbol{\gamma}}^{(k)}\right)}}. \tag{6}$$

To obtain a further better approximation for the ML-estimator for $\boldsymbol{\gamma}$, one step of the Newton-Raphson procedure is used,

$$\hat{\boldsymbol{\gamma}}^{(k+1)} = \hat{\boldsymbol{\gamma}}^{(k)} - \mathbf{H}_{\boldsymbol{\gamma}}\left(\hat{\boldsymbol{\theta}}^{(k)}\right)^{-1} \mathbf{S}_{\boldsymbol{\gamma}}\left(\hat{\boldsymbol{\theta}}^{(k)}\right), \tag{7}$$

with $\mathbf{H}_{\boldsymbol{\gamma}}(\boldsymbol{\theta})$ the matrix of second order partial derivatives of the log-likelihood with respect to $\boldsymbol{\gamma}$. From (3) it is obtained that this matrix is given by

$$\begin{aligned}
\mathbf{H}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) &= -\frac{1}{2}\sum_{i=1}^{n} \frac{\left(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2}{\sigma^2 h^3\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)} \left(h'\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)\right)^2 \mathbf{X}_{i-}\mathbf{X}_{i-}^{\mathrm{T}} \\
&\quad + \frac{1}{2}\sum_{i=1}^{n}\left[\frac{\left(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2}{\sigma^2 h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)} - 1\right]\left[\frac{h''\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)}{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)} - \frac{\left(h'\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)\right)^2}{h^2\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)}\right]\mathbf{X}_{i-}\mathbf{X}_{i-}^{\mathrm{T}} \\
&= -\frac{1}{\sigma^2}\mathbf{X}_{-}^{\mathrm{T}}\mathbf{R}(\boldsymbol{\beta})\mathbf{G}^3(\boldsymbol{\gamma})\mathbf{F}^2(\boldsymbol{\gamma})\mathbf{X}_{-} + \frac{1}{2\sigma^2}\mathbf{X}_{-}^{\mathrm{T}}\mathbf{R}(\boldsymbol{\beta})\mathbf{G}^2(\boldsymbol{\gamma})\mathbf{V}(\boldsymbol{\gamma})\mathbf{X}_{-} \\
&\quad - \frac{1}{2}\mathbf{X}_{-}^{\mathrm{T}}\mathbf{G}(\boldsymbol{\gamma})\mathbf{V}(\boldsymbol{\gamma})\mathbf{X}_{-} + \frac{1}{2}\mathbf{X}_{-}^{\mathrm{T}}\mathbf{G}^2(\boldsymbol{\gamma})\mathbf{F}^2(\boldsymbol{\gamma})\mathbf{X}_{-}.
\end{aligned}$$

Putting all together, the algorithm to compute the ML-estimator for $\boldsymbol{\theta}$ reads as follows.

Algorithm 1:

1. Initialize $\hat{\boldsymbol{\gamma}}^{(0)} = \mathbf{0}_p$, compute $\hat{\boldsymbol{\beta}}^{(0)}$, using (5), and then $\hat{\sigma}^{(0)}$ using (6).
2. Repeat for $k = 0, 1, 2, \ldots$, until convergence:
   (a) Compute $\hat{\boldsymbol{\gamma}}^{(k+1)}$ by performing one step of the Newton-Raphson procedure (7).
   (b) Compute $\hat{\boldsymbol{\beta}}^{(k+1)}$ with (5).
   (c) Compute $\hat{\sigma}^{(k+1)}$ with (6).

It is well known that the maximum likelihood method is not robust to outliers. Trimming (see the next section) is one way towards robustifying a maximum likelihood method.

2.2. Robust estimation in heteroscedastic linear models

2.2.1. Maximum trimmed likelihood estimation

The maximum trimmed likelihood (MTL) estimator was introduced by Hadi and Luceño [15] and Vandev and Neykov [16] and is based on finding, for given q, the subset of $q \leq n$ observations whose maximum likelihood fit has the largest value for the log-likelihood function:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MTL}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{q} \ell(\boldsymbol{\theta})_{i:n},$$

where $\ell(\boldsymbol{\theta})_{1:n} \geq \cdots \geq \ell(\boldsymbol{\theta})_{n:n}$ are the order statistics of $\ell_1(\boldsymbol{\theta}), \ldots, \ell_n(\boldsymbol{\theta})$.

The maximum trimmed likelihood estimator is based on ideas similar to the least trimmed squares estimator. See, for example, Maronna et al. [9]. Cheng [17] proposed an algorithm to obtain the maximum trimmed likelihood estimates, but simulation studies in that paper as well as in Section 4 show that the estimates for $\boldsymbol{\gamma}$ and σ have a large bias.

2.2.2. Heteroscedastic M-estimation

One approach towards robust estimation is to replace the squared loss function $\rho(u) = u^2/2$ in the first order conditions (2)–(4) with a convex slowly increasing loss function which is symmetric with a unique minimum at zero. See for example Carroll and Ruppert [2].

A frequently used function is for example a loss function of Tukey's biweight family,

$$\rho_a(x) = \begin{cases} \frac{a^2}{6}\left(1 - \left(1 - \left(\frac{x}{a}\right)^2\right)^3\right) & \text{if } |x| \le a, \\ \frac{a^2}{6} & \text{if } |x| > a, \end{cases} \tag{8}$$

where the constant a can be tuned for estimation efficiency. Also denote, as in equations (2)–(4), $\psi(u) = \rho'(u)$ and $b = \mathrm{E}(\rho(Z))$ with Z standard normal distributed.

Bianco et al. [3] and Bianco and Boente [4] introduced a one-step version of Weighted generalized M-estimator (short GM-estimator). This method starts from initial high-breakdown point estimators for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and σ and it improves the estimate of $\boldsymbol{\beta}$ by performing one step of the Newton-Raphson procedure. They establish the breakdown properties of the introduced robust estimators, and study via extensive simulations the performances of the one-step versions and related weighted GM-estimators. We refer to these papers for further details on heteroscedastic M-estimation and breakdown properties.

As already mentioned in the introduction it is important in M-regression to introduce a weighting scheme that protects against leverage points. This weighting scheme should be such that observations $\mathbf{X}_{i-}$ that are close to the centre of the data cloud in the covariate space get a weight close to one, whereas observations $\mathbf{X}_{i-}$ that are far away from the center (i.e., leverage points) should get a weight close to zero. We adopt here the weighting scheme provided in Serneels et al. [6]. Denoting

$$w_1(u) = \frac{\psi(u)}{u} = \frac{\rho'(u)}{u},$$

the weight for the observations $\mathbf{X}_{i-}$ is given by

$$W_i = w(\mathbf{X}_{i-}) = w_1\left(\frac{\left\|\mathbf{X}_{i-} - \mathrm{med}_{L_1}(\mathbf{X}_{1-}, \ldots, \mathbf{X}_{n-})\right\|_2}{\mathrm{median}_{1\le i\le n}\left\|\mathbf{X}_{i-} - \mathrm{med}_{L_1}(\mathbf{X}_{1-}, \ldots, \mathbf{X}_{n-})\right\|_2}\right), \tag{9}$$

where $\|\cdot\|_2$ stands for the Euclidean norm, $\mathrm{median}_{1\le i\le n}$ is the sample median of a vector of size n, and med_{L_1} is the L_1-median, a multivariate version of the sample median [18] defined by

$$\mathrm{med}_{L_1}(\mathbf{X}_{1-}, \ldots, \mathbf{X}_{n-}) = \underset{\mathbf{M}}{\mathrm{argmin}} \sum_{i=1}^{n} \|\mathbf{X}_{i-} - \mathbf{M}\|_2,$$

with $\mathbf{M}$ a column vector of size p. Note that, for a given observation $(\mathbf{X}_i^{\mathrm{T}}, Y_i)$ a multiplication of weights $w(\mathbf{X}_{i-})w_1((Y_i - \mathrm{median}(Y_1, \ldots, Y_n))/\mathrm{MAD}(Y_1, \ldots, Y_n))$, with MAD the Median Absolute Deviation of the response observations, i.e., $\mathrm{MAD}(Y_1, \ldots, Y_n) = \mathrm{median}_{1\le i\le n}(|Y_i - \mathrm{median}(Y_1, \ldots, Y_n)|)$, a robust estimator of the standard deviation in the response observations, protects against vertical outliers as well as leverage points.

Recalling (2)–(4) and taking into account the above considerations, the first order conditions for the heteroscedastic M-estimators (he-M-estimator) for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and σ are respectively given by:

$$\begin{aligned}
\mathbf{S}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}) &= \sum_{i=1}^{n} W_i \psi\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}}\right)\frac{\mathbf{X}_i}{\hat{\sigma}\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}} \\
&= \frac{1}{\hat{\sigma}^2}\mathbf{X}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1(\hat{\boldsymbol{\theta}})\mathbf{G}(\hat{\boldsymbol{\gamma}})(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}_{p+1}, \qquad (10) \\
\mathbf{S}_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\theta}}) &= \sum_{i=1}^{n} W_i\left[\rho\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}}\right) - b\right]\frac{h'\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}\mathbf{X}_{i-} \\
&= \mathbf{X}_-^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{A}(\hat{\boldsymbol{\theta}})\mathbf{G}(\hat{\boldsymbol{\gamma}})\mathbf{F}(\hat{\boldsymbol{\gamma}})\mathbf{1}_n = \mathbf{0}_p, \qquad (11) \\
\mathbf{S}_{\sigma}(\hat{\boldsymbol{\theta}}) &= \frac{2}{\hat{\sigma}}\sum_{i=1}^{n} W_i\left[\rho\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}}\right) - b\right] \\
&= \frac{2}{\hat{\sigma}}\sum_{i=1}^{n} W_i\left[w_2\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}}\right)\frac{\left(Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\right)^2}{\hat{\sigma}^2 h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)} - b\right] = 0, \qquad (12)
\end{aligned}$$

with notations as in Table 1.

The solution $\hat{\boldsymbol{\theta}}$ of the first order conditions (10)–(12) is computed again via an iterative procedure. Denote the current value of $\boldsymbol{\theta}$ in the estimation procedure with $\hat{\boldsymbol{\theta}}^{(k)} = (\hat{\boldsymbol{\beta}}^{(k)^{\mathrm{T}}}, \hat{\boldsymbol{\gamma}}^{(k)^{\mathrm{T}}}, \hat{\sigma}^{(k)^{\mathrm{T}}})^{\mathrm{T}}$. The values of $\boldsymbol{\beta}$ and σ in the next iteration step are then

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1\left(\hat{\boldsymbol{\theta}}^{(k)}\right)\mathbf{G}\left(\hat{\boldsymbol{\gamma}}^{(k)}\right)\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1\left(\hat{\boldsymbol{\theta}}^{(k)}\right)\mathbf{G}\left(\hat{\boldsymbol{\gamma}}^{(k)}\right)\mathbf{Y}, \quad (13)$$

$$\hat{\sigma}^{(k+1)} = \sqrt{\frac{\hat{\sigma}^{(k)2}}{b\sum_{i=1}^{n} W_i}\sum_{i=1}^{n} W_i\rho\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(k)}}{\hat{\sigma}^{(k)}\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}^{(k)}\right)}}\right)}, \quad (14)$$

and one step of the Newton-Raphson procedure leads to a better approximation for the solution of (11),

$$\hat{\boldsymbol{\gamma}}^{(k+1)} = \hat{\boldsymbol{\gamma}}^{(k)} - \mathbf{H}_{\boldsymbol{\gamma}}\left(\hat{\boldsymbol{\theta}}^{(k)}\right)^{-1}\mathbf{S}_{\boldsymbol{\gamma}}\left(\hat{\boldsymbol{\theta}}^{(k)}\right), \quad (15)$$

where now

$$\begin{aligned}
\mathbf{H}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) = &-\frac{1}{2}\sum_{i=1}^{n} W_i\psi\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma\sqrt{h(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma})}}\right)\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma\sqrt{h(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma})}}\frac{\left(h'(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma})\right)^2}{h^2(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma})}\mathbf{X}_{i-}\mathbf{X}_{i-}^{\mathrm{T}} \\
&+\sum_{i=1}^{n} W_i\left[\rho\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma\sqrt{h(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma})}}\right) - b\right]\left(\frac{h''(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma})}{h(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma})} - \frac{\left(h'(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma})\right)^2}{h^2(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma})}\right)\mathbf{X}_{i-}\mathbf{X}_{i-}^{\mathrm{T}}
\end{aligned}$$

$$= -\frac{1}{2\sigma^2}\mathbf{X}_-^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1(\boldsymbol{\theta})\mathbf{R}(\boldsymbol{\beta})\mathbf{G}^3(\boldsymbol{\gamma})\mathbf{F}^2(\boldsymbol{\gamma})\mathbf{X}_- + \mathbf{X}_-^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{A}(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\gamma})\mathbf{V}(\boldsymbol{\gamma})\mathbf{X}_- - \mathbf{X}_-^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{A}(\boldsymbol{\theta})\mathbf{G}^2(\boldsymbol{\gamma})\mathbf{F}^2(\boldsymbol{\gamma})\mathbf{X}_-.$$

Note that we use the same notations for the first order conditions for the he-M-estimator and for the Hessian matrix for $\boldsymbol{\gamma}$, as for the ML-estimator of Section 2.1. Although the use of the same notations, for different quantities, might appear somewhat ambiguous, it also shows the similarities between the sets of equations across the methods.

An algorithm to compute the heteroscedastic M-estimator for $\boldsymbol{\theta}$ is as follows.

Algorithm 2:

1. Compute initial weights $W_i^* = w(\mathbf{X}_{i-})w_1\left(\frac{Y_i - \text{median}(Y_1,\ldots,Y_n)}{\text{MAD}(Y_1,\ldots,Y_n)}\right)$.
2. Compute initial estimates $\hat{\boldsymbol{\beta}}^{(0)}$, $\hat{\boldsymbol{\gamma}}^{(0)}$ and $\hat{\sigma}^{(0)}$ by applying algorithm 1 on the weighted data matrices $\mathbf{X}^*$ and $\mathbf{Y}^*$, obtained by multiplying each observation $\mathbf{X}_i$ and Y_i with $\sqrt{W_i^*}$.
3. Repeat for $k = 0, 1, 2, \ldots$, until convergence:
   compute $\hat{\boldsymbol{\gamma}}^{(k+1)}$, $\hat{\boldsymbol{\beta}}^{(k+1)}$ and $\hat{\sigma}^{(k+1)}$ using respectively (15), (13) and (14).

In the estimation of $\boldsymbol{\theta}$, Tukey's biweight loss function (8) is used with two different values for a. For estimating $\boldsymbol{\beta}$, $a = 5.182$ is used to have high efficiency when the error terms come from a normal distribution. Since $\boldsymbol{\gamma}$ and σ are scale parameters, $a = 1.547$ is used to have robust scale estimates with high efficiency (see [9]).

2.2.3. Heteroscedastic S-estimation

Another approach for robust estimation in heteroscedastic linear regression models, is based on the idea of the S-estimator that is available for homoscedastic linear regression models (see [7]). This approach is here adopted to the setting of heteroscedasticity.

Recall a homoscedastic S-estimator: look for coefficients $\boldsymbol{\beta}$ that produce residuals that minimize a robust scale estimator of the residuals, such as the M-scale. More precisely, given residuals $r_i = Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$, $i = 1, \ldots, n$ (in a homoscedastic setting) let $\hat{\sigma}(\mathbf{r}(\boldsymbol{\beta}))$ be an M-scale that solves

$$\frac{1}{n}\sum_{i=1}^{n}\left[\rho\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\hat{\sigma}(\mathbf{r}(\boldsymbol{\beta}))}\right) - b\right] = 0.$$

A simple adaptation of this to the heteroscedastic linear model setting, would be to replace residuals $r_i = Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ by residuals $r_{i*} = (Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})/\sqrt{h(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma})}$. See Slock et al. [19]. Such an adaptation would however not be sufficient. Indeed the resulting estimator would not be robust to leverage points, since large values for the covariates result into small residuals $\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (r_{1*}, \ldots, r_{n*})^{\mathrm{T}}$ and thus in large weights for leverage points. Therefore, additional weights $W_i = w(\mathbf{X}_{i-})$, $i = 1, \ldots, n$, where w is the weight function (9) are used in the estimation procedure to control the effect of leverage points. See also Section 5.2.

In summary, given the residuals $\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma})$, then $\hat{\sigma}(\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma}))$ is an M-scale that solves

$$\sum_{i=1}^{n} W_i\left[\rho\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\hat{\sigma}(\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma}))\sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}\right)}}\right) - b\right] = 0, \tag{16}$$

where ρ is a real function such that

(1) ρ is symmetric, continuously differentiable and $\rho(0) = 0$;
(2) there exists $t > 0$ such that ρ is strictly increasing on $[0, a]$ and constant on $[a, +\infty)$ with $0 < \rho(a) = t < +\infty$;

and $b = \mathrm{E}(\rho(Z))$ with Z standard normal distributed. The loss function ρ that we will use, is Tukey's biweight loss function (8) with $a = 1.547$, which is also used for the homoscedastic S-estimator.

Denote also, in this section, the first-order derivative of ρ with $\psi(u) = \rho'(u)$. The heteroscedastic S-estimator (he-S-estimator) for $\boldsymbol{\beta}$ is then given by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \hat{\sigma}\left(\mathbf{r}_*(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}})\right),$$

where $\hat{\boldsymbol{\gamma}}$ is the heteroscedastic S-estimator for $\boldsymbol{\gamma}$ that solves

$$\mathbf{S}_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} W_i \left[\rho\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}}{\hat{\sigma}\left(\mathbf{r}_*(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\right) \sqrt{h\left(\mathbf{X}_{i-}^{\mathrm{T}} \hat{\boldsymbol{\gamma}}\right)}} \right) - b \right] \frac{h'\left(\mathbf{X}_{i-}^{\mathrm{T}} \hat{\boldsymbol{\gamma}}\right)}{h\left(\mathbf{X}_{i-}^{\mathrm{T}} \hat{\boldsymbol{\gamma}}\right)} \mathbf{X}_{i-} = \mathbf{0}_p. \tag{17}$$

Note that (17) is the first order condition (11) of $\boldsymbol{\gamma}$ where $\hat{\sigma}$ is replaced with $\hat{\sigma}\left(\mathbf{r}_*(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\right)$. This estimation equation is used to compute $\hat{\boldsymbol{\gamma}}$, since $\boldsymbol{\gamma}$ is, similar to σ, a vector of scale parameters and in the estimation procedure of the homoscedastic S-estimator also the M-estimator for scale is used to compute the scale parameter σ.

Analogous to the computation of the homoscedastic S-estimator, the first order derivative of equation (16) with respect to $\boldsymbol{\beta}$ is taken to compute $\partial \hat{\sigma}\left(\mathbf{r}_*(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}})\right) / \partial \boldsymbol{\beta}$. For $\hat{\sigma}\left(\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma})\right) \neq 0$, one can find (for notations see Table 1)

$$\left. \frac{\partial \hat{\sigma}\left(\mathbf{r}_*(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}})\right)}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = -\omega(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \mathbf{X}^{\mathrm{T}} \mathbf{W}(\mathbf{X}_-) \mathbf{W}_1(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \mathbf{G}(\hat{\boldsymbol{\gamma}}) \left(\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}} \right) = \mathbf{0}_{p+1}. \tag{18}$$

The solutions $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ to (18) and (17) have to be computed simultaneously with an iterative procedure. With $\hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\boldsymbol{\gamma}}^{(k)}$ the current values of respectively $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the value of $\boldsymbol{\beta}$ in the next iteration step is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(k+1)} = & \left(\mathbf{X}^{\mathrm{T}} \mathbf{W}(\mathbf{X}_-) \mathbf{W}_1\left(\hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)} \right) \mathbf{G}\left(\hat{\boldsymbol{\gamma}}^{(k)} \right) \mathbf{X} \right)^{-1} \\ & \mathbf{X}^{\mathrm{T}} \mathbf{W}(\mathbf{X}_-) \mathbf{W}_1\left(\hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)} \right) \mathbf{G}\left(\hat{\boldsymbol{\gamma}}^{(k)} \right) \mathbf{Y}, \end{aligned} \tag{19}$$

and one step of the Newton-Raphson procedure is used to obtain a better approximation for $\boldsymbol{\gamma}$. The derivative of $\mathbf{S}_{\boldsymbol{\gamma}}$ with respect to $\boldsymbol{\gamma}$ is

$$\begin{aligned} \mathbf{H}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = & -\frac{1}{\hat{\sigma}^3(\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma}))} \frac{\partial \hat{\sigma}\left(\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma})\right)}{\partial \boldsymbol{\gamma}} \left(\mathbf{X}_-^{\mathrm{T}} \mathbf{W}(\mathbf{X}_-) \mathbf{W}_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) \mathbf{R}(\boldsymbol{\beta}) \mathbf{G}^2(\boldsymbol{\gamma}) \mathbf{F}(\boldsymbol{\gamma}) \mathbf{1}_n \right)^{\mathrm{T}} \\ & - \frac{1}{2\hat{\sigma}^2(\mathbf{r}_*(\boldsymbol{\beta}, \boldsymbol{\gamma}))} \mathbf{X}_-^{\mathrm{T}} \mathbf{W}(\mathbf{X}_-) \mathbf{W}_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) \mathbf{R}(\boldsymbol{\beta}) \mathbf{G}^3(\boldsymbol{\gamma}) \mathbf{F}^2(\boldsymbol{\gamma}) \mathbf{X}_- \\ & + \mathbf{X}_-^{\mathrm{T}} \mathbf{W}(\mathbf{X}_-) \mathbf{A}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \mathbf{G}(\boldsymbol{\gamma}) \mathbf{V}(\boldsymbol{\gamma}) \mathbf{X}_- - \mathbf{X}_-^{\mathrm{T}} \mathbf{W}(\mathbf{X}_-) \mathbf{A}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \mathbf{G}^2(\boldsymbol{\gamma}) \mathbf{F}^2(\boldsymbol{\gamma}) \mathbf{X}_-, \end{aligned}$$

with

$$\frac{\partial \hat{\sigma}(\mathbf{r}_*(\boldsymbol{\beta},\boldsymbol{\gamma}))}{\partial \boldsymbol{\gamma}} = -\frac{1}{2}\omega(\boldsymbol{\beta},\boldsymbol{\gamma})\mathbf{X}_-^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1(\boldsymbol{\beta},\boldsymbol{\gamma})\mathbf{R}(\boldsymbol{\beta})\mathbf{G}^2(\boldsymbol{\gamma})\mathbf{F}(\boldsymbol{\gamma})\mathbf{1}_n,$$

which can be computed by taking the first order derivative of (16) with respect to $\boldsymbol{\gamma}$, and hence,

$$\hat{\boldsymbol{\gamma}}^{(k+1)} = \hat{\boldsymbol{\gamma}}^{(k)} - \mathbf{H}_{\boldsymbol{\gamma}}\left(\hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)}\right)^{-1}\mathbf{S}_{\boldsymbol{\gamma}}\left(\hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)}\right). \tag{20}$$

For computing homoscedastic S-regression estimates fast computing algorithms have been developed. See Salibain-Barrera and Yohai [20]. Adaptation of a fast algorithm for homoscedastic S-regression estimates leads to an algorithm for computing the heteroscedastic S-regression estimates. See also Slock et al. [19]. This algorithm is based on improvements steps or I-steps. In an I-step one step of the iterative procedure for computing the S-regression estimates is performed. Hence, if $\hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\boldsymbol{\gamma}}^{(k)}$ are approximations of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, this I-step is defined as follows:

(1) Compute $\hat{\sigma}^{(k)} = \hat{\sigma}(\mathbf{r}_*(\hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)}))$ with (14).
(2) Compute $\hat{\boldsymbol{\gamma}}^{(k+1)}$ by performing one step of the Newton-Raphson procedure (20).
(3) Compute $\hat{\boldsymbol{\beta}}^{(k+1)}$ by performing one step of (19).

Since $\hat{\sigma}(\mathbf{r}_*(\boldsymbol{\beta},\boldsymbol{\gamma}))$ is in general non-convex, different starting points may converge to different critical points. A large number N of starting values is therefore taken. To find a starting point $\boldsymbol{\beta}^{(0)}$ for $\boldsymbol{\beta}$, draw a random subsample of size $p+1$ of the data set, denote the subsample by $(\mathbf{Y}_*, \mathbf{X}_*)$, and let $\boldsymbol{\beta}^{(0)}$ be the ordinary least-squares estimator on which κ I-steps of the homoscedastic S-estimator are applied. After repeating this for a large number of starting values, one selects as a final estimate the one that leads to the smallest estimated value of the M-scale (among the candidate estimates). This leads to the following algorithm to compute the heteroscedastic S-regression estimators.

Algorithm 3:
(1) Initialize $\boldsymbol{\gamma}^{(0)} = \mathbf{0}_p$ and let $\boldsymbol{\beta}_1^{(0)}, \ldots, \boldsymbol{\beta}_N^{(0)}$ be initial candidates. For each $(\boldsymbol{\beta}_\ell^{(0)}, \boldsymbol{\gamma}^{(0)})$, $\ell = 1, \ldots, N$,
  (a) Carry out κ I-steps and denote the improved candidate with $(\boldsymbol{\beta}_\ell^{(1)}, \boldsymbol{\gamma}_\ell^{(1)})$.
  (b) Compute the M-scale $\hat{\sigma}_\ell = \hat{\sigma}(\mathbf{r}_*(\boldsymbol{\beta}_\ell^{(1)}, \boldsymbol{\gamma}_\ell^{(1)}))$.
(2) Keep the t improved candidates with the lowest values for $\hat{\sigma}(\mathbf{r}_*(\boldsymbol{\beta},\boldsymbol{\gamma}))$.
(3) For each $(\boldsymbol{\beta}_\ell^{(1)}, \boldsymbol{\gamma}_\ell^{(1)})$, $\ell = 1, \ldots, t$, carry out I-steps until convergence and denote the final candidate with $(\boldsymbol{\beta}_\ell^{F}, \boldsymbol{\gamma}_\ell^{F})$.
(4) The heteroscedastic S-regression estimate of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is the candidate $(\boldsymbol{\beta}_\ell^{F}, \boldsymbol{\gamma}_\ell^{F})$ with the lowest value for $\hat{\sigma}(\mathbf{r}_*(\boldsymbol{\beta},\boldsymbol{\gamma}))$.

More details on this fast implementation approach in the homoscedastic setting are provided in Gijbels and Vrinssen [21].

3. Robust variable selection in linear regression models

The methods described in Section 2 do not perform a variable selection task, but they can be used as initial estimators for the heteroscedastic nonnegative garrote method that will

perform simultaneously the variable selection method. We first recall some essential ingredients of the robust homoscedastic nonnegative garrote method that was established in Gijbels and Vrinssen [21] and further studied in Gijbels et al. [14], i.e., when considering (1) with $\sigma_i = \sigma$ for all $i = 1, \ldots, n$.

3.1. Robust variable selection in homoscedastic linear regression models

3.1.1. Non-robust nonnegative garrote method

Under a homoscedatic linear regression model, the ordinary least-squares estimator $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (\hat{\beta}_0^{\mathrm{OLS}}, \ldots, \hat{\beta}_p^{\mathrm{OLS}})^{\mathrm{T}}$ solves the optimization problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2, \tag{21}$$

leading to, if the inverse of the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ exists, the ordinary least-squares estimator

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

One approach for variable selection is to add a penalty that measures the size of the coefficients $\boldsymbol{\beta}$, to the goodness-of-fit term in (21). The nonnegative garrote (NNG) method [22] uses a penalty on shrinkage factors of the regression coefficients. Starting from the ordinary least-squares estimator, the method shrinks or puts some coefficients of $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ equal to zero (i.e., eliminates some covariate components) using the nonnegative garrote shrinkage factors $\mathbf{c} = (c_1, \ldots, c_p)^{\mathrm{T}}$. Since such a shrinkage should not alter the sign of a covariate's influence in the linear model, and should also be globally a real shrinkage (i.e., $\sum_{j=1}^{p} c_j < p$) of the original regression coefficients, the NNG shrinkage factors $\widehat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_p)^{\mathrm{T}}$ are found by solving

$$\min_{\mathbf{c}} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0^{\mathrm{OLS}} - \sum_{j=1}^{p} c_j \hat{\beta}_j^{\mathrm{OLS}} X_{ij} \right)^2$$
$$\text{subject to } c_j \geq 0 \text{ for } j = 1, \ldots, p, \text{ and } \sum_{j=1}^{p} c_j \leq s,$$

for given s (with $s \leq p$), which is also equivalent to the optimization problem

$$\min_{\mathbf{c}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0^{\mathrm{OLS}} - \sum_{j=1}^{p} c_j \hat{\beta}_j^{\mathrm{OLS}} X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} c_j \right\}$$
$$\text{subject to } c_j \geq 0 \text{ for } j = 1, \ldots, p, \tag{22}$$

for given $\lambda > 0$.

The NNG estimator of the regression coefficient β_j, for $j = 1, \ldots, p$, is then given by

$$\hat{\beta}_j^{\mathrm{NNG}} = \hat{c}_j \hat{\beta}_j^{\mathrm{OLS}}, \quad j = 1, \ldots, p,$$

which, since $\sum_{j=1}^{p} c_j < p$ clearly reveals the shrinking effect. If, for a given j, $\hat{c}_j = 0$, $\hat{\beta}_j^{\mathrm{NNG}} = 0$ and the associated covariate is not selected as being part of the linear model.

Obviously this estimation and selection method is not robust. In the next section, we briefly review how to get to robust nonnegative garrote methods.

3.1.2. Robust nonnegative garrote method

The original nonnegative garrote method (22) involves three essential parts: the initial estimation of $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$, the estimation of the nonnegative shrinkage factors $\mathbf{c} = (c_1, \ldots, c_p)^{\mathrm{T}}$, and the choice of a regularization parameter λ. Robustifying the original nonnegative garrote method can be done by (i) starting from an initial robust estimator for the vector $\boldsymbol{\beta}$; (ii) considering robust versions of the quadratic loss goodness-of-fit term in (22); (iii) using an appropriate robust method to select the regularization parameter λ. A detailed discussion on resulting robust nonnegative garrote methods for homoscedastic linear regression can be found in Gijbels and Vrinssen [21]. We here only provide some very brief discussion.

Denote the initial robust estimator of β_j by $\hat{\beta}_j^{\mathrm{R}}$, $j = 0, \ldots, p$. Then let $Z_{ij}^{\mathrm{R}} = \hat{\beta}_j^{\mathrm{R}} X_{ij}$ for $j = 1, \ldots, p$, $i = 1, \ldots, n$, such that $\mathbf{Z}^{\mathrm{R}} = (Z_{ij}^{\mathrm{R}})_{1 \le i \le n, 1 \le j \le p}$ and $\mathbf{Z}_i^{\mathrm{R}} = (Z_{i1}^{\mathrm{R}}, \ldots Z_{ip}^{\mathrm{R}})^{\mathrm{T}}$, $i = 1, \ldots, n$. Also denote the residuals of the robust nonnegative garrote shrinkage factors $\mathbf{c}$ with $\mathbf{r}_{\mathrm{R}}(\mathbf{c}) = (r_{\mathrm{R},1}, \ldots, r_{\mathrm{R},n})^{\mathrm{T}}$, where $r_{\mathrm{R},i} = Y_i - \hat{\beta}_0^{\mathrm{R}} - \sum_{j=1}^{p} c_j Z_{ij}^{\mathrm{R}}$, $i = 1, \ldots, n$. Robust nonnegative garrote shrinkage factors $\widehat{\mathbf{c}}$ are found by solving the optimization problem

$$\widehat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmin}} \left\{ \mathrm{GOF}(\mathbf{r}_{\mathrm{R}}(\mathbf{c})) + \lambda \sum_{j=1}^{p} c_j \right\}$$
$$\text{s.t. } c_j \ge 0 \ (j = 1, \ldots, p),$$

where $\mathrm{GOF}(\mathbf{r}_{\mathrm{R}}(\mathbf{c}))$ is a robust alternative to the mean squared residuals in (22). For example, in the robust M-nonnegative garrote method $\mathrm{GOF}(\mathbf{r}_{\mathrm{R}}(\mathbf{c})) = (1/n) \sum_{i=1}^{n} \rho((Y_i - \hat{\beta}_0^{\mathrm{R}} - \sum_{j=1}^{p} c_j Z_{ij}^{\mathrm{R}})/\hat{\sigma})$, where $\hat{\sigma}$ is a robust scale estimator of the residuals $\mathbf{r}(\hat{\boldsymbol{\beta}}^{\mathrm{R}})$ of the initial estimators of the coefficients $\boldsymbol{\beta}$, and ρ is a nonnegative, symmetric and slowly increasing function with a unique minimum at zero.

3.2. Variable selection in heteroscedastic linear regression models

We now return to the heteroscedastic linear regression model (1) where the unknown coefficients/parameters are collected into the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \sigma)^{\mathrm{T}}$. Since the covariates possibly come into play in the mean part, but also in the error variance part, we now will have two vectors of shrinkage factors, denoted by $\mathbf{c} = (c_1, \ldots, c_p)^{\mathrm{T}}$ and $\mathbf{d} = (d_1, \ldots, d_p)^{\mathrm{T}}$. Consequently we will also have two penalty terms, as well as two regularization parameters to be chosen. With also an unknown variance factor σ^2 to be estimated, the (robust) heteroscedastic nonnegative garrote method starts from an initial estimator for $\boldsymbol{\theta}$, and then

it shrinks or puts some coefficients of this initial estimator equal to zero using the heteroscedastic nonnegative garrote shrinkage factors $\hat{\boldsymbol{\alpha}} = (\widehat{\mathbf{c}}^{\mathrm{T}}, \widehat{\mathbf{d}}^{\mathrm{T}}, \hat{\sigma})^{\mathrm{T}}$ with $\widehat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_p)^{\mathrm{T}}$ and $\widehat{\mathbf{d}} = (\hat{d}_1, \ldots, \hat{d}_p)^{\mathrm{T}}$. Suppose that we have initial estimators $\hat{\beta}_j^{\text{init}}$ and $\hat{\gamma}_j^{\text{init}}$ of the coefficients β_j and γ_j. The heteroscedastic nonnegative garrote estimate for β_0 remains $\hat{\beta}_0^{\text{init}}$, the estimator for β_j, for $j = 1, \ldots, p$, is $\hat{c}_j \hat{\beta}_j^{\text{init}}$; the estimate for γ_j is $\hat{d}_j \hat{\gamma}_j^{\text{init}}$, for $j = 1, \ldots, p$, and the estimate for σ is given by $\hat{\sigma}$. Further, denote $\hat{\beta}_j^{\text{init}} X_{ij}$ with Z_{ij}, and $\hat{\gamma}_j^{\text{init}} X_{ij}$ with Q_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\mathrm{T}}$ where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^{\mathrm{T}}$ and $\mathbf{Q} = (\mathbf{Q}_1, \ldots, \mathbf{Q}_n)^{\mathrm{T}}$ where $\mathbf{Q}_i = (Q_{i1}, \ldots, Q_{ip})^{\mathrm{T}}$, for $i = 1, \ldots, n$. Hence, $\mathbf{Z}$ collects the transformed covariates that explain the mean response and $\mathbf{Q}$ collects the transformed covariates that explain the error variance.

Since maximum likelihood estimation is a commonly-used technique, we present first an extension of the nonnegative garrote method for the heteroscedastic linear model (1) in Section 3.2.1. Thereafter we discuss three robust heteroscedastic nonnegative garrote methods.

3.2.1. A (non-robust) heteroscedastic nonnegative garrote method

The heteroscedastic nonnegative garrote method combines the ideas behind the ML-estimator of Section 2.1 and the nonnegative garrote method. We constrain the ML-estimator: the heteroscedastic nonnegative garrote shrinkage factors have to globally shrink the initial estimators, but they are not allowed to change the sign of these initial estimators. Hence, if the conditional density funtion of ε_i given $\mathbf{X}_i$ is denoted with f, then the heteroscedastic nonnegative garrote (he-NNG) shrinkage factors $\hat{\boldsymbol{\alpha}}$ are found by solving

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\log(\sigma) + \frac{1}{2} \log\left(h\left(\mathbf{Q}_i^{\mathrm{T}} \mathbf{d} \right) \right) - \log\left(f\left(\frac{Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}} \mathbf{c}}{\sigma \sqrt{h\left(\mathbf{Q}_i^{\mathrm{T}} \mathbf{d} \right)}} \right) \right) \right] + \lambda_1 \sum_{j=1}^{p} c_j + \lambda_2 \sum_{j=1}^{p} d_j \right\}$$

$$\text{s.t. } c_j \geq 0 \ (j = 1, \ldots, p), \text{ and } d_j \geq 0 \ (j = 1, \ldots, p),$$

where $\boldsymbol{\alpha} = (\mathbf{c}^{\mathrm{T}}, \mathbf{d}^{\mathrm{T}}, \sigma)^{\mathrm{T}}$ with $\mathbf{c} = (c_1, \ldots, c_p)^{\mathrm{T}}$ and $\mathbf{d} = (d_1, \ldots, d_p)^{\mathrm{T}}$ and $\lambda_1 > 0$ and $\lambda_2 > 0$ are given regularization parameters. If the error terms are standard normally distributed, this optimization problem can be written as

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \, \ell(\boldsymbol{\alpha})$$

$$\text{s.t. } c_j \geq 0 \ (j = 1, \ldots, p), \text{ and } d_j \geq 0 \ (j = 1, \ldots, p), \tag{23}$$

where

$$\ell(\boldsymbol{\alpha}) = \frac{1}{2} \log(2\pi) + \log(\sigma) + \frac{1}{2n} \sum_{i=1}^{n} \log\left(h\left(\mathbf{Q}_i^{\mathrm{T}} \mathbf{d} \right) \right) + \frac{1}{2n} \sum_{i=1}^{n} \frac{\left(Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}} \mathbf{c} \right)^2}{\sigma^2 h\left(\mathbf{Q}_i^{\mathrm{T}} \mathbf{d} \right)} + \lambda_1 \sum_{j=1}^{p} c_j + \lambda_2 \sum_{j=1}^{p} d_j.$$

The first order conditions for the heteroscedastic nonnegative garrote shrinkage factors are

$$\mathbf{S_c}(\hat{\boldsymbol{\alpha}}) = \left.\frac{\partial \ell(\boldsymbol{\alpha})}{\partial \mathbf{c}}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}} = -\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{c}}}{\hat{\sigma}\sqrt{h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}}\right) \frac{\mathbf{Z}_i}{\hat{\sigma}\sqrt{h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}} + \lambda_1 \mathbf{1}_p$$

$$= -\frac{1}{n\hat{\sigma}^2}\mathbf{Z}^{\mathrm{T}}\mathbf{G}(\widehat{\mathbf{d}})(\widetilde{\mathbf{Y}} - \mathbf{Z}\widehat{\mathbf{c}}) + \lambda_1 \mathbf{1}_p = \mathbf{0}_p, \tag{24}$$

$$\mathbf{S_d}(\hat{\boldsymbol{\alpha}}) = \left.\frac{\partial \ell(\boldsymbol{\alpha})}{\partial \mathbf{d}}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}} = -\frac{1}{n}\sum_{i=1}^{n}\left[\rho\left(\frac{Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{c}}}{\hat{\sigma}\sqrt{h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}}\right) - b\right] \frac{h'\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}{h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}\mathbf{Q}_i + \lambda_2 \mathbf{1}_p$$

$$= -\frac{1}{2n\hat{\sigma}^2}\mathbf{Q}^{\mathrm{T}}\mathbf{R}(\widehat{\mathbf{c}})\mathbf{G}^2(\widehat{\mathbf{d}})\mathbf{F}(\widehat{\mathbf{d}})\mathbf{1}_n + \frac{1}{2n}\mathbf{Q}^{\mathrm{T}}\mathbf{G}(\widehat{\mathbf{d}})\mathbf{F}(\widehat{\mathbf{d}})\mathbf{1}_n + \lambda_2 \mathbf{1}_p = \mathbf{0}_p, \tag{25}$$

$$\mathbf{S}_\sigma(\hat{\boldsymbol{\alpha}}) = \left.\frac{\partial \ell(\boldsymbol{\alpha})}{\partial \sigma}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}} = \frac{2}{n\hat{\sigma}}\sum_{i=1}^{n}\left[\rho\left(\frac{Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{c}}}{\hat{\sigma}\sqrt{h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}}\right) - b\right] = 0, \tag{26}$$

where $\rho(u) = u^2/2$, $\psi(u) = \rho'(u) = u$ and $b = \mathrm{E}(\rho(Z))$ with Z standard normal distributed, and we denoted $\widetilde{\mathbf{Y}} = \mathbf{Y} - \hat{\beta}_0^{\text{init}}\mathbf{1}_n$, and with further notations as in Table 1.

An iterative procedure is needed. Suppose that $\hat{\boldsymbol{\alpha}}^{(k)} = (\widehat{\mathbf{c}}^{(k)^{\mathrm{T}}}, \widehat{\mathbf{d}}^{(k)^{\mathrm{T}}}, \hat{\sigma}^{(k)})^{\mathrm{T}}$ is the current value for $\boldsymbol{\alpha}$ in this iterative procedure, the value for σ in the next iteration step is given by

$$\hat{\sigma}^{(k+1)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{\left(Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{c}}^{(k)}\right)^2}{h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}^{(k)}\right)}}, \tag{27}$$

and the value for $\mathbf{c}$ is

$$\widehat{\mathbf{c}}^{(k+1)} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}\right)^{-1}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\widetilde{\mathbf{Y}} - n\hat{\sigma}^{(k)^2}\lambda_1\mathbf{1}_p\right). \tag{28}$$

Note that (28) is a critical point of the quadratic function

$$\frac{1}{2}\mathbf{c}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}\mathbf{c} - \left(\mathbf{Z}^{\mathrm{T}}\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\widetilde{\mathbf{Y}} - n\hat{\sigma}^{(k)^2}\lambda_1\mathbf{1}_p\right)^{\mathrm{T}}\mathbf{c}.$$

Hence, if the constraints of optimization problem (23) are also taken into account, the value for $\mathbf{c}$ in the next iteration step can be found by solving the problem

$$\widehat{\mathbf{c}}^{(k+1)} = \underset{\mathbf{c}}{\operatorname{argmin}}\left\{\frac{1}{2}\mathbf{c}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}\mathbf{c} - \left(\mathbf{Z}^{\mathrm{T}}\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)(\widetilde{\mathbf{Y}} - n\hat{\sigma}^{(k)^2}\lambda_1\mathbf{1}_p\right)^{\mathrm{T}}\mathbf{c}\right\}$$
$$\text{s.t. } c_j \geq 0 \ (j = 1, \ldots, p). \tag{29}$$

To obtain a better approximation for $\mathbf{d}$, one step of the Newton-Raphson procedure is used:

$$\widehat{\mathbf{d}}^{(k+1)} = \widehat{\mathbf{d}}^{(k)} - \mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)^{-1}\mathbf{S_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right) = \mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)^{-1}\left[\mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\widehat{\mathbf{d}}^{(k)} - \mathbf{S_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\right], \tag{30}$$

in which the derivative of $\mathbf{S_d}$ with respect to $\mathbf{d}$ is

$$\begin{aligned}\mathbf{H_d}(\boldsymbol{\alpha}) &= \frac{1}{2n}\sum_{i=1}^{n}\frac{\left(Y_i-\hat\beta_0^{\text{init}}-\mathbf{Z}_i^{\mathrm{T}}\mathbf{c}\right)^2}{\sigma^2 h^3\left(\mathbf{Q}_i^{\mathrm{T}}\mathbf{d}\right)}\left(h'\left(\mathbf{Q}_i^{\mathrm{T}}\mathbf{d}\right)\right)^2\mathbf{Q}_i\mathbf{Q}_i^{\mathrm{T}}\\ &\quad-\frac{1}{2n}\sum_{i=1}^{n}\left[\frac{\left(Y_i-\hat\beta_0^{\text{init}}-\mathbf{Z}_i^{\mathrm{T}}\mathbf{c}\right)^2}{\sigma^2 h\left(\mathbf{Q}_i^{\mathrm{T}}\mathbf{d}\right)}-1\right]\left[\frac{h''\left(\mathbf{Q}_i^{\mathrm{T}}\mathbf{d}\right)}{h\left(\mathbf{Q}_i^{\mathrm{T}}\mathbf{d}\right)}\mathbf{Q}_i\mathbf{Q}_i^{\mathrm{T}}-\frac{\left(h'\left(\mathbf{Q}_i^{\mathrm{T}}\mathbf{d}\right)\right)^2}{h^2\left(\mathbf{Q}_i^{\mathrm{T}}\mathbf{d}\right)}\mathbf{Q}_i\mathbf{Q}_i^{\mathrm{T}}\right].\\ &= \frac{1}{n\sigma^2}\mathbf{Q}^{\mathrm{T}}\mathbf{R}(\mathbf{c})\mathbf{G}^3(\mathbf{d})\mathbf{F}^2(\mathbf{d})\mathbf{Q}-\frac{1}{2n\sigma^2}\mathbf{Q}^{\mathrm{T}}\mathbf{R}(\mathbf{c})\mathbf{G}^2(\mathbf{d})\mathbf{V}(\mathbf{d})\mathbf{Q}\\ &\quad+\frac{1}{2n}\mathbf{Q}^{\mathrm{T}}\mathbf{G}(\mathbf{d})\mathbf{V}(\mathbf{d})\mathbf{Q}-\frac{1}{2n}\mathbf{Q}^{\mathrm{T}}\mathbf{G}^2(\mathbf{d})\mathbf{F}^2(\mathbf{d})\mathbf{Q}.\end{aligned}$$

Remark that (30) is a critical point of the quadratic function

$$\frac{1}{2}\mathbf{d}^{\mathrm{T}}\mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{d}-\left[\mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\widehat{\mathbf{d}}^{(k)}-\mathbf{S_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\right]^{\mathrm{T}}\mathbf{d}.$$

Hence, if the constraints of optimization problem (23) are also taken into account, a better approximation for $\mathbf{d}$ than $\widehat{\mathbf{d}}^{(k)}$ can be found by solving the quadratic programming problem

$$\begin{aligned}&\widehat{\mathbf{d}}^{(k+1)}=\underset{\mathbf{d}}{\operatorname{argmin}}\left\{\frac{1}{2}\mathbf{d}^{\mathrm{T}}\mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{d}-\left[\mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\widehat{\mathbf{d}}^{(k)}-\mathbf{S_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\right]^{\mathrm{T}}\mathbf{d}\right\}\\ &\text{s.t. } d_j\geq 0\ (j=1,\ldots,p).\end{aligned}\tag{31}$$

Finally, an algorithm for computing he-NNG estimates for $\mathbf{c}$ and $\mathbf{d}$ together, for given regularization parameters λ_1 and λ_2, reads as follows.

Algorithm 4:
(1) Initialize $\widehat{\mathbf{d}}^{(0)}=\mathbf{1}_p$ and compute $\widehat{\mathbf{c}}^{(0)}$ and $\hat\sigma^{(0)}$ using optimization problem (29) and equation (27) respectively.
(2) Repeat for $k=0,1,2,\ldots$, until convergence:
Compute $\widehat{\mathbf{d}}^{(k+1)}$, $\widehat{\mathbf{c}}^{(k+1)}$ and $\hat\sigma^{(k+1)}$ using respectively (31), (29) and (27).

The regularization parameters λ_1 and λ_2 have to be selected in a proper way. This can be done by minimizing, with respect to (λ_1,λ_2), a BIC or AIC criterion defined as

$$\begin{aligned}\text{BIC}(\lambda_1,\lambda_2)&=\sum_{i=1}^{n}\log\left(\hat\sigma^2 h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)\right)+\sum_{i=1}^{n}\frac{\left(Y_i-\hat\beta_0^{\text{init}}-\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{c}}\right)^2}{\hat\sigma^2 h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}+\log(n)\text{df},\\ \text{AIC}(\lambda_1,\lambda_2)&=\sum_{i=1}^{n}\log\left(\hat\sigma^2 h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)\right)+\sum_{i=1}^{n}\frac{\left(Y_i-\hat\beta_0^{\text{init}}-\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{c}}\right)^2}{\hat\sigma^2 h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}+2\text{df},\end{aligned}\tag{32}$$

where the degrees of freedom (df) are estimated by the number of nonzero he-NNG estimates of the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

The heteroscedastic nonnegative garrote method can be used when it is known that there are no outliers in the data. When there are outliers in the data, robust alternatives for the heteroscedastic nonnegative garrote method are needed. A first robust heteroscedastic nonnegative garrote method that we introduce, is the heteroscedastic trimmed nonnegative garrote method. This method is based on the MTL-estimator of Section 2.2.1.

3.2.2. The heteroscedastic trimmed nonnegative garrote method

Let H be the set of indices that belongs to the subset of size q that maximizes the objective function of the maximum trimmed likelihood method or let H be the set of indices that belongs to the observations with non-zero weights (for the residuals and the covariates) for the heteroscedastic M- or S-estimator. In this manner, we already have a subset of the data without outliers. The heteroscedastic trimmed nonnegative garrote (he-TNNG) shrinkage factors $\hat{\boldsymbol{\alpha}}$ are

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i \in H} \left[\frac{1}{2} \log(2\pi) + \log(\sigma) + \frac{1}{2} \log\left(h\left(\mathbf{Q}_i^{\mathrm{T}} \mathbf{d} \right) \right) + \frac{1}{2} \frac{\left(Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}} \mathbf{c} \right)^2}{\sigma^2 h\left(\mathbf{Q}_i^{\mathrm{T}} \mathbf{d} \right)} \right] + \lambda_1 \sum_{j=1}^{p} c_j + \lambda_2 \sum_{j=1}^{p} d_j \right\}$$

$$\text{s.t. } c_j \geq 0 \ (j = 1, \ldots, p), \text{ and } d_j \geq 0 \ (j = 1, \ldots, p), \tag{33}$$

for given regularization parameters $\lambda_1 > 0$ and $\lambda_2 > 0$. This vector $\hat{\boldsymbol{\alpha}}$ is found by solving the optimization problem (23) of the heteroscedastic nonnegative garrote method for observations $(\mathbf{X}_i^{\mathrm{T}}, Y_i)$, with $i \in H$. Thus Algorithm 4 can be applied here. We use the set of indices H obtained from the initial robust estimator to reduce the computation time, but note that this set of indices belongs to a subset of size q that is possibly not the optimal subset for minimizing the objective function in (33) and thus that there might exist a subset of size q that provides a smaller value for this objective function.

The regularization parameters λ_1 and λ_2 are selected by minimizing a robustified version of a BIC criterion, denoted by RBIC, and defined as (32) but instead of the sum over all $i \in \{1, \ldots, n\}$ one takes the sum over $i \in H$, and with now the degrees of freedom (df) estimated by the number of nonzero he-TNNG estimates of the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Alternatively one can use a robustified version of an AIC criterion, an expression analogously to the expression for RBIC, but with $\log(n)\text{df}$ replaced by 2df.

In the following section we propose a robust heteroscedastic nonnegative garrote method that is based on the he-M-estimator of Section 2.2.2.

3.2.3. The heteroscedastic M-nonnegative garrote method

Similar to the heteroscedastic M-estimator, one can replace in the first order conditions (24)–(26) the function $\rho(u) = u^2/2$ with a slowly increasing loss function which is symmetric with a unique minimum at zero, to robustify the heteroscedastic nonnegative garrote method. Also let $\psi(u) = \rho'(u)$ and $b = \mathrm{E}(\rho(Z))$, with Z standard normal distributed.

The hetorescedastic M-nonnegative garrote (he-MNNG) shrinkage factors $\hat{\boldsymbol{\alpha}}$ are nonnegative and their first order conditions are

$$\mathbf{S_c}(\hat{\boldsymbol{\alpha}}) = -\frac{1}{n\hat{\sigma}^2}\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1(\hat{\boldsymbol{\alpha}})\mathbf{G}(\widehat{\mathbf{d}})(\widetilde{\mathbf{Y}} - \mathbf{Z}\widehat{\mathbf{c}}) + \lambda_1\mathbf{1}_p = \mathbf{0}_p, \tag{34}$$

$$\mathbf{S_d}(\hat{\boldsymbol{\alpha}}) = -\frac{1}{n}\mathbf{Q}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{A}(\hat{\boldsymbol{\alpha}})\mathbf{G}(\widehat{\mathbf{d}})\mathbf{F}(\widehat{\mathbf{d}})\mathbf{1}_n + \lambda_2\mathbf{1}_p = \mathbf{0}_p, \tag{35}$$

$$\mathbf{S}_\sigma(\hat{\boldsymbol{\alpha}}) = \frac{2}{n\hat{\sigma}}\sum_{i=1}^{n} W_i\left[w_2\left(\frac{Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{c}}}{\hat{\sigma}\sqrt{h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}}\right)\frac{\left(Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{c}}\right)^2}{\hat{\sigma}^2 h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)} - b\right] = 0,$$

with notations as indicated in Table 1.

Note that these first order conditions are similar to these of the he-M-estimator (10)–(12), but that $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ is replaced with $\hat{\beta}_0^{\text{init}} + \mathbf{Z}_i^{\mathrm{T}}\mathbf{c}$ and $\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}$ with $\mathbf{Q}_i^{\mathrm{T}}\mathbf{d}$. In equations (34) and (35) there are also the extra terms $\lambda_1\mathbf{1}_p$ and $\lambda_2\mathbf{1}_p$ terms that come from the shrinkage operation.

An iterative procedure has to be used to compute the heteroscedastic M-nonnegative garrote shrinkage factors $\hat{\boldsymbol{\alpha}}$. With $\hat{\boldsymbol{\alpha}}^{(k)} = (\widehat{\mathbf{c}}^{(k)^{\mathrm{T}}}, \widehat{\mathbf{d}}^{(k)^{\mathrm{T}}}, \hat{\sigma}^{(k)})^{\mathrm{T}}$ the current value for $\boldsymbol{\alpha}$, the value of σ in the next iteration step is given by

$$\hat{\sigma}^{(k+1)} = \sqrt{\frac{\left(\hat{\sigma}^{(k)}\right)^2}{b\sum_{i=1}^{n} W_i}\sum_{i=1}^{n} W_i\rho\left(\frac{Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{c}}^{(k)}}{\hat{\sigma}^{(k)}\sqrt{h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}^{(k)}\right)}}\right)}, \tag{36}$$

and the value of $\mathbf{c}$ is given by

$$\begin{aligned}\widehat{\mathbf{c}}^{(k+1)} = &\left(\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}\right)^{-1}\\ &\times\left(\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\widetilde{\mathbf{Y}} - n\left(\hat{\sigma}^{(k)}\right)^2\lambda_1\mathbf{1}_p\right).\end{aligned} \tag{37}$$

Since (37) is a critical point of the quadratic function

$$\begin{aligned}&\frac{1}{2}\mathbf{c}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}\mathbf{c}\\ &-\left(\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\widetilde{\mathbf{Y}} - n\hat{\sigma}^{(k)^2}\lambda_1\mathbf{1}_p\right)^{\mathrm{T}}\mathbf{c},\end{aligned}$$

and, when the constraints of optimization problem (23) are also taken into account, the next value of $\mathbf{c}$ in the iteration procedure can be found by solving the optimization problem

$$\begin{aligned}\widehat{\mathbf{c}}^{(k+1)} = \underset{\mathbf{c}}{\text{argmin}}&\left\{\frac{1}{2}\mathbf{c}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}\mathbf{c}\right.\\ &\left.-\left(\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\widetilde{\mathbf{Y}} - n\hat{\sigma}^{(k)^2}\lambda_1\mathbf{1}_p\right)^{\mathrm{T}}\mathbf{c}\right\}\\ &\text{s.t. } c_j \geq 0\ (j = 1, \ldots, p).\end{aligned} \tag{38}$$

To obtain an approximation for $\mathbf{d}$, one step of the Newton-Raphson procedure is used. The first order partial derivative of $\mathbf{S_d}$ with respect to $\mathbf{d}$ is

$$\mathbf{H_d}(\boldsymbol{\alpha}) = \frac{1}{2n\sigma^2}\mathbf{Q}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1(\boldsymbol{\alpha})\mathbf{R}(\mathbf{c})\mathbf{G}^3(\mathbf{d})\mathbf{F}^2(\mathbf{d})\mathbf{Q} - \frac{1}{n}\mathbf{Q}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{A}(\boldsymbol{\alpha})\mathbf{G}(\mathbf{d})\mathbf{V}(\mathbf{d})\mathbf{Q} + \frac{1}{n}\mathbf{Q}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{A}(\boldsymbol{\alpha})\mathbf{G}^2(\mathbf{d})\mathbf{F}^2(\mathbf{d})\mathbf{Q}.$$

Hence, the iterative step for $\mathbf{d}$ becomes

$$\widehat{\mathbf{d}}^{(k+1)} = \widehat{\mathbf{d}}^{(k)} - \mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)^{-1}\mathbf{S_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right) = \mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)^{-1}\left[\mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\widehat{\mathbf{d}}^{(k)} - \mathbf{S_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\right].$$

Since this is a critical point of the quadratic function

$$\frac{1}{2}\mathbf{d}^{\mathrm{T}}\mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{d} - \left[\mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\widehat{\mathbf{d}}^{(k)} - \mathbf{S_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\right]^{\mathrm{T}}\mathbf{d},$$

and taking into account the constraints of optimization problem (23), a better approximation for $\mathbf{d}$ can be found by solving the quadratic programming problem

$$\widehat{\mathbf{d}}^{(k+1)} = \underset{\mathbf{d}}{\operatorname{argmin}}\left\{\frac{1}{2}\mathbf{d}^{\mathrm{T}}\mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{d} - \left[\mathbf{H_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\widehat{\mathbf{d}}^{(k)} - \mathbf{S_d}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\right]^{\mathrm{T}}\mathbf{d}\right\}$$
$$\text{s.t. } d_j \geq 0 \ (j = 1, \ldots, p). \tag{39}$$

The next algorithm computes the heteroscedastic M-nonnegative garrote shrinkage factors for $\boldsymbol{\alpha}$.

Algorithm 5:

1. Compute initial weights $W_i^* = w(\mathbf{X}_{i-})w_1\left(\frac{Y_i - \text{median}(Y_1, \ldots, Y_n)}{\text{MAD}(Y_1, \ldots, Y_n)}\right)$.
2. Compute initial estimates $\widehat{\mathbf{c}}^{(0)}$, $\widehat{\mathbf{d}}^{(0)}$ and $\hat{\sigma}^{(0)}$ by applying Algorithm 4 on the weighted data matrices $\mathbf{X}^*$ and $\mathbf{Y}^*$, obtained by multiplying each observation $\mathbf{X}_i$ and Y_i with $\sqrt{W_i^*}$.
3. Repeat for $k = 0, 1, 2, \ldots$, until convergence:
   Compute $\widehat{\mathbf{d}}^{(k+1)}$, $\widehat{\mathbf{c}}^{(k+1)}$ and $\hat{\sigma}^{(k+1)}$ from respectively (39), (38) and (36).

The regularization parameters λ_1 and λ_2 can be selected by minimizing a RBIC criterion

$$\text{RBIC}(\lambda_1, \lambda_2) = \sum_{i=1}^{n} w(\mathbf{X}_{i-}) \log\left(\hat{\sigma}^2 h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)\right) + 2\sum_{i=1}^{n} w(\mathbf{X}_{i-})\rho\left(\frac{Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{c}}}{\hat{\sigma}\sqrt{h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}}\right) + \log(n)\text{df},$$

where the degrees of freedom (df) are estimated by the number of nonzero he-MNNG estimates of the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. An alternative is to rely on a RAIC criterion, in which $\log(n)\text{df}$ would be replaced by 2df.

In the following section a method to estimate the heteroscedastic nonnegative garrote shrinkage factors based on the he-S-estimator is proposed.

3.2.4. The heteroscedastic S-nonnegative garrote method

The heteroscedastic S-nonnegative garrote method looks for the nonnegative coefficients $\mathbf{c}$ that produce residuals that minimize a penalized robust scale estimator of the residuals. Given the residuals $\mathbf{r}_*(\mathbf{c},\mathbf{d}) = (r_{1*},\ldots,r_{n*})^{\mathrm{T}}$ with $r_{i*} = (Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}}\mathbf{c})/\sqrt{h(\mathbf{Q}_i^{\mathrm{T}}\mathbf{d})}$, for $i = 1,\ldots,n$, then $\hat{\sigma}(\mathbf{r}_*(\mathbf{c},\mathbf{d}))$ is an M-scale that solves

$$\frac{1}{n}\sum_{i=1}^{n} W_i\left[\rho\left(\frac{Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}}\mathbf{c}}{\hat{\sigma}(\mathbf{r}_*(\mathbf{c},\mathbf{d}))\sqrt{h\left(\mathbf{Q}_i^{\mathrm{T}}\mathbf{d}\right)}}\right) - b\right] = 0, \tag{40}$$

where ρ and b are as in Section 2.2.3. The heteroscedastic S-nonnegative garrote (he-SNNG) shrinkage factor for $\mathbf{c}$ is thus given by

$$\begin{aligned}\widehat{\mathbf{c}} &= \underset{\mathbf{c}}{\operatorname{argmin}}\left\{\hat{\sigma}(\mathbf{r}_*(\mathbf{c},\widehat{\mathbf{d}})) + \lambda_1\sum_{j=1}^{p} c_j\right\},\\ &\text{s.t. } c_j \geq 0\ (j = 0,\ldots,p),\end{aligned} \tag{41}$$

where $\widehat{\mathbf{d}}$ is nonnegative and the positive coefficients of $\widehat{\mathbf{d}}$ solve

$$\mathbf{S}_{\mathbf{d}}(\widehat{\mathbf{c}},\widehat{\mathbf{d}}) = -\frac{1}{n}\sum_{i=1}^{n} W_i\left[\rho\left(\frac{Y_i - \hat{\beta}_0^{\text{init}} - \mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{d}}}{\hat{\sigma}(\mathbf{r}_*(\widehat{\mathbf{c}},\widehat{\mathbf{d}}))\sqrt{h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}}\right) - b\right]\frac{h'\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}{h\left(\mathbf{Q}_i^{\mathrm{T}}\widehat{\mathbf{d}}\right)}\mathbf{Q}_i + \lambda_2\mathbf{1}_p = \mathbf{0}_p. \tag{42}$$

Similar as for the heteroscedastic S-estimator, the weights $W_i = w(\mathbf{X}_{i-})$, $i = 1,\ldots,n$, where w is weight function (9) are added to the procedure to control the effect of leverage points.

Note that (40) and (42) are obtained by replacing in (16) and (17) $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ with $\hat{\beta}_0^{\text{init}} + \mathbf{Z}_i^{\mathrm{T}}\mathbf{c}$ and $\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}$ with $\mathbf{Q}_i^{\mathrm{T}}\mathbf{d}$, where $\mathbf{c}$ and $\mathbf{d}$ have to be nonnegative. Furthermore, the terms $\lambda_1\sum_{j=1}^{p} c_j$ and $\lambda_2\mathbf{1}_p$ are added to (41) and (42) to obtain shrinkage of the initial regression coefficients estimates. Therefore, the computation of the heteroscedastic S-nonnegative garrote shrinkage factors are similar to the he-S-estimates: in the computation of the heteroscedastic S-nonnegative garrote schrinkage factors $\widehat{\mathbf{c}}$, the first order partial derivative of $\hat{\sigma}(\mathbf{r}_*(\mathbf{c},\mathbf{d}))$ with respect to $\mathbf{c}$ is needed. One can find (with notations as in Table 1)

$$\mathbf{0}_p = \left.\frac{\partial\hat{\sigma}(\mathbf{r}_*(\mathbf{c},\widehat{\mathbf{d}}))}{\partial\mathbf{c}}\right|_{\mathbf{c}=\widehat{\mathbf{c}}} + \lambda_1\mathbf{1}_p = -\omega(\widehat{\mathbf{c}},\widehat{\mathbf{d}})\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1(\widehat{\mathbf{c}},\widehat{\mathbf{d}})\mathbf{G}(\widehat{\mathbf{d}})\left(\widetilde{\mathbf{Y}} - \mathbf{Z}\widehat{\mathbf{c}}\right) + \lambda_1\mathbf{1}_p.$$

The heteroscedastic S-nonnegative garrote (he-SNNG) shrinkage factors $\mathbf{c}$ and $\mathbf{d}$ are computed together using an iterative procedure. If $\widehat{\mathbf{c}}^{(k)}$ and $\widehat{\mathbf{d}}^{(k)}$ are the current values of $\mathbf{c}$ and $\mathbf{d}$ in the iteration procedure, then the value for $\mathbf{c}$ in the next iteration step is given by

$$\begin{aligned}\widehat{\mathbf{c}}^{(k+1)} = &\left(\omega\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}\right)^{-1}\\ &\cdot\left(\omega\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_-)\mathbf{W}_1\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\widetilde{\mathbf{Y}} - \lambda_1\mathbf{1}_p\right),\end{aligned}$$

which is a critical point of the quadratic function

$$\frac{\omega\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)}{2}\mathbf{c}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_{-})\mathbf{W}_1\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}\mathbf{c}$$
$$-\left(\omega\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_{-})\mathbf{W}_1\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\widetilde{\mathbf{Y}}-\lambda_1\mathbf{1}_p\right)^{\mathrm{T}}\mathbf{c}.$$

With the constraints of optimization problem (41) taken into account, the value of $\mathbf{c}$ in the next step of the iteration procedure can be found by solving the quadratic programming problem

$$\begin{aligned}\widehat{\mathbf{c}}^{(k+1)} = \underset{\mathbf{c}}{\operatorname{argmin}}\Bigg\{&\frac{\omega\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)}{2}\mathbf{c}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_{-})\mathbf{W}_1\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}\mathbf{c}\\ &-\left(\omega\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_{-})\mathbf{W}_1\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\mathbf{G}\left(\widehat{\mathbf{d}}^{(k)}\right)\widetilde{\mathbf{Y}}\right.\\ &\left.-\lambda_1\mathbf{1}_p\right)^{\mathrm{T}}\mathbf{c}\Bigg\},\\ \text{s.t. } c_j \geq 0\ (j=1,\ldots,p).&\end{aligned} \tag{43}$$

One step of the Newton-Raphson procedure is used, to obtain the value of $\mathbf{d}$ in the $(k+1)$th iteration step. The derivative of $\mathbf{S_d}$ with respect to $\mathbf{d}$ is

$$\begin{aligned}\mathbf{H_d}(\mathbf{c},\mathbf{d}) = &\frac{1}{n\hat{\sigma}^3(\mathbf{r}_*(\mathbf{c},\mathbf{d}))}\frac{\partial\hat{\sigma}(\mathbf{r}_*(\mathbf{c},\mathbf{d}))}{\partial\mathbf{d}}\left(\mathbf{Q}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_{-})\mathbf{W}_1(\mathbf{c},\mathbf{d})\mathbf{R}(\mathbf{c})\mathbf{G}^2(\mathbf{d})\mathbf{F}(\mathbf{d})\mathbf{1}_n\right)^{\mathrm{T}}\\ &+\frac{1}{2n\hat{\sigma}^2(\mathbf{r}_*(\mathbf{c},\mathbf{d}))}\mathbf{Q}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_{-})\mathbf{W}_1(\mathbf{c},\mathbf{d})\mathbf{R}(\mathbf{c})\mathbf{G}^3(\mathbf{d})\mathbf{F}^2(\mathbf{d})\mathbf{Q}\\ &-\frac{1}{n}\mathbf{Q}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_{-})\mathbf{A}(\mathbf{c},\mathbf{d})\mathbf{G}(\mathbf{d})\mathbf{V}(\mathbf{d})\mathbf{Q}+\frac{1}{n}\mathbf{Q}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_{-})\mathbf{A}(\mathbf{c},\mathbf{d})\mathbf{G}^2(\mathbf{d})\mathbf{F}^2(\mathbf{d})\mathbf{Q},\end{aligned}$$

with

$$\frac{\partial\hat{\sigma}(\mathbf{r}_*(\mathbf{c},\mathbf{d}))}{\partial\mathbf{d}} = -\frac{1}{2}\omega(\mathbf{c},\mathbf{d})\mathbf{Q}^{\mathrm{T}}\mathbf{W}(\mathbf{X}_{-})\mathbf{W}_1(\mathbf{c},\mathbf{d})\mathbf{R}(\mathbf{c})\mathbf{G}^2(\mathbf{d})\mathbf{F}(\mathbf{d})\mathbf{1}_n$$

which can be computed by taking the first order derivative of (40) with respect to $\mathbf{d}$. Note again the similarity between $\mathbf{H_d}(\mathbf{c},\mathbf{d})$ and $\mathbf{H}_{\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ of Section 2.2.3. Hence, the value of $\mathbf{d}$ becomes,

$$\begin{aligned}\widehat{\mathbf{d}}^{(k+1)} &= \widehat{\mathbf{d}}^{(k)} - \mathbf{H_d}\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)^{-1}\mathbf{S_d}\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\\ &= \mathbf{H_d}\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)^{-1}\left(\mathbf{H_d}\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\widehat{\mathbf{d}}^{(k)} - \mathbf{S_d}\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\right),\end{aligned}$$

which is a critical point of the function

$$\frac{1}{2}\mathbf{d}^{\mathrm{T}}\mathbf{H_d}\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\mathbf{d} - \left(\mathbf{H_d}\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\widehat{\mathbf{d}}^{(k)} - \mathbf{S_d}\left(\widehat{\mathbf{c}}^{(k)},\widehat{\mathbf{d}}^{(k)}\right)\right)^{\mathrm{T}}\mathbf{d}.$$

Thus, if we also include the constraint that the shrinkage factors have to be nonnegative, the value of $\mathbf{d}$ in the $(k+1)$th iteration step can be found by solving the programming

problem

$$\widehat{\mathbf{d}}^{(k+1)} = \underset{\mathbf{d}}{\operatorname{argmin}} \left\{ \frac{1}{2}\mathbf{d}^{\mathrm{T}}\mathbf{H}_{\mathbf{d}}\left(\widehat{\mathbf{c}}^{(k)}, \widehat{\mathbf{d}}^{(k)}\right)\mathbf{d} - \left(\mathbf{H}_{\mathbf{d}}\left(\widehat{\mathbf{c}}^{(k)}, \widehat{\mathbf{d}}^{(k)}\right)\widehat{\mathbf{d}}^{(k)} - \mathbf{S}_{\mathbf{d}}\left(\widehat{\mathbf{c}}^{(k)}, \widehat{\mathbf{d}}^{(k)}\right)\right)^{\mathrm{T}}\mathbf{d} \right\}$$
$$\text{s.t. } d_j \geq 0 \ (j = 1, \ldots, p). \tag{44}$$

The fast algorithm for heteroscedastic S-regression estimates of Section 2.2.3 is adapted to get an algorithm to compute the heteroscedastic S-nonnegative garrote estimates. In an I-step of the algorithm one step of the iterative procedure for computing the heteroscedastic S-nonnegative garrote shrinkage factors is implemented. Thus, if $\widehat{\mathbf{c}}^{(k)}$ and $\widehat{\mathbf{d}}^{(k)}$ are approximations of $\mathbf{c}$ and $\mathbf{d}$ respectively, this I-step is defined as follows:

(1) Compute $\hat{\sigma}^{(k)} = \hat{\sigma}(\mathbf{r}_*(\widehat{\mathbf{c}}^{(k)}, \widehat{\mathbf{d}}^{(k)}))$ with (36).
(2) Compute $\widehat{\mathbf{d}}^{(k+1)}$ by performing one step of the iterative procedure (44).
(3) Compute $\widehat{\mathbf{c}}^{(k+1)}$ by performing one step of the iterative procedure (43).

Different starting points may converge to different critical points, since $\hat{\sigma}(\mathbf{r}_*(\mathbf{c}, \mathbf{d}))$ is in general non-convex. Therefore, a large number N of starting values is needed. To find a starting point $\mathbf{c}^{(0)}$ for $\mathbf{c}$, draw a random subsample of size p of the data set, denote the subsample by $(\mathbf{Y}_*, \mathbf{Z}_*)$, and let $\mathbf{c}^{(0)}$ be solution of problem (29) on which κ I-steps of the homoscedastic S-nonnegative garrote method are applied.

The following algorithm computes the heteroscedastic S-regression estimates of $\mathbf{c}$ and $\mathbf{d}$:

Algorithm 6:
1. Initialize $\mathbf{d}^{(0)} = \mathbf{1}_p$ and let $\mathbf{c}_1^{(0)}, \ldots, \mathbf{c}_N^{(0)}$ be initial candidates. For each $(\mathbf{c}_\ell^{(0)}, \mathbf{d}^{(0)})$,
   (a) Carry out κ I-steps and denote the improved candidate with $(\mathbf{c}_\ell^{(1)}, \mathbf{d}_\ell^{(1)})$.
   (b) Compute the M-scale $\hat{\sigma}_\ell = \hat{\sigma}(\mathbf{r}_*(\mathbf{c}_\ell^{(1)}, \mathbf{d}_\ell^{(1)}))$.
2. Keep the t improved candidates with the lowest values for the objective function of problem (41).
3. For each $(\mathbf{c}_\ell^{(1)}, \mathbf{d}_\ell^{(1)})$, $\ell = 1, \ldots, t$, carry out I-steps until convergence and denote the final candidate with $(\mathbf{c}_\ell^F, \mathbf{d}_\ell^F)$.
4. The estimate of the heteroscedastic S-nonnegative garrote shrinkage factors is the candidate $(\mathbf{c}_\ell^F, \mathbf{d}_\ell^F)$ with the lowest value for the objective function of problem (41).

A possible way to choose the regularization parameters λ_1 and λ_2 is by minimizing a robust BIC criterion defined as

$$\mathrm{RBIC}(\lambda_1, \lambda_2) = \hat{\sigma}(\mathbf{r}_*(\widehat{\mathbf{c}}, \widehat{\mathbf{d}})) + \log(n)\mathrm{df},$$

where the degrees of freedom (df) are estimated by the number of nonzero he-SNNG estimates of the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. A RAIC criterion is obtained by replacing $log(n)$df with 2df.

The finite-sample performances of the different heteroscedastic nonnegative garrote methods are compared in the following section. We investigate the performances in the estimation *and* the variable selection tasks and how well the methods can handle different types of outliers.

4. Simulation study

With this simulation study we investigate the finite-sample performances of the various (heteroscedastisc) nonnegative garrote methods, indicated in Table 2 (see also the abbreviations). Some computer codes for the nonnegative garrote methods written by the authors are available at https://wis.kuleuven.be/stat/stat-inferen/codes. For the heteroscedastic S-nonnegative garrote method of Section 3.2.4 we took $N = 500$, $\kappa = 2$ and $t = 5$.

4.1. Simulation examples and settings

In a first example, presented in Section 4.2, we considered a linear regression model with 10 covariates of which three are relevant in explaining the response (but not all with equal impact/importance) and there is no intercept. In Section S.2 in the Supplemental Material, we present results on a second example, a latent factor model with twenty-five covariates, in which four variables are of equal importance in explaining the mean response. In both examples two (or no) covariates influence the error variance. See Table 3 for summary information.

Table 2. Methods included in the simulation study.

Method with abbreviation				
Method	Abbreviation	Described in	Heteroscedasticity	Robust
homoscedastic non-robust nonnegative garrote method	NNG	Section 3.1.1		
homoscedastic robust nonnegative garrote method	MM-NNG	Section 3.1.2 Gijbels & Vrinssen [21]		X
heteroscedastic nonnegative garrrote method	he-NNG	Section 3.2.1	X	X
heteroscedastic trimmed nonnegative garrote method	he-TNNG	Section 3.2.2	X	X
heteroscedastic M-nonnegative garrote method	he-MNNG	Section 3.2.3	X	X
heteroscedastic S-nonnegative garrote method	he-SNNG	Section 3.2.4	X	X

Table 3. Examples in the simulation study.

	# significant covariates for the			Coefficient vectors	
Ex.	Mean (/out of)	Variance (/out of)	Settings	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$
1	3/10	2/10	1	$\boldsymbol{\beta} = (0, 3, 2, 1.5, 0, 0, 0, 0, 0, 0)^T$	$\boldsymbol{\gamma}_1 = (0, 0, 0, 0, 0, 0, 0, 0, 1, 1)^T$
			2		$\boldsymbol{\gamma}_2 = (1, 1, 0, 0, 0, 0, 0, 0, 0, 0)^T$
			3		$\boldsymbol{\gamma}_3 = (1, 0, 0, 0, 0, 0, 0, 0, 0, 1)^T$
			4		$\boldsymbol{\gamma}_4 = \mathbf{0}_{10}$
2	4/25	2/25	1	$\boldsymbol{\beta} = (1, 1, 1, 1, 0, 0)^T$	$\boldsymbol{\gamma}_1 = (0, 0, 0, 0, 1, 1)^T$
				following 19 components of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are zero	
			2		$\boldsymbol{\gamma}_2 = (1, 1, 0, 0, 0, 0)^T$
			3		$\boldsymbol{\gamma}_3 = (1, 0, 0, 0, 0, 1)^T$
			4		$\boldsymbol{\gamma}_4 = \mathbf{0}_6$

In both simulation examples we consider four different situations/settings for $\boldsymbol{\gamma}$ (see Table 3): three heteroscedastic settings (Settings 1–3) and one homoscedastic setting (Setting 4):

- `Setting 1`: the variables that explain the error variance are different from those that explain the mean response;
- `Setting 2`: some variables that explain the mean response also explain the error variance;
- `Setting 3`: there is one covariate that impacts both the mean response and the error variance, and one extra covariate that impacts only the error variance;
- `Setting 4`: a homoscedastic setting with $\boldsymbol{\gamma} = \mathbf{0}_p$.

Different types of outliers are induced in both simulation designs according to three contamination schemes:

- `CScheme 1`. No contamination.
- `CScheme 2`. 10% of the error terms (in general $N(0,1)$ distributed) are normally distributed with mean 20 and standard deviation 1.
- `CScheme 3`. Same as in `CScheme 2`, but the 10% contaminated observations have high-leverage points generated from independent $N(50,1)$ distributions.

So, under `CScheme 1` there is no contamination, under `CScheme 2` there are only vertical outliers, whereas under `CScheme 3` there are vertical outliers as well as leverage points.

We generate 100 independent samples of size $n=200$ from the different simulation designs and compute, for each of the different nonnegative garrote methods, the evaluation criteria given in Table 4. The estimation errors for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and σ are respectively computed using the criteria

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}-\mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\right)^2, \quad \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}-\mathbf{X}_{i-}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}\right)^2, \quad \text{and} \quad (\sigma-\widehat{\sigma})^2. \tag{45}$$

Table 5 indicates which (natural) initial estimators we used for the various nonnegative garrote methods.

Table 4. Evaluation criteria.

ER	the estimation errors, defined in (45), averaged over all simulations, but calculated without the contaminated data
MTZ	median of the number of zero coefficients restricted to the true zero coefficients (true zero's)
MFZ	median of the number of zero coefficients restricted to the true non-zero coefficients (false zero's)
MTP	median of the number of non-zero coefficients restricted to the true non-zero coefficients (true positives)
MFP	median of the number of non-zero coefficients restricted to the true zero coefficients (false positives)

Table 5. Methods included in the simulation study, with their initial estimators.

Method	Initial estimators via
NNG	ordinary least-squares
MM-NNG	τ-estimator (see [8])
he-NNG	maximum likelihood method of Section 3.2.1
he-TNNG	maximum trimmed likelihood method of Section 2.2.1 (trimming: usage of 75% of the observations)
he-MNNG	heteroscedastic M-estimator of Section 2.2.2
he-SNNG	heteroscedastic S-estimator of Section 2.2.3

4.2. Simulation results for Example 1

We consider a simulation model with 10 covariates, given by

$$Y_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + 3\sqrt{\exp\left(\mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma}_k\right)}\varepsilon_i$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_k$, $k = 1,2,3,4$, are given in Table 3. The covariates $X_{i1}, \ldots, X_{i10}$ are generated from a multivariate normal distribution with mean 0 and $\mathrm{Corr}(X_{ij}, X_{ik}) = 0.5^{|j-k|}$ for $j, k = 1, \ldots, 10$. The error terms ϵ_i are standard normal distributed.

The results for the heteroscedastic `Setting 1` for the different methods when the discussed BIC-type criterion is used to select the regularization parameter(s), are shown in Figures 1–10. Results for the homoscedastic `Setting 4` are provided in Figures S.3–S.14 of the Supplemental Material. The results for heteroscedastic `Settings 2` and `3` are similar to these of `Setting 1`, and therefore are not included (but are available from the authors upon request).

We first discuss the results for heteroscedastic Setting 1, with respect to the quality of estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_1$. Figures 1 and 2 present the means and standard deviations of the estimation errors for the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively. Comparing the results for the he-MNNG and the he-NNG methods, we see that the estimation errors are comparable when the data are not contaminated. When the data contain outliers, the estimation errors of he-NNG are large. The estimation errors for the second contamination scheme are, for estimation of $\boldsymbol{\beta}$, outside the boundary of the plotting range, indicated with the vertical line in the plot.

It can also be seen that the homoscedastic nonnegative garrote methods, namely NNG and MM-NNG have a larger estimation error for the coefficients $\boldsymbol{\beta}$ due to the (ignored) heteroscedasticity in the model. Comparing the robust heteroscedastic nonnegative garrote methods, he-MNNG has the smallest estimation errors for the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (it has the highest efficiency compared to the he-TNNG and he-SNNG methods). For estimation of $\boldsymbol{\beta}$ the he-TNNG method has smaller estimation errors than he-SNNG. Since the latter estimation procedure uses 75% of the observations, it has a lower breakdown point, but higher efficiency than he-SNNG.

The estimation errors for the coefficients $\boldsymbol{\gamma}_1$ are low for he-MNNG and he-SNNG in all three contamination schemes, and for he-NNG in contamination scheme 1. The estimation errors of the coefficients $\boldsymbol{\gamma}_1$ for he-TNNG are now higher than these for he-SNNG. This is because he-SNNG provides estimates for $\boldsymbol{\gamma}$ that have higher efficiency than these of he-TNNG.

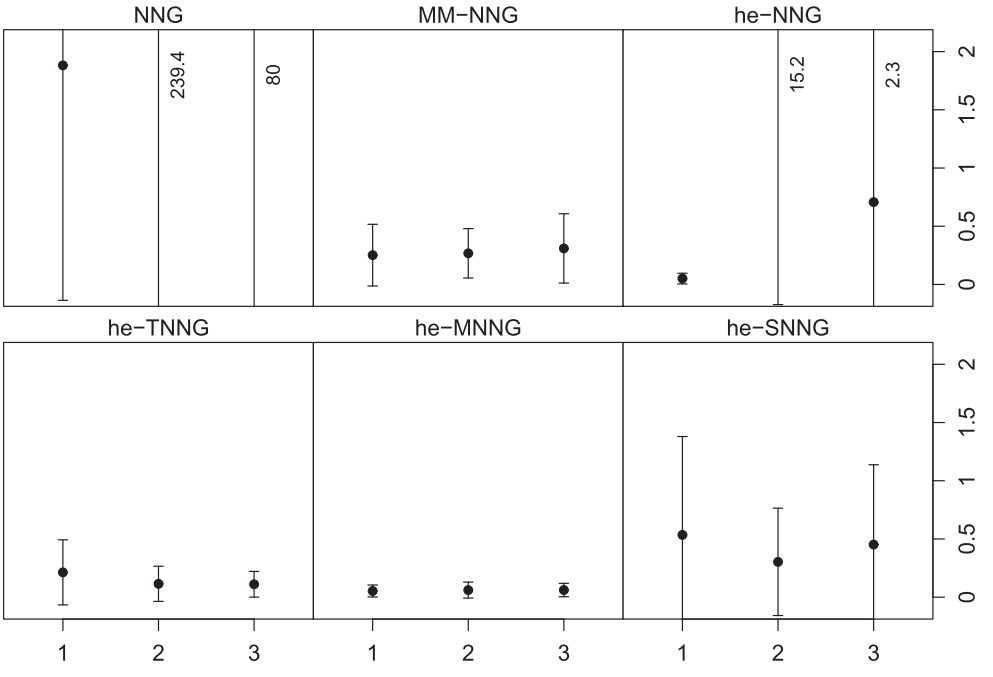

Figure 1. Mean and standard deviation of the estimation error for the regression coefficients $\boldsymbol{\beta}$ for the 3 contamination schemes (`CScheme 1`, `CScheme 2` and `CScheme 3`) for the heteroscedastic `Setting 1` of Example 1 and the different nonnegative garrote methods.

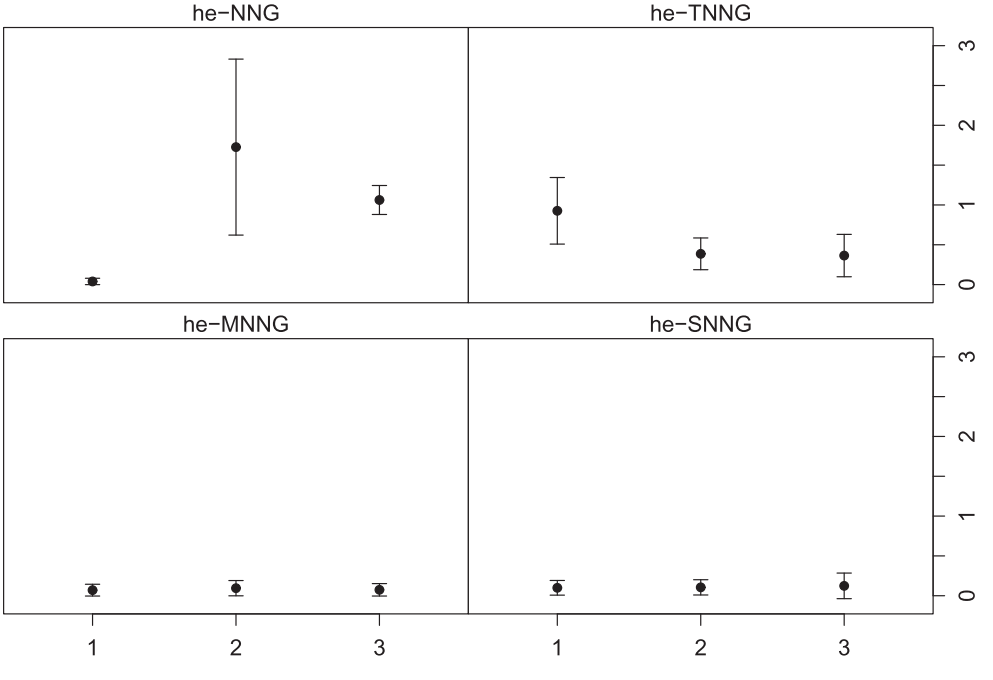

Figure 2. Mean and standard deviation of the estimation error for the regression coefficients $\boldsymbol{\gamma}_1$ of Example 1 for the 3 contamination schemes and the different nonnegative garrote methods.

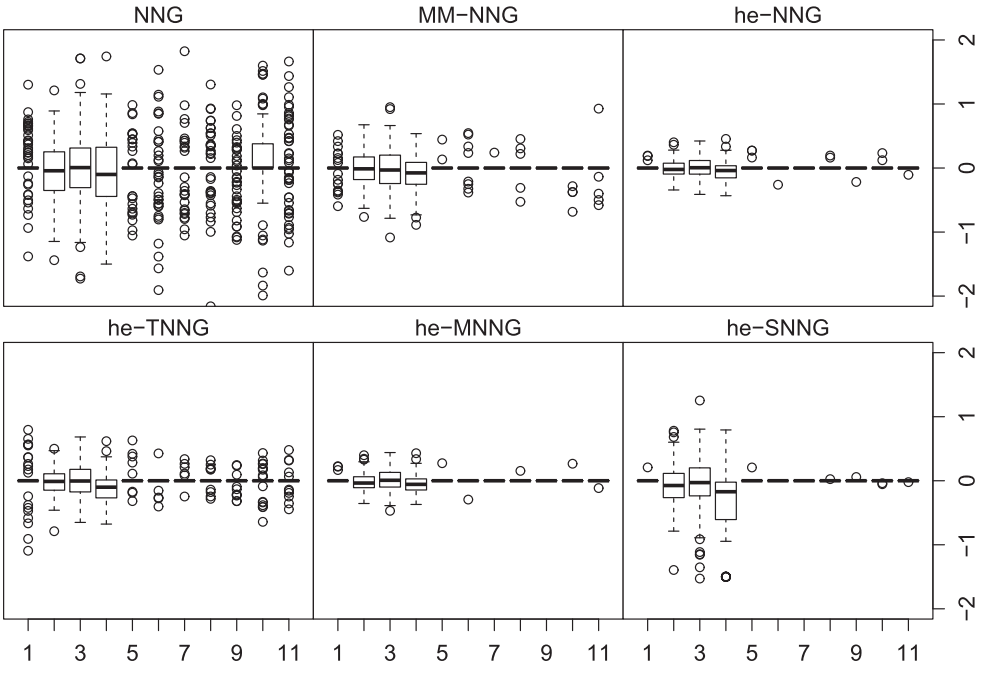

Figure 3. Difference between the estimated value and the true value of the regression coefficients β for contamination scheme 1 for heteroscedastic `Setting 1` of Example 1.

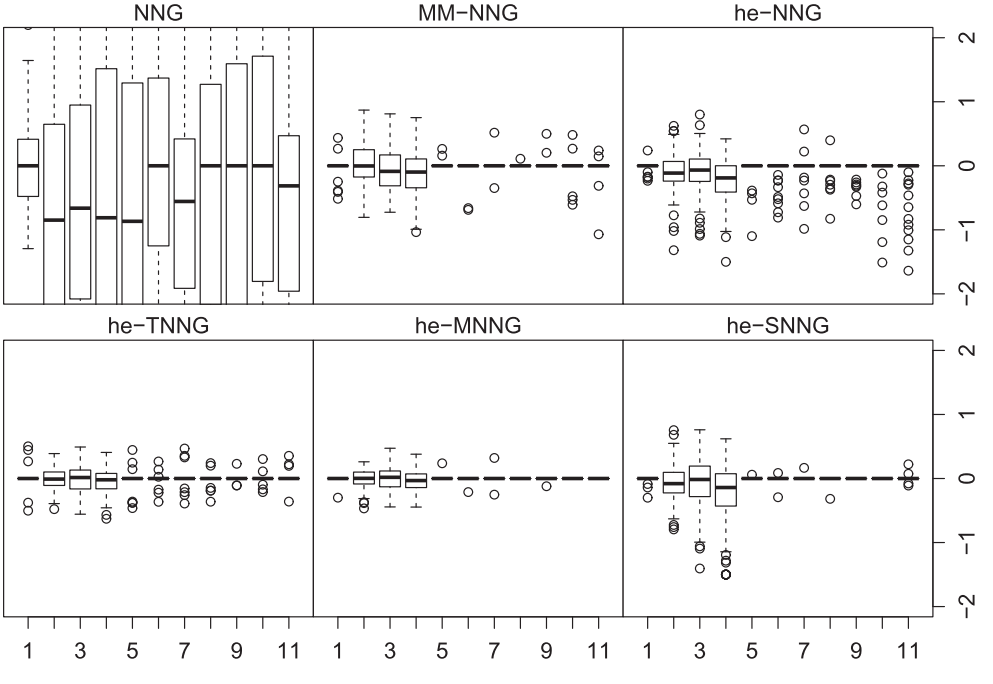

Figure 4. Difference between the estimated value and the true value of the regression coefficients β for contamination scheme 3 for heteroscedastic `Setting 1` of Example 1.

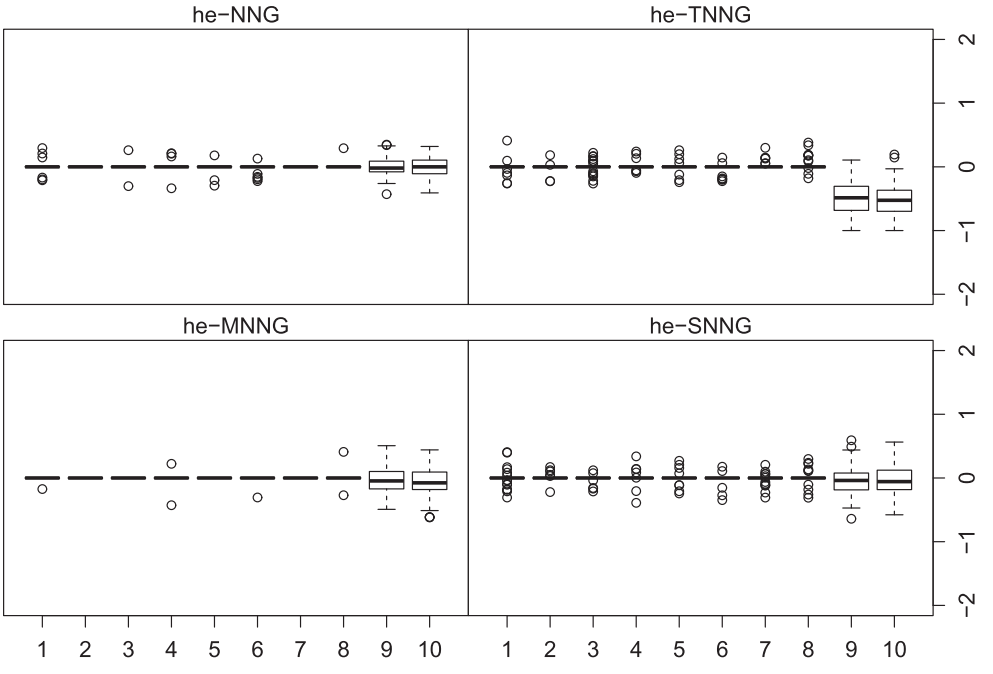

Figure 5. Difference between the estimated value and the true value of the regression coefficients γ_1 of Example 1 for contamination scheme 1.

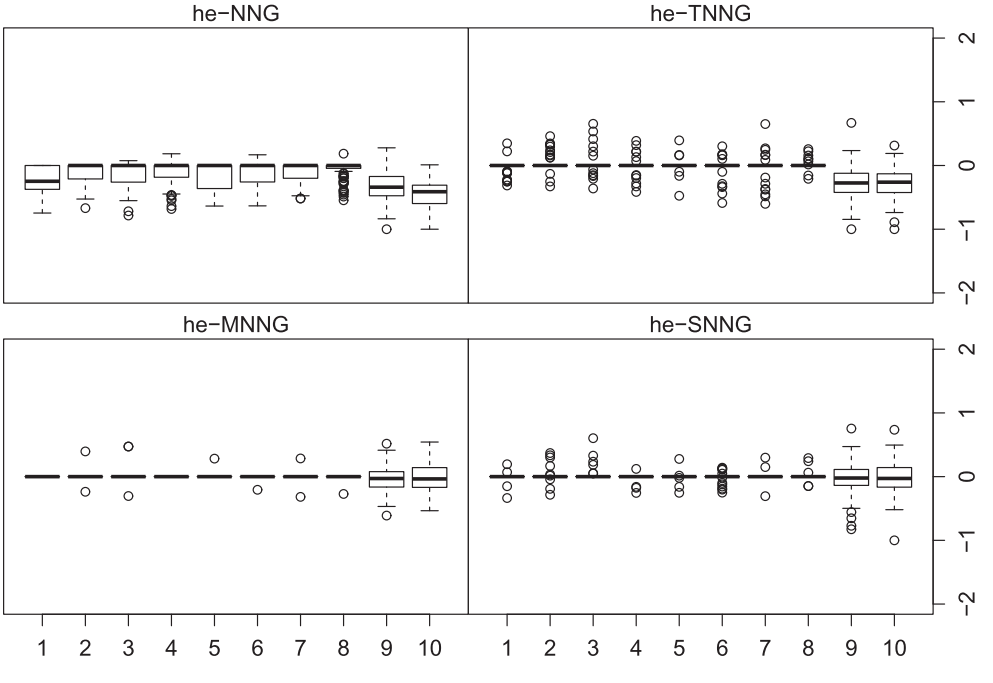

Figure 6. Difference between the estimated value and the true value of the regression coefficients γ_1 of Example 1 for contamination scheme 3.

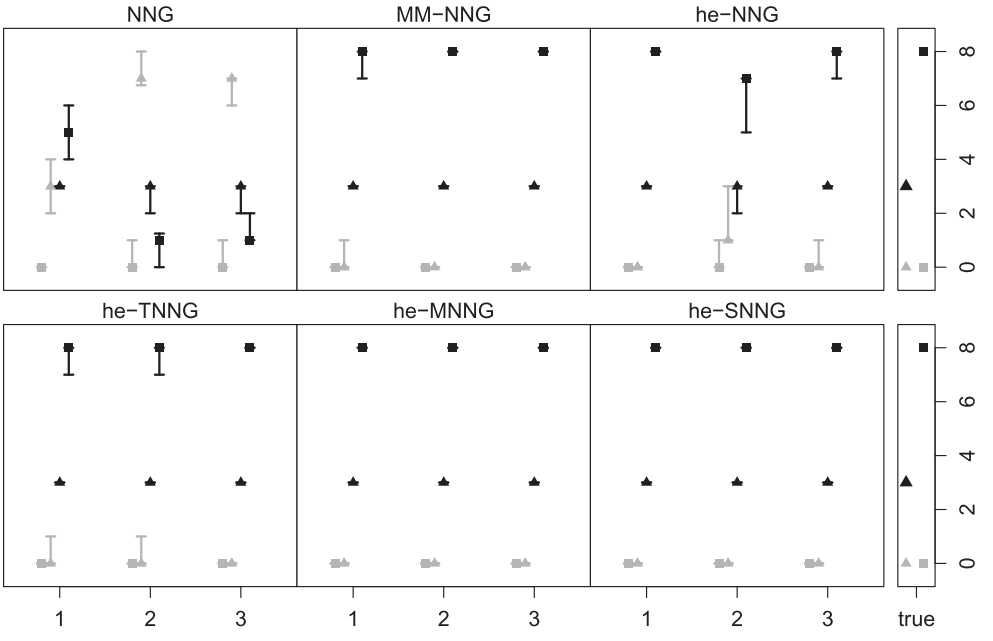

Figure 7. Median, first and third quartiles of the true zero's (), false zero's ( ), true positives () and false positives ( ) for the regression coefficients $\boldsymbol{\beta}$ for the 3 contamination schemes for heteroscedastic `Setting 1` of Example 1 and the different nonnegative garrote methods. The optimal values are plotted in the side bar.

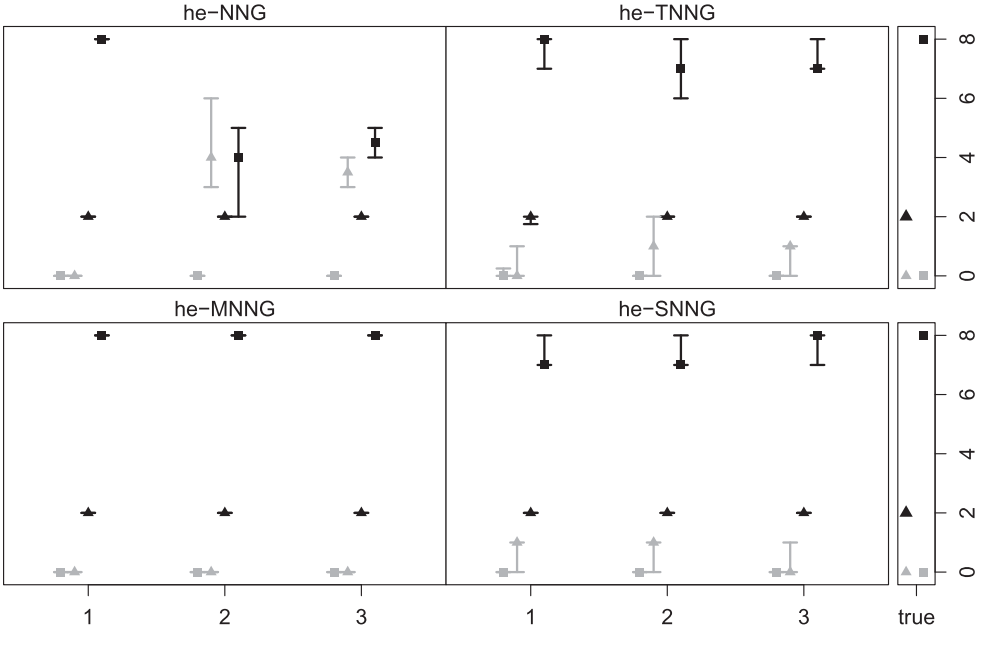

Figure 8. Median, first and third quartiles of the true zero's (), false zero's ( ), true positives () and false positives ( ) for the regression coefficients γ_1 of Example 1 for the 3 contamination schemes and the different nonnegative garrote methods. The optimal values are plotted in the side bar.

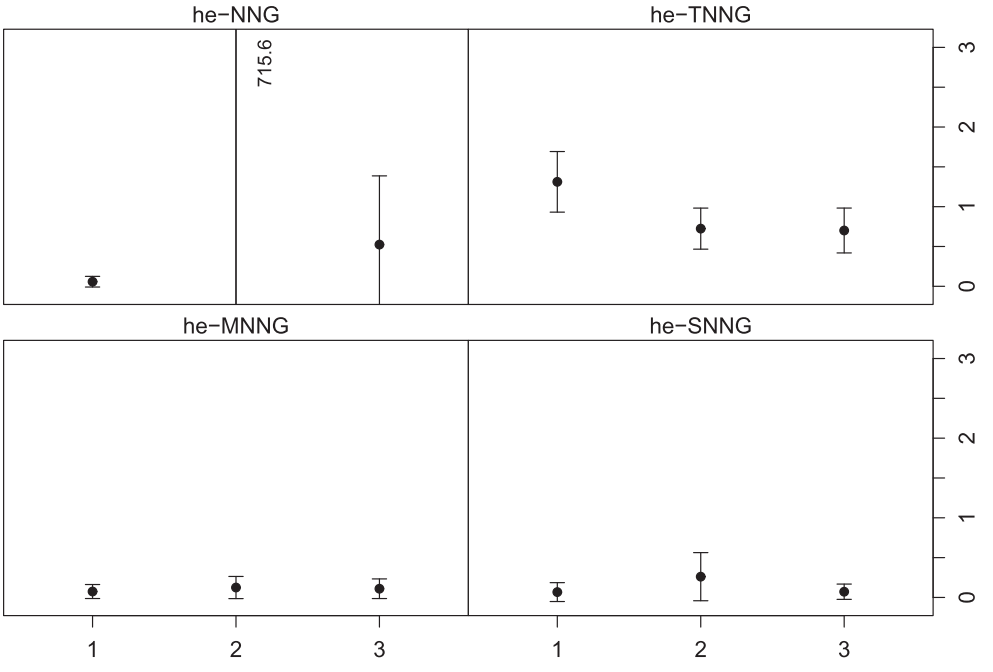

Figure 9. Mean and standard deviation of the estimation error for the regression coefficients σ for the 3 contamination schemes for heteroscedastic `Setting 1` of Example 1 and the different nonnegative garrote methods.

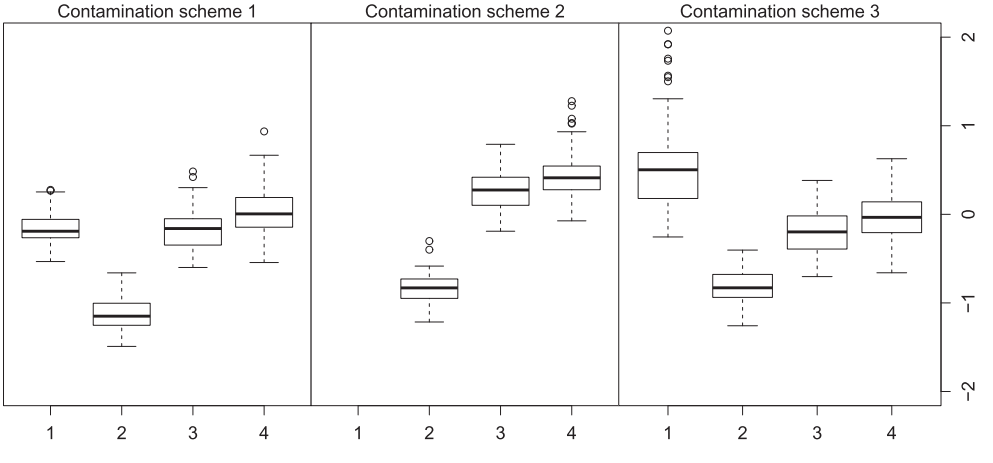

Figure 10. Difference between the estimated value and the true value of the regression coefficients σ for the 3 contamination schemes for heteroscedastic `Setting 1` of Example 1 with 1 = he-NNG, 2 = he-TNNG, 3 = he-MNNG and 4 = he-SNNG.

More detailed views of the quality of estimation for two of the three contamination schemes for the eleven coefficients of $\boldsymbol{\beta}$ are in Figures 3 and 4. The boxplots for contamination `Scheme 2` are in Figure S.1 in the Supplemental Material. Note that only coefficients 2, 3 and 4 are different from zero in the true model. Similar conclusions as for Figure 1

can be drawn. A slightly higher variability is noticeable for the he-SNNG method when compared to the he-MNNG method.

Figures 5 and 6 provide similar detailed boxplots for the estimation of γ_1 for contamination `Scheme 1` and `Scheme 3`, respectively. See Figure S.2 in the Supplementary Material part for the boxplots for contamination `Scheme 2`. Note that only variables 9 and 10 are included in the error variance in the true model. From the boxplots it can be seen again that he-MNNG performs very well. The variability in the estimates of he-SNNG and he-MNNG is the same for all three contamination schemes and is slightly bigger than the variability in the estimates of he-NNG for the first contamination scheme. Note from Figures 5 and 6 that the estimates of the he-TNNG method are biased. It is known that the homoscedastic least trimmed squares estimator underestimates the scale of the residuals. Since the coefficients $\boldsymbol{\gamma}$ are coefficients for the variance, the he-TNNG produces biased estimates for these coefficients.

We next look at the quality of the variable selection task, using the criteria from Table 4. Figures 7 and 8 present the medians, first and third quartiles of the true positives, true zero's, false positives and false zero's, for respectively the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, for the 3 contamination schemes for the heteroscedastic `Setting 1`. The nonnegative garrote methods work well when the MTP, MTZ, MFP and MFZ values are close to 3, 8, 0 and 0 respectively (the optimal values) for $\boldsymbol{\beta}$, and close to 2, 8, 0 and 0 for $\boldsymbol{\gamma}$. These optimal values are plotted in the side bars of the figures. Figure 7 reveals that NNG selects too many variables and that, when the data are contaminated, it sometimes also selects the wrong variables. The non-robust he-NNG method has the same problem for the contaminated data. The robust heteroscedastic nonnegative garrote methods as well as MM-NNG perform well in variable selection, but he-TNNG tends to select more variables than necessary, which is to a lesser extent also the case for the S-NNG method for the selection task for the error variance. In the latter case, he-SNNG also selects more false zero's than he-MNNG.

Finally, the means and standard deviations of the estimation errors for σ for the regression model with $\boldsymbol{\gamma}_1$ are presented in Figure 9 and the boxplots of the differences between the estimated values and the true value of σ are shown in Figure 10 for the four heteroscedastic methods. From these figures, we can conclude that he-MNNG and he-SNNG perform well for all three contamination schemes and that he-NNG also performs well for uncontaminated data sets. The estimation errors for he-NNG are large for the second contamination scheme (and outside the boundary of the plotting range for the boxplots). Further, the estimation errors of he-TNNG are larger due to the biased estimated values for σ.

From this simulation example we can conclude that he-MNNG works best, followed closely by the second best method, he-SNNG.

4.3. Conclusions

Conclusions drawn from both simulation examples are

- The he-MNNG method performs better than he-TNNG and he-SNNG in variable selection and estimation. Hence, it is recommended to use this method. The he-SNNG method is second best.
- If the data sets do not contain outliers, he-NNG can be used too.

- Since we use 75% of the observations in the estimation procedure of he-TNNG, the method has a higher efficiency for the estimates of the coefficients $\boldsymbol{\beta}$ than he-SNNG, but a smaller breakdown point. If a larger breakdown point is needed and only 50% of the observations are used, the efficiency of he-TNNG estimates for $\boldsymbol{\beta}$ will decrease.
- The he-TNNG method has a larger bias in the estimates for the variance coefficients $\boldsymbol{\gamma}$ and σ than he-MNNG and he-SNNG. This is a known disadvantage of these types of trimmed estimators.

5. Influence functions

The aim of this section is to investigate the influence functions of the different (robust) heteroscedastic nonnegative garrote methods that were introduced in Section 3. Since each method also involves a specific method for the initial estimators, the influence function of each nonnegative garrote method also depends on that of the method used in the initial estimation step. We derive expressions for influence functions of (i) the maximum likelihood estimator of Section 2.1, in Theorem S.3.1; (ii) the heteroscedastic S-estimator of Section 2.2.3, in Theorem S.3.2; (iii) the heteroscedastic nonnegative garrote estimator of Section 3.2.1, in Theorem S.3.3; and (iv) the heteroscedastic S-nonnegative garrote estimator of Section 3.2.4, in Theorem A.1. None of these influence functions are available in the literature. Due to the high-technicality of the results, we only state in the Appendix the influence function for the heteroscedastic S-nonnegative garrote estimator he-SNNG, being one of the recommended methods. Expressions of the influence functions for the other estimators can be found in the Supplemental Material. The proof of the theorems are very technical and not included here, but can be found in the PhD thesis of Vrinssen [23]. The proofs start along similar lines as the proof in the homoscedastic setting published in Gijbels et al. [14].

A first step in deriving an influence function is to write the functional form of the considered estimator, as is done for the he-SNNG method in Section 5.1. The expression for its influence function is given in Theorem A.1. In Section 5.2 we illustration these influence functions for some concrete example, shedding as such light on their interpretation.

5.1. Functionals

Consider the population form of the heteroscedastic linear regression model

$$Y = \mathcal{X}^{\mathrm{T}}\boldsymbol{\beta} + \sigma\sqrt{h\left(\mathcal{X}_{-}^{\mathrm{T}}\boldsymbol{\gamma}\right)}\varepsilon, \tag{46}$$

where Y is the response, $\mathcal{X} = (1, X_1, \ldots, X_p)^{\mathrm{T}}$ and $\mathcal{X}_{-} = (X_1, \ldots, X_p)^{\mathrm{T}}$ are the vectors with the p covariates, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^{\mathrm{T}}$ are vectors of unknown regression coefficients and ε is the error term with conditional mean 0 and conditional variance 1. Further, denote the cumulative distribution function of $(\mathcal{X}^{\mathrm{T}}, Y)$ by F.

The functional form of an estimator is then the functional $T(F)$ which describes the estimator on population level. By replacing F by the sample empirical distribution function denoted by F_n, based on $(\mathbf{X}_1^T, Y_1), \ldots, (\mathbf{X}_n^T, Y_n)$, one then obtains $T(F_n)$ the estimator.

The influence function of an estimator described by a functional $T(F)$ measures the effect of an infinitesimal contamination of the distribution F with the point mass

distribution δ_{P_0}, at a single point $P_0 = (\mathbf{X}_0^{\mathrm{T}}, Y_0)$ with $\mathbf{X}_0 = (1, X_{01}, \ldots, X_{0p})^{\mathrm{T}}$. Denoting the contaminated distribution by

$$F_{\epsilon,P_0} = (1-\epsilon)F + \epsilon\delta_{P_0}, \quad 0 < \epsilon < 1,$$

the influence function is defined as

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

For the functional form of the heteroscedastic S-nonnegative garrote estimator, we assume that we already have the functional forms of the initial estimator (in the numerical illustrations we used the heteroscedastic S-estimator here; see Theorem S.3.2 for its influence function). The functional form of the heteroscedastic S-nonnegative garrote shrinkage factors is then used to shrink or put to zero some components of the functional forms of the initial estimator. Let $\boldsymbol{\beta}^{\text{init}}(F) = (\beta_0^{\text{init}}(F), \ldots, \beta_p^{\text{init}}(F))^{\mathrm{T}}$ and $\boldsymbol{\gamma}^{\text{init}}(F) = (\gamma_1^{\text{init}}(F), \ldots, \gamma_p^{\text{init}}(F))^{\mathrm{T}}$ be the functional forms of the initial estimators and denote $\beta_j^{\text{init}}(F)X_j$ with $Z_j(F)$ for $j = 1, \ldots, p$ and $\gamma_j^{\text{init}}(F)X_j$ with $Q_j(F)$ for $j = 1, \ldots, p$. Also let $\mathcal{Z}(F) = (Z_1(F), \ldots, Z_p(F))^{\mathrm{T}}$ and $\mathcal{Q}(F) = (Q_1(F), \ldots, Q_p(F))^{\mathrm{T}}$.

The functional form of the heteroscedastic S-nonnegative garrote shrinkage factors $\boldsymbol{\alpha}^{\text{he-SNNG}}(F) = ((\mathbf{c}^{\text{he-SNNG}}(F))^{\mathrm{T}}, (\mathbf{d}^{\text{he-SNNG}}(F))^{\mathrm{T}}, \sigma^{\text{he-SNNG}}(F))^{\mathrm{T}}$, with $\mathbf{c}^{\text{he-SNNG}}(F) = (c_1^{\text{he-SNNG}}(F), \ldots, c_p^{\text{he-SNNG}}(F))^{\mathrm{T}}$ and $\mathbf{d}^{\text{he-SNNG}}(F) = (d_1^{\text{he-SNNG}}(F), \ldots, d_p^{\text{he-SNNG}}(F))^{\mathrm{T}}$, can be found by minimizing

$$\sigma + \lambda_1 \sum_{j=1}^{p} c_j \quad \text{such that } c_j \geq 0 \ (j = 1, \ldots, p),$$

for $\mathbf{c} \in \mathbb{R}_+^p$, where $\sigma \in \mathbb{R}_+ \setminus \{0\}$ solves

$$\int w(\mathcal{X}_-) \left[\rho\left(\frac{Y - \beta_0^{\text{init}}(F) - \mathcal{Z}^{\mathrm{T}}(F)\mathbf{c}}{\sigma\sqrt{h\left(\mathcal{Q}^{\mathrm{T}}(F)\mathbf{d}\right)}} \right) - b \right] \mathrm{d}F(\mathcal{X}^{\mathrm{T}}, Y) = 0,$$

and $\mathbf{d} = (d_1, \ldots, d_p)^{\mathrm{T}} \in \mathbb{R}_+^p$, where the positive d_j solve

$$-\int w(\mathcal{X}_-) \left[\rho\left(\frac{Y - \beta_0^{\text{init}}(F) - \mathcal{Z}^{\mathrm{T}}(F)\mathbf{c}}{\sigma\sqrt{h\left(\mathcal{Q}^{\mathrm{T}}(F)\mathbf{d}\right)}} \right) - b \right] \frac{h'\left(\mathcal{Q}^{\mathrm{T}}(F)\mathbf{d}\right)}{h\left(\mathcal{Q}^{\mathrm{T}}(F)\mathbf{d}\right)} Q_j(F)\mathrm{d}F(\mathcal{X}^{\mathrm{T}}, Y) + \lambda_2 = 0.$$

The functional form of the heteroscedastic S-nonnegative garrote estimator is then given by $((\boldsymbol{\beta}^{\text{he-SNNG}}(F))^{\mathrm{T}}, (\boldsymbol{\gamma}^{\text{he-SNNG}}(F))^{\mathrm{T}}, \sigma^{\text{he-SNNG}}(F))^{\mathrm{T}}$ with $\boldsymbol{\beta}^{\text{he-SNNG}}(F) = (\beta_0^{\text{he-SNNG}}(F), \ldots, \beta_p^{\text{he-SNNG}}(F))^{\mathrm{T}}$ and $\boldsymbol{\gamma}^{\text{he-SNNG}}(F) = (\gamma_1^{\text{he-SNNG}}(F), \ldots, \gamma_p^{\text{he-SNNG}}(F))^{\mathrm{T}}$ where $\beta_j^{\text{he-SNNG}}(F) = c_j^{\text{he-SNNG}}(F)\beta_j^{\text{init}}(F)$, for $j = 1, \ldots, p$, $\boldsymbol{\beta}_0^{\text{he-SNNG}}(F) = \beta_0^{\text{init}}(F)$, and $\gamma_j^{\text{he-SNNG}}(F) = d_j^{\text{he-SNNG}}(F)\gamma_j^{\text{init}}(F)$, for $j = 1, \ldots, p$.

We assume that all the functional forms defined in this section are continuous in F. Note that if $F = F_n$, is the empirical distribution function corresponding to the sample $\mathbf{P}_n = \{P_1, \ldots, P_n\}$ with $P_i = (1, X_{1i}, \ldots, X_{pi}, Y_i)$, these optimization problems are equivalent to the optimization problems at sample level of Section 3.2.4.

5.2. Illustrations of the influence functions

Although the expressions for the influence functions are rather complex, they are in fact quite useful as will be seen from the examples provided in this section.

For these illustrations we consider the heteroscedastic linear regression model (46) with only one covariate X that is independent from the error term ε. Both X and ε are standard normal distributed. The values for the coefficients are $\boldsymbol{\beta} = (0, 2)^{\mathrm{T}}$, $\gamma = 1$ and $\sigma = 1$, and the exponential function is used to model the variance, i.e., $h(x) = \exp(x)$. For the regularization parameters we used $(\lambda_1, \lambda_2) = (0.1, 0.01)$ in the illustrations.

For each of the four methods for which the expressions for the influence functions are provided (see Theorem A.1 and Theorems S.3.1–S.3.3), the influence functions for the coefficients $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and σ are plotted in respectively Figures 11–13. The blue lines in these plots correspond to the true regression line ($y = 2x$). Note the differences in the scales for the figures of the robust methods (the subfigures (a)& (c)) and nonrobust heteroscedastic methods (the subfigures (b)& (d)). From these figures it can be clearly seen that the influence functions of the ML-estimator and he-NNG estimator are unbounded when x_0 is negative. When x_0 is negative, the term $h(x_0\gamma)$ goes to zero, which results in large standardized residuals $u_0 = (y_0 - x_0\beta)/\sigma\sqrt{h(x_0\gamma)}$. For positive values for x_0, the term $h(x_0\gamma)$ increases fast, such that the standardized residuals are small.

The influence functions for the heteroscedastic S-estimator and he-SNNG estimator are bounded for regression outliers (i.e., points far away from the regression line). The fluctuations in the influence function for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for values for x_0 ranging from -5 to 5 come from the term $w(x_{0-})\psi(u_0)u_0$ in the expression of the influence function for $\boldsymbol{\gamma}$. These fluctuations are stronger for the he-SNNG method than for the heteroscedastic S-estimator (compare subfigures (c) and (a) in Figures 11 and 12). From the influence functions for the coefficients $\boldsymbol{\beta}$ it can be seen that the observations (x_0, y_0) with small residuals $y_0 - x_0\beta$ (the observations that follow the true regression line) have more influence on the heteroscedastic S-estimator and he-SNNG estimator than those observations with large residuals. Especially, the observations with negative values for x_0 that follow the true regression line have a large influence on the robust heteroscedastic estimators. The influence functions for σ of the robust heteroscedastic methods are bounded for regression outliers and they stay rather flat for all observations.

Figure S.19 in the Supplemental Material depicts the differences in the influence functions of the he-SNNG estimators when the additional weights $w(x_{0-})$ are not included in the estimation procedure. This figure clearly demonstrates the importance of including the weights $w(x_{0-})$.

6. Real data application

As an illustration we consider the data set on progression of diabetes available in the R-package *lars*. This data set contains 442 observations and 10 covariates. The 10 covariates

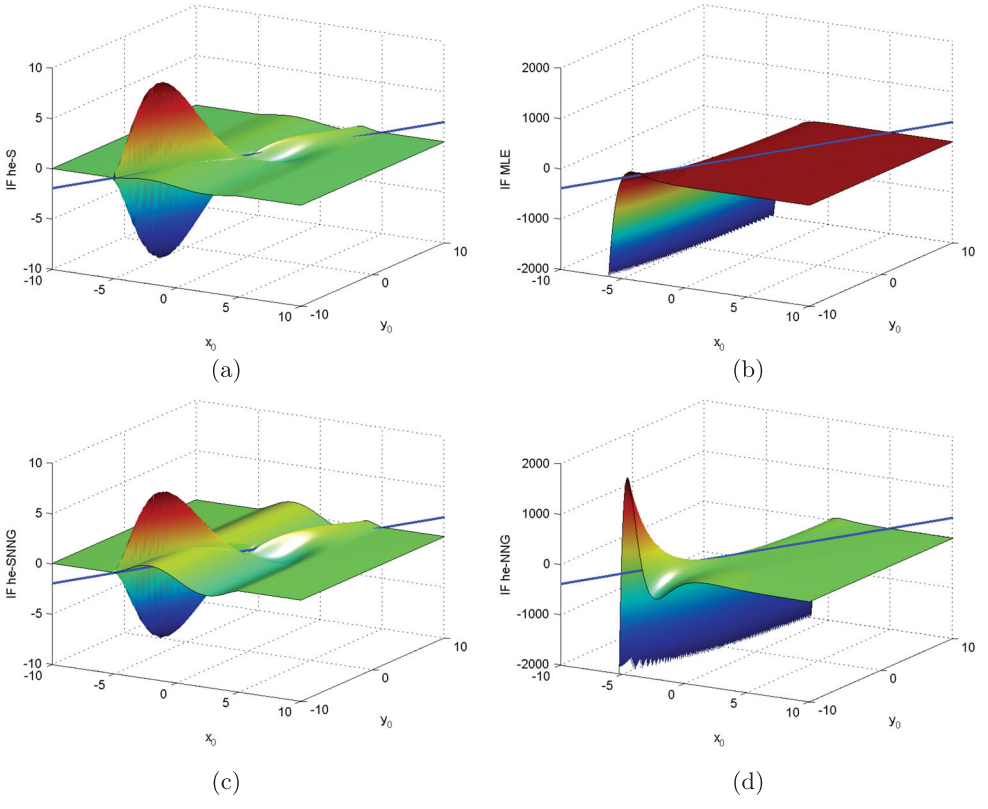

Figure 11. Influence functions for the heteroscedastic S-estimator, the maximum likelihood estimator, the heteroscedastic S-nonnegative garrote estimator and the heteroscedastic nonnegative garrote estimator for coefficients $\boldsymbol{\beta}$. (a) heteroscedastic S-estimator, (b) ML-estimator, (c) he-SNNG estimator, (d) he-NNG estimator.

are age (*age*), sex (*sex*), body mass index (*bmi*), average blood pressure (*map*) and six blood serum measurements (*tc*, *ldl*, *hdl*, *tch*, *ltg* and *glu*); and the response variable is a quantitative measure of disease progression one year after the first measurements. The response variable as well as the covariates were all standardized before our analysis. A possible test for testing heteroscedasticity is the Breusch-Pagan test (see [24]). This test is performed as follows. First estimate the model $Y_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i$ with OLS and compute the regression residuals $u_i = Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$. Then perform OLS regression on the second model $\log(u_i^2) = \mathbf{X}_{i-}^{\mathrm{T}}\boldsymbol{\gamma} + \eta$. The null hypothesis states that $\boldsymbol{\gamma} = \mathbf{0}_p$ and the test statistic is now given by $\chi^2 = nR^2 \sim \chi_p^2$ (under the null hypothesis), where R^2 is the R-squared value of the second model. Hence, the test statistic is the ratio of the explained variance to the total variance in the second model. The null hypothesis is rejected if too much of the variance is explained by the covariates. The Breusch-Pagan test has for this linear regression model a p-value of 0.000527, which means that the error terms are heteroscedastic. Therefore, we apply the maximum likelihood method on the heteroscedastic regression

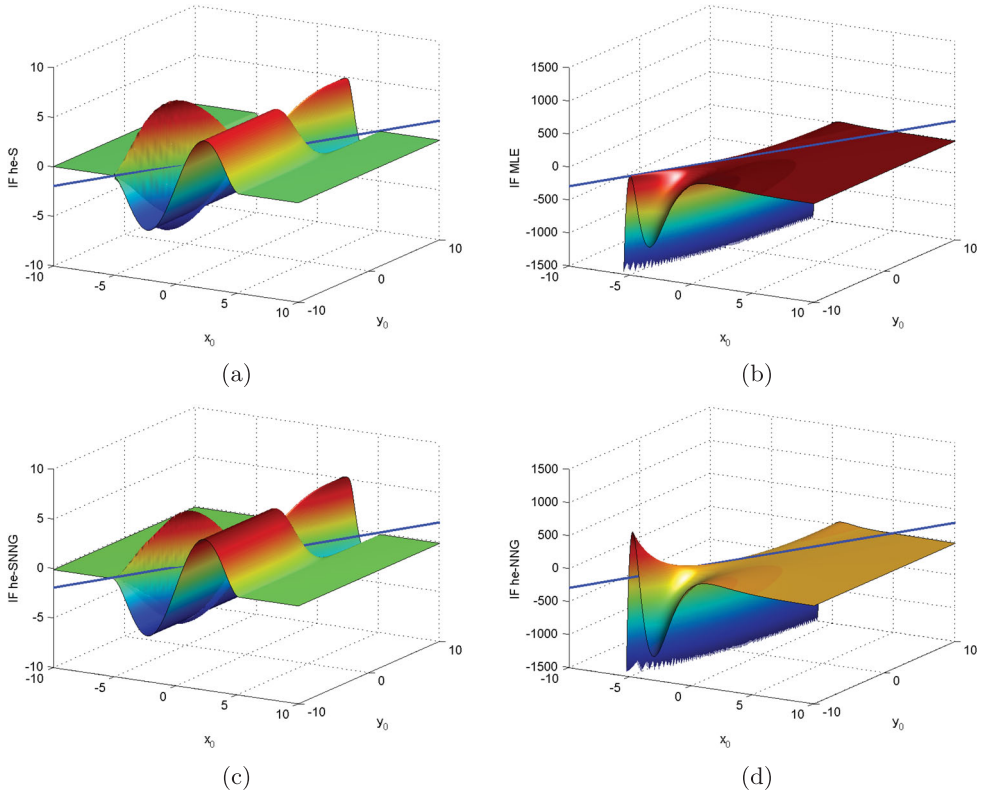

Figure 12. Influence functions for the heteroscedastic S-estimator, the maximum likelihood estimator, the heteroscedastic S-nonnegative garrote estimator and the heteroscedastic nonnegative garrote estimator for coefficients $\boldsymbol{\gamma}$. (a) heteroscedastic S-estimator, (b) ML-estimator, (c) he-SNNG estimator, (d) he-NNG estimator.

model (1) (with $h(x) = \exp(x)$) and we transform the diabetes data set to a (hopefully) homoscedastic data set by dividing the response and all the covariate measurements of each observation with the estimated error standard deviation that belongs to that observation. The Breusch-Pagan test has for this transformed data set a p-value of 0.88953, which means that we cannot reject the null hypothesis that this data set is homoscedastic.

The regularization paths of the nonnegative garrote method for the original and transformed data set are shown in Figure 14. On the horizontal axis the value of s is plotted. Smaller values of s (i.e., when moving from right to left) lead to less variables included (more shrinkage). From these plots it can be seen that the regularization paths are different for both data sets and that the variable *hdl* is selected for larger s in the original data set and not in the transformed data set. So, because of the heteroscedasticity in the data set the nonnegative garrote method tends to select too many variables.

To further improve the model we also include the quadratic terms for 9 covariates (not for the dummy variable *sex*) in the linear regression model. Nott et al. [25] analyzed this data set using a heteroscedastic linear regression model, but they included also the interaction terms between the covariates. However, the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is then almost singular,

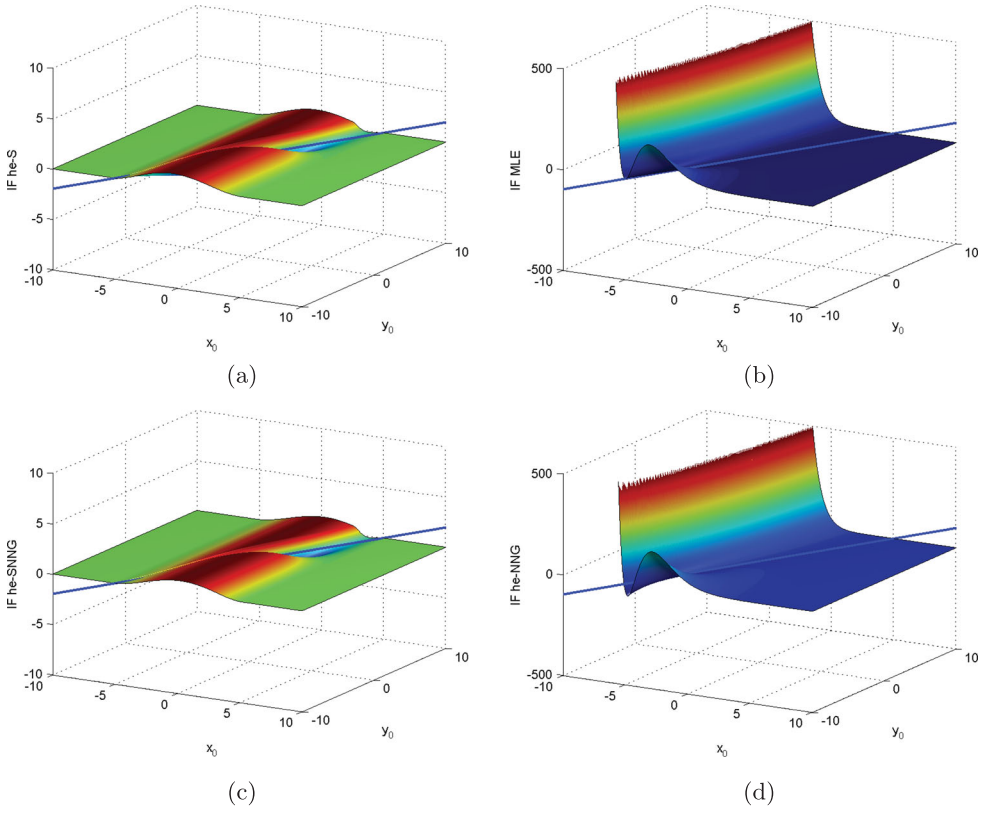

Figure 13. Influence functions for the heteroscedastic S-estimator, the maximum likelihood estimator, the heteroscedastic S-nonnegative garrote estimator and the heteroscedastic nonnegative garrote estimator for σ. (a) heteroscedastic S-estimator, (b) ML-estimator, (c) he-SNNG estimator, (d) he-NNG estimator.

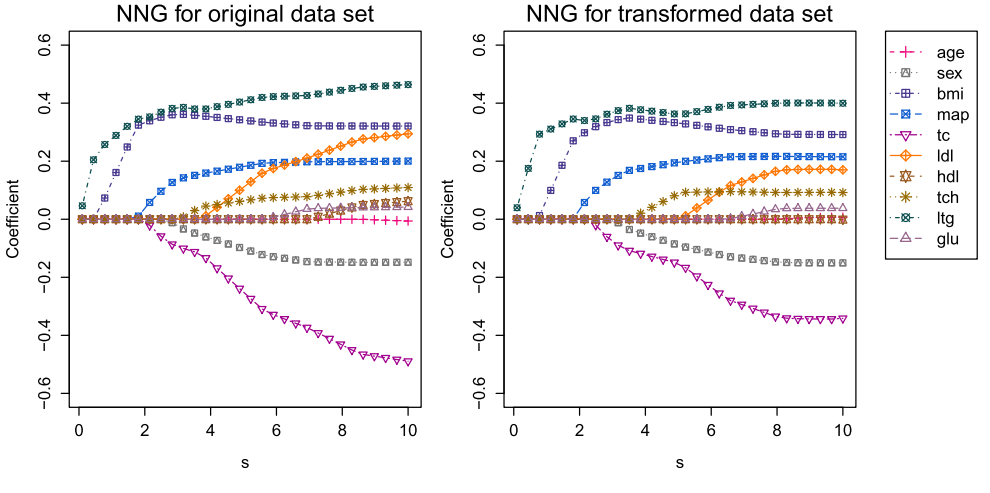

Figure 14. Diabetes data: estimated coefficients in function of the regularization parameter for NNG for the original and the transformed data set.

which leads to problems in computing the initial estimators. Therefore, we do not include the interaction terms in this analysis. A (robust) heteroscedastic ridge estimator could be used so that also the interaction terms can be included in the model, but to the best of our knowledge such a method does not exist so far.

We next apply the methods listed in Table 2 to these data, using the appropriate (robust) BIC criterion for each method, to select the regularization parameter(s). In Table 6 the selected variables for explaining the mean response (indicated with ✓) and the variance (indicated with ⋆) can be found in the columns entitled 'original data'. The variables *sex*, *bmi*, *map*, *tc*, *ldl*, *ltg* and ltg^2 are selected to explain the response by all the heteroscedastic nonnegative garrote methods, but he-TNNG selects many more variables. The difference between NNG and the heteroscedastic nonnegative garrote methods is that NNG selects the variable glu^2 instead of ltg^2, MM-NNG selects both variables.

The variables that are selected to explain the variance differ among the four heteroscedastic nonnegative garrote methods. If the Breusch-Pagan test is used to test for heteroscedasticity for each variable separately in the full model (all variables are used for explaining the response, but only one variable is used for explaining the variance), one can find that the null hypothesis for testing homoscedasticity is rejected on a significance level of 1% for the variables *hdl*, *tch*, *ltg* and hdl^2, and the variable *hdl* has the smallest *p*-value. The method he-MNNG selects of these variables only the variable *hdl* and he-NNG selects the variable *ltg*, but also two variables for which the Breusch-Pagan test is not rejected. If the Breusch-Pagan test is applied to the model found by he-MNNG, the null hypothesis for testing homoscedasticity is not rejected with a *p*-value of 0.1093. If the same is done for the model found by he-NNG, the null hypothesis is not rejected with a *p*-value of 0.0834.

Table 7 gives the RSS (Residual Sum of Squares) for the homoscedastic and the heteroscedastic methods respectively, i.e. ,

$$\text{RSS} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\right)^2 \qquad \text{RSS} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\right)^2}{\exp\left(\mathbf{X}_{i-}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right)}.$$

Comparing the RSS values, we see that he-MNNG has the lowest value for the RSS. Hence, using a heteroscedastic regression model improves the results for this data set. The values of RSS for he-SNNG and he-TNNG are larger than these of other methods, probably because they have selected too many variables for explaining the variance. From the simulation study in Section 4 we already concluded that he-NNG and he-MNNG perform better than he-SNNG and he-TNNG in case of no contamination.

We will now compare the robustness of the different methods by introducing vertical outliers in the diabetes data. The values of the response are in the range from 25 to 346 with a mean of 152.13 and a standard deviation of 77.09. Therefore, we introduce vertical outliers in the data set by randomly replacing 10% of the values of the response with values coming from a $N(500, 1)$-distribution. Results are in Table 6, in the columns entitled 'data with outliers added', and in the last column of Table 7 (where the RSS are computed without the vertical outliers). From the last column of Table 7 it can be seen that he-MNNG performs better than the other methods, since it has the lowest value for the RSS and it is the only method that selects the same variables as for the original data set (see Table 6).

Table 6. Diabetes data: selected covariates for explaining the response.

	Original data										Data with outliers added									
			Heteroscedastic methods										Heteroscedastic methods							
Variable	NNG		MM-NNG		he-NNG		he-MNNG		he-SNNG		he-TNNG		NNG		MM-NNG		he-NNG		he-MNNG		he-SNNG		he-TNNG	
age									⋆	✓	⋆										⋆			
sex	✓		✓		✓		✓		⋆	✓	⋆		✓		✓				✓		✓	⋆	✓	⋆
bmi	✓		✓		✓		✓		⋆	✓			✓		✓				✓		✓	⋆	✓	⋆
map	✓		✓		✓		✓		⋆	✓			✓		✓				✓			⋆	✓	⋆
tc	✓		✓		✓	⋆	✓		⋆	✓	⋆		✓		✓		⋆		✓		✓	⋆	✓	⋆
ldl	✓		✓		✓	⋆	✓		⋆	✓	⋆		✓		✓		⋆		✓		✓	⋆	✓	⋆
hdl								⋆	⋆	✓										⋆		⋆	✓	
tch											⋆		✓										✓	⋆
ltg	✓		✓		✓	⋆	✓			✓			✓		✓		✓	⋆	✓		✓		✓	
glu										✓	⋆		✓					⋆				⋆		⋆
age^2																								
bmi^2										✓													✓	
map^2									⋆	✓			✓									⋆	✓	⋆
tc^2										✓													✓	⋆
ldl^2										✓														⋆
hdl^2										✓	⋆											⋆		⋆
tch^2									⋆													⋆		
ltg^2			✓		✓		✓		⋆	✓	⋆						⋆		✓		✓	⋆	✓	⋆
glu^2	✓		✓							✓	⋆												✓	⋆

Table 7. Diabetes data: RSS.

Method	RSS (original data)	RSS (contaminated data)
NNG	2872.02	4284.82
MM-NNG	2869.54	2978.99
he-NNG	2870.85	4839.05
he-MNNG	2842.95	2903.69
he-SNNG	3519.91	4258.98
he-TNNG	4328.08	5402.66

7. Conclusion and discussion

In this paper we have studied an appropriate S-estimation method in a heteroscedastic linear regression model, and proposed several (robust) methods for variable selection in such models. For each discussed method we provide algorithms. Although there is no formal guarantee that the optimization problems have only one global minimum, we never experienced any problem in the numerical studies.

Throughout the paper we assumed that the function h which describes how the error variance depends on the linear term (the $\mathbf{X}_{-i}^T\boldsymbol{\gamma}$ term), is known. It would be interesting to look into the situation when h is not known. One approach could be to approximate σ_i by a P-spline approximation of the covariates and use ideas of flexible P-spline estimation in variable selection. See for example Antoniadis et al. [12,13]. This is part of future research.

Acknowledgments

The authors are grateful to the anonymous reviewers for their comments on an earlier version of the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research is supported by Federaal Wetenschapsbeleid IAP Research Networks [Network P7/06] of the Belgian State (Belgian Science Policy), the Research Fund of the KU Leuven [project GOA/12/014], and the Research Foundation Flanders (FWO) [grant 1.5.137.13N].

References

[1] Jobson JD, Fuller WA. Least squares estimation when the covariance matrix and parameter vector are functionally related. J Am Stat Assoc. 1980;75:176–181.
[2] Carroll RJ, Ruppert D. Robust estimation in heteroscedastic linear models. Ann Stats. 1982;10:429–441.
[3] Bianco A, Boente G, di Rienzo J.. Some results for robust GM-based estimators in heteroscedastic regression models. J Stat Plan Inference. 2000;89:215–242.

[4] Bianco A, Boente G. On the asymptotic behavior of one-step estimates in heteroscedastic regression models. Stat Probab Lett. 2002;60:33–47.
[5] Atkinson AC, Riani M, Torti F. Robust methods for heteroskedastic regression. Comput Stat Data Anal. 2016;104:209–222.
[6] Serneels S, Croux C, Filzmoser P, et al. Partial robust M-regression. Chemometr Intell Lab Syst. 2005;79:55–64.
[7] Rousseeuw PJ, Yohai VJ. Robust regression by means of Sestimators. Robust and nonlinear time series analysis (Heidelberg, 1983). New York: Springer; 1984. p. 256–272. (Lecture Notes in Statistics, Vol. 26).
[8] Yohai VJ, Zamar RH. High breakdown-point estimates of regression by means of the minimization of an efficient scale. J Am Stat Assoc. 1988;83:406–413.
[9] Maronna RA, Martin RD, Yohai VJ. Theory and methods. In: Robust statistics. Chichester: John Wiley & Sons, Ltd.; 2006. (Wiley Series in Probability and Statistics).
[10] Box GEP, Hill WJ. Correcting inhomogeneity of variance with power transformation weighting. Technometrics. 1974;16:385–389.
[11] Bickel PJ. Using residuals robustly I. Test for heteroscedasticity, nonlinearity. Ann Stat. 1978;6:266–291.
[12] Antoniadis A, Gijbels I, Verhasselt A. Variable selection in varying coefficient models using P-splines. J Comput Graph Stat. 2012;21:638–661.
[13] Antoniadis A, Gijbels I, Verhasselt A. Variable selection in additive models using P-splines. Technometrics. 2012;54(4):425–438.
[14] Gijbels I, Verhasselt A, Vrinssen I. Consistency and robustness properties of the S-nonnegative garrote estimator. Statistics. 2017;51:921–947.
[15] Hadi AS, Luceño A.. Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. Comput Stat Data Anal. 1997;25:251–272.
[16] Vandev DL, Neykov NM. About regression estimators with high breakdown point. Statistics. 1998;32:111–129.
[17] Cheng T-C. Robust diagnostics for the heteroscedastic regression model. Comput Stat Data Anal. 2011;55:1845–1866.
[18] Hössjer O, Croux C. Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. J Nonparametr Stat. 1995;4:293–308.
[19] Slock P, Van Aelst S, Salibian-Barrera M. A fast algorithm for *S*-estimation in robust heteroscedastic regression. Presentation at the 6th International Conference of the ERCIM WG on Computing & Statistics, ERCIM 2013, London, UK, December 14–16, 2013.
[20] Salibian-Barrera M, Yohai VJ. A fast algorithm for S-regression estimates. J Comput Graph Stat. 2006;15:414–427.
[21] Gijbels I, Vrinssen I. Robust nonnegative garrote variable selection in linear regression. Comput Stat Data Anal. 2015;85:1–22.
[22] Breiman L. Better subset regression using the nonnegative garrote. Technometrics. 1995;37:373–384.
[23] Vrinssen I. Robust nonnegative garrote variable selection in (heteroscedastic) linear regression [doctoral dissertation]. Belgium: KU Leuven; 2015.
[24] Breusch TS, Pagan AR. A simple test for heteroskedasticity and random coefficient variation. Econometrica. 1979;47(5):1287–1294.
[25] Nott DJ, Tran M-N, Leng C. Variational approximation for heteroscedastic linear models and matching pursuit algorithms. Stat Comput. 2012;22:497–512.

Appendix. Influence function for the heteroscedastic S-nonnegative garrote estimator

To simplify the notation, we will use $\mathrm{d}F$ for $\mathrm{d}F(\mathcal{X}^{\mathrm{T}}, Y)$ in the sequel. We introduce the following notations:

$$u_F = \frac{Y - \mathcal{X}^{\mathrm{T}}\boldsymbol{\beta}^{\mathrm{he-SNNG}}(F)}{\sigma^{\mathrm{he-SNNG}}(F)\sqrt{h\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\mathrm{he-SNNG}}(F)\right)}}, \quad u_0 = \frac{Y_0 - \mathbf{X}_0^{\mathrm{T}}\boldsymbol{\beta}^{\mathrm{he-SNNG}}(F)}{\sigma^{\mathrm{he-SNNG}}(F)\sqrt{h\left(\mathbf{X}_{0-}^{\mathrm{T}}\boldsymbol{\gamma}^{\mathrm{he-SNNG}}(F)\right)}}.$$

Further, the following sets of indices are introduced. For given $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$, we denote the set of indices containing the non-zero regression coefficients of $\boldsymbol{\beta}^{\mathrm{he-SNNG}}(F)$ and $\boldsymbol{\gamma}^{\mathrm{he-SNNG}}(F)$ respectively with $\mathcal{S}_{\lambda_1} = \{j : \beta_j^{\mathrm{he-SNNG}}(F) \neq 0\}$ and $\mathcal{S}_{\lambda_2} = \{j : \gamma_j^{\mathrm{he-SNNG}}(F) \neq 0\}$ and the set of indices containing the zero regression coefficients of $\boldsymbol{\beta}^{\mathrm{he-SNNG}}(F)$ and $\boldsymbol{\gamma}^{\mathrm{he-SNNG}}(F)$ respectively with $\mathcal{N}_{\lambda_1} = \{j : \beta_j^{\mathrm{he-SNNG}}(F) = 0\}$ and $\mathcal{N}_{\lambda_2} = \{j : \gamma_j^{\mathrm{he-SNNG}}(F) = 0\}$.

The expressions for the influence functions of the functional forms of the heteroscedastic S-nonnegative garrote estimators are in Theorem A.1, and use the following notations:

$$\mu_1 = \int w(\mathcal{X}_-)\psi(u_F)u_F\mathrm{d}F, \quad \nu_2 = \int w(\mathcal{X}_-)\psi'(u_F)u_F^2\mathrm{d}F, \tag{A1}$$

$$\mathbf{a}_F = \int w(\mathcal{X}_-)\psi(u_F)\frac{\mathcal{X}}{\sigma^{\text{he-SNNG}}(F)\sqrt{h\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}}\mathrm{d}F,$$

$$\mathbf{b}_F = \int w(\mathcal{X}_-)\psi'(u_F)\frac{u_F\mathcal{X}}{\sqrt{h\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}}\mathrm{d}F,$$

$$\mathbf{c}_F = \int w(\mathcal{X}_-)\psi(u_F)u_F\frac{h'\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}{h\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}\mathcal{X}_-\mathrm{d}F,$$

$$\mathbf{d}_F = \int w(\mathcal{X}_-)\psi'(u_F)\frac{u_F^2h'\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}{h\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}\mathcal{X}_-\mathrm{d}F, \tag{A2}$$

and

$$\mathbf{A} = \int w(\mathcal{X}_-)\psi'(u_F)\frac{\mathcal{X}\mathcal{X}^{\mathrm{T}}}{\sigma^{\text{he-SNNG}}(F)h\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}\mathrm{d}F,$$

$$\mathbf{B} = \int w(\mathcal{X}_-)\psi(u_F)\frac{h'\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}{\sigma^{\text{he-SNNG}}(F)h^{3/2}\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}\mathcal{X}_-\mathcal{X}^{\mathrm{T}}\mathrm{d}F,$$

$$\mathbf{C} = \int w(\mathcal{X}_-)\psi(u_F)u_F\frac{\left(h'\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)\right)^2}{h^2\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}\mathcal{X}_-\mathcal{X}_-^{\mathrm{T}}\mathrm{d}F,$$

$$\mathbf{D} = \int w(\mathcal{X}_-)\left(\rho(u_F) - b\right)\left(\frac{h''\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}{h\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)} - \frac{\left(h'\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)\right)^2}{h^2\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}\right)\mathcal{X}_-\mathcal{X}_-^{\mathrm{T}}\mathrm{d}F,$$

$$\mathbf{E} = \int w(\mathcal{X}_-)\psi'(u_F)u_F\frac{h'\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}{h^{3/2}\left(\mathcal{X}_-^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}\mathcal{X}\mathcal{X}_-^{\mathrm{T}}\mathrm{d}F, \tag{A3}$$

Theorem A.1: *Let* $\lambda_1 \geq 0$ *and* $\lambda_2 \geq 0$. *The influence function of the functional form of the heteroscedastic S-nonnegative garrote estimator for* $\boldsymbol{\gamma}$ *is given by* $\mathrm{IF}(P_0, \boldsymbol{\gamma}^{\text{he-SNNG}}, F) = (\mathrm{IF}_1(P_0, \boldsymbol{\gamma}^{\text{he-SNNG}}, F), \ldots, \mathrm{IF}_p(P_0, \boldsymbol{\gamma}^{\text{he-SNNG}}, F))^{\mathrm{T}}$, *with*

$$
\begin{aligned}
&\mathrm{IF}_j(P_0, \boldsymbol{\gamma}^{\text{he-SNNG}}, F) \\
&= \begin{cases}
0 & \text{if } j \in \mathcal{N}_{\lambda_2}, \\
\Pi_{\boldsymbol{\gamma} j} \Big\{ -\lambda_2 \left(\boldsymbol{\gamma}^{\text{init}}(F)\right)^{-1} + \lambda_2 \mathrm{diag}\left(\boldsymbol{\gamma}^{\text{init}}(F)\right)^{-2} \mathrm{IF}(P_0, \gamma^{\text{init}}, F) \\
\quad + \left(\mathbf{B}_{\mathcal{S}_{\lambda_1}} - \dfrac{1}{\mu_1} \mathbf{c}_F \mathbf{a}_{F\mathcal{S}_{\lambda_1}}^{\mathrm{T}} \right) \Pi_{\boldsymbol{\beta}\mathcal{S}_{\lambda_1}\mathcal{S}_{\lambda_1}} \left[\frac{1}{\mu_1^2} \mathbf{b}_{F\mathcal{S}_{\lambda_1}} w(\mathbf{X}_{0-}) \left(\rho\left(u_0\right) - b\right) \right. \\
\quad - \lambda_1 \mathrm{diag}\left(\beta_{\mathcal{S}_{\lambda_1}}^{\text{init}}(F)\right)^{-2} \mathrm{IF}(P_0, \beta_{\mathcal{S}_{\lambda_1}}^{\text{init}}, F) \\
\quad \left. - \dfrac{1}{\mu_1} w(\mathbf{X}_{0-}) \psi(u_0) \dfrac{X_{0\mathcal{S}_{\lambda_1}}}{\sqrt{h\left(\mathbf{X}_{0-}^{\mathrm{T}} \boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}} \right] \\
\quad - \dfrac{\sigma^{\text{he-SNNG}}(F)}{\mu_1^2} \left(\mathbf{B}_{\mathcal{S}_{\lambda_1}} - \dfrac{1}{\mu_1} \mathbf{c}_F \mathbf{a}_{F\mathcal{S}_{\lambda_1}}^{\mathrm{T}} \right) \Pi_{\boldsymbol{\beta}\mathcal{S}_{\lambda_1}\mathcal{S}_{\lambda_1}} \mathbf{a}_{F\mathcal{S}_{\lambda_1}} w(\mathbf{X}_{0-}) \\
\quad \left[\left(\dfrac{\nu_2}{\mu_1} + 1 \right) \left(\rho\left(u_0\right) - b\right) - \psi(u_0) u_0 \right] \\
\quad + \left(-\dfrac{1}{\mu_1} \mathbf{c}_F + \dfrac{h'\left(\mathbf{X}_{0-}^{\mathrm{T}} \boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}{h\left(\mathbf{X}_{0-}^{\mathrm{T}} \boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)} X_{0-} \right) w(\mathbf{X}_{0-}) \left(\rho\left(u_0\right) - b\right) \Big\} & \text{otherwise,}
\end{cases}
\end{aligned}
$$

for $j = 1, \ldots, p$, *where* $(\boldsymbol{\gamma}^{\text{init}}(F))^{-1}$ *is a vector with the jth element equal to* $(\gamma_j^{\text{init}}(F))^{-1}$, $\Pi_{\boldsymbol{\gamma} j}$ *denotes the jth row of*

$$
\begin{aligned}
\Pi_{\boldsymbol{\gamma}} = &\left(\frac{1}{2\mu_1} \left(\mathbf{B}_{\mathcal{S}_{\lambda_1}} - \frac{1}{\mu_1} \mathbf{c}_F \mathbf{a}_{F\mathcal{S}_{\lambda_1}}^{\mathrm{T}} \right) \Pi_{\boldsymbol{\beta}\mathcal{S}_{\lambda_1}\mathcal{S}_{\lambda_1}} \left(\frac{1}{\mu_1} \mathbf{b}_{F\mathcal{S}_{\lambda_1}} \mathbf{c}_F^{\mathrm{T}} - \mathbf{E}_{\mathcal{S}_{\lambda_1}} \right) \right. \\
&+ \frac{\sigma^{\text{he-SNNG}}(F)}{2\mu_1} \left(\mathbf{B}_{\mathcal{S}_{\lambda_1}} - \frac{1}{\mu_1} \mathbf{c}_F \mathbf{a}_{F\mathcal{S}_{\lambda_1}}^{\mathrm{T}} \right) \Pi_{\boldsymbol{\beta}\mathcal{S}_{\lambda_1}\mathcal{S}_{\lambda_1}} \left(\frac{\nu_2}{\mu_1^2} \mathbf{a}_{F\mathcal{S}_{\lambda_1}} \mathbf{c}_F^{\mathrm{T}} - \mathbf{B}_{\mathcal{S}_{\lambda_1}}^{\mathrm{T}} + \frac{1}{\mu_1} \mathbf{a}_{F\mathcal{S}_{\lambda_1}} \mathbf{d}_F^{\mathrm{T}} \right) \\
&\left. - \frac{1}{2\mu_1} \mathbf{c}_F \mathbf{c}_F^{\mathrm{T}} + \frac{1}{2} \mathbf{C} - \mathbf{D} \right)^{-1},
\end{aligned}
$$

with

$$
\Pi_{\boldsymbol{\beta}} = \left(\frac{1}{\mu_1} \mathbf{A} - \frac{1}{\mu_1^2} \mathbf{b}_F \mathbf{a}_F^{\mathrm{T}} - \frac{1}{\mu_1^2} \mathbf{a}_F \mathbf{b}_F^{\mathrm{T}} + \frac{\nu_2}{\mu_1^3} \sigma^{\text{he-SNNG}}(F) \mathbf{a}_F \mathbf{a}_F^{\mathrm{T}} \right)^{-1}.
$$

The influence function of the functional form of the heteroscedastic S-nonnegative garrote estimator for $\boldsymbol{\beta}$ *is given by* $\mathrm{IF}(P_0, \boldsymbol{\beta}^{\text{he-SNNG}}, F) = (\mathrm{IF}_0(P_0, \boldsymbol{\beta}^{\text{he-SNNG}}, F), \ldots, \mathrm{IF}_p(P_0, \boldsymbol{\beta}^{\text{he-SNNG}}, F))^{\mathrm{T}}$, *with*

$$
\begin{aligned}
&\mathrm{IF}_j(P_0, \boldsymbol{\beta}^{\text{he-SNNG}}, F) \\
&= \begin{cases}
0 & \text{if } j \in \mathcal{N}_{\lambda_1}, \\
\Pi_{\boldsymbol{\beta} j} \Big[\lambda_1 \mathrm{diag}\left(\beta^{\text{init}}(F)\right)^{-2} \mathrm{IF}(P_0, \beta^{\text{init}}, F) - \dfrac{1}{2\mu_1} \mathbf{E} \mathrm{IF}(P_0, \boldsymbol{\gamma}^{\text{he-SNNG}}, F) \\
\quad + \dfrac{1}{2\mu_1^2} \mathbf{b}_F \left(-2 w(\mathbf{X}_{0-}) \left(\rho\left(u_0\right) - b\right) + \mathbf{c}_F^{\mathrm{T}} \mathrm{IF}(P_0, \boldsymbol{\gamma}^{\text{he-SNNG}}, F) \right) \\
\quad - \dfrac{1}{2\mu_1} \sigma^{\text{he-SNNG}}(F) \left(\mathbf{B}^{\mathrm{T}} - \dfrac{\nu_2}{\mu_1^2} \mathbf{a}_F \mathbf{c}_F^{\mathrm{T}} - \dfrac{1}{\mu_1} \mathbf{a}_F \mathbf{d}_F^{\mathrm{T}} \right) \mathrm{IF}(P_0, \boldsymbol{\gamma}^{\text{he-SNNG}}, F) \\
\quad + \dfrac{\sigma^{\text{he-SNNG}}(F)}{\mu_1^2} \mathbf{a}_F w(\mathbf{X}_{0-}) \left[\left(\dfrac{\nu_2}{\mu_1} + 1 \right) \left(\rho\left(u_0\right) - b\right) - \psi(u_0) u_0 \right] \\
\quad + \dfrac{1}{\mu_1} w(\mathbf{X}_{0-}) \psi(u_0) \dfrac{\mathbf{X}_0}{\sqrt{h\left(\mathbf{X}_{0-}^{\mathrm{T}} \boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}} & \text{otherwise,}
\end{cases}
\end{aligned}
$$

for $j = 0, \ldots, p$, and the influence function of the functional form of the heteroscedastic S-nonnegative garrote estimator for σ is given by

$$
\begin{aligned}
&\mathrm{IF}(P_0, \sigma^{\text{he-SNNG}}, F) \\
&\quad = \frac{\sigma^{\text{he-SNNG}}(F)}{\mu_1}\left[w(\mathbf{X}_{0-})\,(\rho\,(u_0) - b) - \frac{1}{2}\mathbf{c}_F^{\mathrm{T}}\mathrm{IF}(P_0, \boldsymbol{\gamma}^{\text{he-SNNG}}, F)\right] \\
&\qquad + \frac{\sigma^{\text{he-SNNG}}(F)}{\mu_1^3}\mathbf{a}_{F\mathcal{S}_{\lambda_1}}^{\mathrm{T}}\,\Pi_{\boldsymbol{\beta}\mathcal{S}_{\lambda_1}\mathcal{S}_{\lambda_1}}\mathbf{b}_{F\mathcal{S}_{\lambda_1}}\left(w(\mathbf{X}_{0-})\,(\rho\,(u_0) - b) - \frac{1}{2}\mathbf{c}_F^{\mathrm{T}}\mathrm{IF}(P_0, \boldsymbol{\gamma}^{\text{he-SNNG}}, F)\right) \\
&\qquad + \frac{\sigma^{\text{he-SNNG}}(F)}{2\mu_1^2}\mathbf{a}_{F\mathcal{S}_{\lambda_1}}^{\mathrm{T}}\,\Pi_{\boldsymbol{\beta}\mathcal{S}_{\lambda_1}\mathcal{S}_{\lambda_1}}\left[\left(\mathbf{E}_{\mathcal{S}_{\lambda_1}} + \mathbf{B}_{\mathcal{S}_{\lambda_1}}^{\mathrm{T}}\right)\mathrm{IF}(P_0, \boldsymbol{\gamma}^{\text{he-SNNG}}, F)\right. \\
&\qquad \left. - \,2w(\mathbf{X}_{0-})\psi(u_0)\frac{X_{0\mathcal{S}_{\lambda_1}}}{\sqrt{h\left(\mathbf{X}_{0-}^{\mathrm{T}}\boldsymbol{\gamma}^{\text{he-SNNG}}(F)\right)}}\right] \\
&\qquad - \lambda_1\frac{\sigma^{\text{he-SNNG}}(F)}{\mu_1}\mathbf{a}_{F\mathcal{S}_{\lambda_1}}^{\mathrm{T}}\,\Pi_{\boldsymbol{\beta}\mathcal{S}_{\lambda_1}\mathcal{S}_{\lambda_1}}\mathrm{diag}\left(\beta_{\mathcal{S}_{\lambda_1}}^{\text{init}}(F)\right)^{-2}\mathrm{IF}(P_0, \beta_{\mathcal{S}_{\lambda_1}}^{\text{init}}, F) \\
&\qquad + \frac{\left(\sigma^{\text{he-SNNG}}(F)\right)^2}{2\mu_1^3}\mathbf{a}_{F\mathcal{S}_{\lambda_1}}^{\mathrm{T}}\,\Pi_{\boldsymbol{\beta}\mathcal{S}_{\lambda_1}\mathcal{S}_{\lambda_1}}\mathbf{a}_{F\mathcal{S}_{\lambda_1}}\left[\left(\frac{\nu_2}{\mu_1}\mathbf{c}_F^{\mathrm{T}} - \mathbf{d}_F^{\mathrm{T}}\right)\mathrm{IF}(P_0, \boldsymbol{\gamma}^{\text{he-SNNG}}, F)\right. \\
&\qquad \left. -2w(\mathbf{X}_{0-})\left(\left(\frac{\nu_2}{\mu_1} + 1\right)(\rho\,(u_0) - b) - \psi(u_0)u_0\right)\right].
\end{aligned}
$$