# Automatic bias correction for testing in high dimensional linear models

**Jing Zhou and Gerda Claeskens**

ORStat and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium
jing.zhou@kuleuven.be; gerda.claeskens@kuleuven.be

**Abstract**

Hypothesis testing is challenging due to the test statistic's complicated asymptotic distribution when it is based on a regularised estimator in high dimensions. We propose a robust testing framework for $\ell_1$-regularised M-estimators to cope with non-Gaussian distributed regression errors, using the robust approximate message passing algorithm. The proposed framework enjoys an automatically built-in bias correction and is applicable with general convex nondifferentiable loss functions which also allows inference when the focus is a conditional quantile instead of the mean of the response. The estimator compares numerically well with the debiased and desparsified approaches while using the least squares loss function. Use of Huber's loss function demonstrates that the proposed construction provides stable confidence intervals under different regression error distributions.

Keywords: Approximate message passing algorithm, confidence interval, high-dimensional linear model, hypothesis testing, $\ell_1$-regularisation, loss function.

## 1 Introduction

In a sparse high dimensional linear model of the form $Y = X\beta + \varepsilon$, we wish to perform hypothesis testing and construct confidence intervals for components of the $p$-vector $\beta$ when $p$ grows with the sample $n$ such that $n/p \to \delta \in (0, 1)$. Rather than using $\ell_1$-regularised estimation and desparsifying or debiasing such estimators, we use the approximate message passing algorithm to take the selection uncertainty into account. This means that we consider the full vector of estimated coefficients and do not theoretically restrict to an assumed perfectly selected subset of nonzero coefficients.

Our main contribution is a general framework of testing, using $\ell_1$-regularised M-estimators, which allows to incorporate a broad group of convex loss functions, e.g., (i) least squares (LS) loss: $\rho(z) = z^2$, (ii) Huber loss with $\rho_u(z) = z^2/2$ if $|z| \leq u$ and $\rho_u(z) = u|z| - u^2/2$ if $|z| > u$, (iii) least absolute deviation (LAD) loss with $\rho(z) = |z|$ and (iv) quantile loss with $\rho_\tau(z) = z(\tau - I\{z \leq 0\})$. In particular, loss functions that provide robustness in case of outliers in the regression errors are included. However, an investigation about how to use the approximate message passing algorithm for the detection of outliers in the high-dimensional predictive variables, remains a future research topic. This general framework distinguishes our approach from the papers based on bias correction for $\ell_1$-regularized estimators that deal with a specific loss function or that impose differentiability assumptions on the loss functions (van de Geer et al., 2014). When taking the least squares loss function, our approach does not require to use the Karush-Kuhn-Tucker characterisation of the Lasso estimators (Javanmard and Montanari, 2014; van de Geer et al., 2014).

Instead, our method relies on an asymptotically normal distributed estimator obtained from the robust approximate message passing (RAMP) algorithm (Donoho et al., 2009; Bayati and Montanari, 2011a; Bradic, 2016; Donoho and Montanari, 2016; Zhou et al., 2020). The estimator in the last step of the RAMP algorithm is obtained by applying a soft-thresholding function to the estimator that is our main focus. This estimator from the robust approximate message passing algorithm does not require additional computations and enjoys by construction the debiasing properties that have earlier been studied for $\ell_1$-regularised estimators. The RAMP algorithm assumes the covariates $X_{\cdot 1}, \ldots, X_{\cdot p}$ to be independent Gaussian, which is inherited from the AMP framework (see for example Donoho et al., 2009; Bayati and Montanari, 2011a). This assumption is also compulsory in the convex Gaussian mini-max theorem (Gordon, 1985, 1988; Thrampoulidis et al., 2018) in compressed sensing. The two mentioned approaches provide insight into the asymptotic behaviour of the estimator $\widehat{\beta}$ under the i.i.d. assumption. In this paper, we do not deviate far from the AMP framework regarding the assumptions. Instead, the objective is to provide a simple computational tool for robust testing built on Bradic (2016); Zhou et al. (2020). To broaden the scope we incorporate a decorrelation step as in Wang et al. (2016) to adapt the proposed testing procedure to correlated Gaussian designs.

Our simulations show that this estimator compares well to the debiased (Javanmard and Montanari, 2014) and desparsified estimators (van de Geer et al., 2014) in case the least squares loss is used. For other loss functions our method applies in an equal fashion. We illustrate it for quantile loss and for the Huber loss function.

It is important to point out that this method provides inference for the *full* vector of coefficients as opposed to only working with the selected components. In this sense our results are not comparable to those obtained by selective inference (e.g., Lee et al., 2016) where a conditioning on the event of selection takes place and inference is restricted to only the coefficients appearing in the selected model after regularisation. Our target of inference is different since we are interested in results for the full vector, zeros included. This work also differs substantially from the earlier results on regularised estimation (e.g., Zou, 2006) which discusses the properties of the null and nonnull subvectors separately by selection consistency of the null subvector and the asymptotic normality of the nonnull subvector. The RAMP approach includes the selection uncertainty via its specific algorithm.

More details about the estimator and the RAMP algorithm is given in Section 2. Its use for hypothesis testing and the construction of confidence intervals is contained in Section 3. Simulation results showing its advantageous behaviour are in Section 4 and for a data analysis see Section 5. Section 6 concludes.

# 2 Notation, assumptions and estimators

We wish to estimate the parameter vector $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ in the model $Y = X\beta + \boldsymbol{\varepsilon}$ with a known design matrix $X \in \mathbb{R}^{n\times p}$ and a response vector $Y \in \mathbb{R}^n$. We denote the rows of $X$ by $X_{i\cdot}$ corresponding to $n$ independent samples, $i = 1, \ldots, n$; and denote the columns of $X$ by $X_{\cdot j}$ representing $p$ predictive variables, $j = 1, \ldots, p$. The covariates with non-zero coefficients are 'relevant' to the response vector. In addition, we assume $p > n$ and denote the number of non-zero coefficients of $\beta$ by $s$.

## 2.1 Regularised estimators with general loss functions

The parameter vector $\beta$ is often estimated by solving a minimisation problem combining a loss function $\rho(\cdot)$ and an $\ell_1$-regulariser with parameter $\lambda$ as follows

$$\widehat{\beta}(\lambda) = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n} \rho(Y_i - X_{i\cdot}^\top \beta) + \lambda \|\beta\|_1 \right\}. \tag{1}$$

The most straightforward approach to estimate $\widehat{\beta}(\lambda)$ is by solving the right-hand-side of (1). Intensive research has been devoted to this approach with different loss functions $\rho$, see Tibshirani (1996); Donoho and Johnstone (1994); Donoho (1995) with the LS loss function; Wang et al. (2007) with the LAD loss function; Belloni and Chernozhukov (2011) with the quantile loss, etc.

Due to the regularisation, the asymptotic distribution of an estimator of $\beta$ in high dimensions becomes complicated and is no longer Gaussian due to the presence of many exact zeros for the unselected components. One popular approach to state asymptotic theory for the regularised estimators is using the 'oracle property', which states the selection consistency of the subvector of zeros, the non-selected components, and the asymptotic normality of the estimator of the truly nonzero components of $\beta$, see Zou (2006); Fan and Li (2001); Sun et al. (2020); Bradic et al. (2011). Although the 'oracle property' provides promising asymptotic guarantees, a perfect selection is hard to achieve for finite samples. In addition, to achieve selection consistency, a so-called 'beta-min' assumption requiring that the magnitude of $\beta$ is sufficiently large, is often made in the literature (e.g. Bühlmann and Van De Geer, 2011).

Subsequent inference following selection attracted attention. One way to achieve valid inference is via sample splitting in which the estimators are obtained from a sample independent of the sample that is used for inference. Wasserman and Roeder (2009) split the sample in three parts: a first one-third of the data is used for $\ell_1$-regularised estimation for a grid of regularization values; a second one-third of the data determines a suitable regularisation parameter via cross-validation and the third part of the data uses a least squares estimator for the selected set of variables as determined by the estimation on the first one-third of the data using the cross-validated regularisation found from the second data part. Independence between the different parts of the dataset guarantees valid inference. Rather than a single time splitting the dataset, Meinshausen et al. (2009) split the sample into two pieces multiple times, and perform aggregated inference using the second subset based on the selected variables by the first subset. With a large number of such random splits of the data, they show that asymptotically there is a control of the familywise error rate and the false discovery rate. For datasets of smaller size, sample splitting has not been attractive. The method in this paper does not need to split the sample in order to achieve valid inference.

Alternatively, one can approximate $\widehat{\beta}(\lambda)$ via the approximate message passing (AMP) algorithm (Donoho et al., 2009; Bayati and Montanari, 2011a; Donoho and Montanari, 2016); this AMP algorithm and its generalisation is a crucial tool throughout this paper. The importance is that an asymptotic representation of the mean squared error (MSE) of the regularised estimator (the full vector) can be obtained, thus effectively taking the selection effects into account. This is in contrast with so-called oracle approaches that assume perfect selection. Further, the mean squared difference of the AMP approximation and $\widehat{\beta}(\lambda)$ converges to 0 almost surely (Bayati and Montanari, 2011b, Theorem 1.8) when $p \to \infty$ and similar results are obtained for other loss functions (Bradic, 2016; Zhou et al., 2020).

Studies of the asymptotic distribution of the full vector $\beta$ in high dimensions started with the settings where $p/n \leq 1$ and without regularisation, see El Karoui et al. (2013); Lei et al. (2018); Donoho and Montanari (2016); El Karoui (2013) studied $\ell_2$-regularised M-estimators for $p/n < \infty$ and showed that the M-estimators without regularisation can be seen as a limiting case letting the tuning parameter of the $\ell_2$-regularisation be 0, while assuming additionally that the loss function is strongly convex.

While returning to settings where $p/n > 1$ and regularisation is a prevailing solution for estimating $\beta$, some alternative options of constructing confidence intervals and hypothesis testing have been carefully investigated on the Lasso estimator in (1) which is obtained by choosing $\rho(\cdot)$ to be the least squares loss. Instead of focusing directly on the regularised estimators with complex asymptotic distribution, the desparsifying (van de Geer et al., 2014) and debiasing (Javanmard and Montanari, 2014) approaches construct estimators based on the Lasso estimators under an i.i.d. Gaussian assumption on the regression error, $\varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$. The desparsification rewrites the Karush–Kuhn–Tucker (KKT) condition satisfied by the Lasso estimator, while the debiasing estimator is obtained by adding a term compensating the bias introduced by the $\ell_1$-regulariser. Although derived through different approaches under slightly different assumptions, the estimators of interest for inference in van de Geer et al. (2014); Javanmard and Montanari (2014) follow identical expressions and are denoted by $\widehat{\beta}_{\text{U}}$. With $M$ being an approximation of the inverse of the sample covariance matrix $\widehat{\Sigma} = X^\top X/n$, the estimator $\widehat{\beta}_{\text{U}}$ is defined as

$$\widehat{\beta}_{\text{U}} = \widehat{\beta}_{\text{LS}}(\lambda) + MX^\top(Y - X\widehat{\beta}_{\text{LS}}(\lambda))/n. \tag{2}$$

However, the methods use different approaches to approximate the matrix $M$. More specifically, Javanmard and Montanari (2014) proposed an algorithm to approximate a sparse matrix $M$; and at the same time the algorithm controls the non-Gaussianity, bias and variance of the estimator $\widehat{\beta}_{\mathrm{U}}$. On the contrary, van de Geer et al. (2014) only requires a suitable approximation of $M$, which is estimated by a nodewise Lasso running $p$ times on each $X_{\cdot j}, j = 1, \ldots, p$ as the response variable in a Gaussian regression model with all $p-1$ other covariates, thus except for $X_{\cdot j}$, in the design matrix. Consequently, inference based on the estimator $\widehat{\beta}_{\mathrm{U}}$ uses the asymptotic normality

$$\sqrt{n}(\widehat{\beta}_{\mathrm{U}} - \beta) = W + o_P(1), \quad W|X \sim N(0, \sigma_\varepsilon^2 M\widehat{\Sigma}M^\top). \tag{3}$$

Extensions of this approach include Caner and Kock (2018) on the conservative desparsified Lasso and Gueuning and Claeskens (2018) on the focused information criterion based on the desparsified Lasso for a high-dimensional linear model.

Since the least squares loss function is sensitive to non-Gaussian distributed errors, robust loss functions that deal with outliers in regression errors and error distributions with heavy-tails are preferred, among which a popular choice is the quantile loss function (Koenker and Bassett, 1978; Koenker, 2005). Debiasing the regularised quantile estimator was investigated in Zhao et al. (2014, 2019); Bradic and Kolar (2017). Simultaneous confidence intervals are constructed by using a Gaussian multiplier bootstrap, see also Zhang and Cheng (2017); Dezeure et al. (2017) for bootstrapping the debiased Lasso. The high dimensional rank score was developed to estimate the sparsity function in Bradic and Kolar (2017) while uniform confidence bands were constructed based on a Bahadur representation of the debiased estimator. Our method differs from the literature mentioned above since we do not develop theory applicable for only a specific loss function, and further, our construction does not rely on bootstrapping, which can be computationally intensive. By our construction, a relatively convenient switch between loss functions can be realised without a heavy computational burden. However, our proposed method can only cope with outliers in the regression errors; robust inference when outliers exist in the predictive variables is beyond the scope of this paper.

## 2.2 Assumptions

We assume the design matrix $X \in \mathbb{R}^{n\times p}$, the error vector $\boldsymbol{\varepsilon}$, and the coefficient vector $\beta$ to satisfy Assumptions (A1)-(A5) from Zhou et al. (2020, Appendix A), which we repeat here for completeness.

(A1) A standard Gaussian design: for $i = 1, \ldots, p$ and $j = 1, \ldots, n$, the $X_{ij} \sim N(0, 1/n)$ are independent and identically distributed.

(A2) For the $p$-vector $\beta$ it holds that for $p$ tending to infinity a sequence of uniform distributions that is placed on its components converges to a distribution with a bounded $(2k-2)$th moment for $k \geq 2$. We denote by $B_0$ a random variable with this limiting distribution function $F_{B_0}$.

(A3) Loss function $\rho$ such that: (i) the subgradient $\partial\rho(x) = \sum_{j=1}^{3} v_j(x)$ such that $v_1$ has an absolutely continuous derivative, $v_2$ is continuous, consisting of piecewise linear parts and is constant outside a bounded interval, and $v_3$ is a non-decreasing step function. Define $v_2'(u) = \alpha_l$ and $v_3(u) = \gamma_l$ when $u \in (r_l, r_{l+1}]$ where $\alpha_0 = \alpha_L = 0$, $-\infty = r_0 < r_1 < \ldots < r_L < r_{L+1} = \infty$ and $-\infty = \gamma_0 < \gamma_1 < \ldots < \gamma_L < \gamma_{L+1} = \infty$. (ii) $|\partial\rho(u)|$ is bounded for all $u \in \mathbb{R}$. (iii) $\int \rho(z - t)dF_\varepsilon(z)$ has a unique minimum at $t = 0$. (iv) There exists a value $\delta > 0$ and $\eta > 1$ such that $E[\{\sup_{|u|\leq\delta} |v_1''(z + u)|\}^\eta]$ is finite.

(A4) For some $\kappa > 1$, (i) $\lim_{p\to\infty} E_{\widehat{f}_\beta}(B_0^{2\kappa-2}) = E_{f_{B_0}}(B_0^{2\kappa-2}) < \infty$; (ii) $\lim_{p\to\infty} E_{\widehat{f}_\varepsilon}(\varepsilon^{2\kappa-2}) = E_{f_\varepsilon}(\varepsilon^{2\kappa-2}) < \infty$; (iii) $\lim_{p\to\infty} E_{\widehat{f}_{q_0}}(B_0^{2\kappa-2}) < \infty$.

(A5) $\varepsilon_1, \ldots, \varepsilon_n$ and $\varepsilon$ are i.i.d. random variables with mean zero, a finite 2nd moment, cumulative distribution function $F_\varepsilon$ and probability density function $f_\varepsilon$. The $F_\varepsilon$ has bounded derivatives $f_\varepsilon$ and $\partial f_\varepsilon$ and $f_\varepsilon > 0$ in the neighbourhood of $r_1, \ldots, r_L$ in (A3).

In addition, we denote the set of indices of the non-zero components of $\beta$ by $S$ consisting of $s$ elements. Its complement is denoted $S^{\mathrm{c}} = \{1, \ldots, p\} \backslash S$ with size $p - s$. We assume that the ratios $n/p \to \delta \in (0, 1)$, $n/s \to a \in (1, \infty)$, $s/p \to \omega = P(B_0 \neq 0)$ when $n, p, s \to \infty$.

Assumption (A1) is a standard assumption on the design matrix used in Bayati and Montanari (2011a); Donoho et al. (2009); Donoho and Montanari (2016); Bradic (2016). Assumption (A1) is implicit but critical in the construction. In the limit the $X_{ij}$'s are represented by a $N(0, 1)$ distributed random variable $Z$ (see for example in Eq.(3.46) Bayati and Montanari, 2011a). The variable $Z$ determines the asymptotic normality of $\widetilde{\beta}$, see the discussion in Section 3.1. Assumption (A2) defines a random variable $B_0$ with distribution $F_{B_0}$ to which the sequence $\beta_j, j = 1, \ldots, p$ converges as $n$ tends to infinity by assigning $1/p$ point mass to each component of the vector $\beta$. For (A2), see also Bayati and Montanari (2011a); Bradic (2016). Assumption (A3) is first stated in Bradic (2016) to incorporate convex possibly non-differentiable loss functions by using the subgradient of the loss function instead of gradient. In general, Assumption (A3) states that the RAMP algorithm allows for loss functions of which the subgradient can be decomposed into combinations of three types of functions on the intervals as specified in Assumption (A3): (1) functions with absolutely continuous first derivative; (2) piecewise continuous linear functions; (3) non-decreasing step functions. This assumption is fundamental for deriving a consistent estimator in the RAMP algorithm (Bradic, 2016, Lemma 3). Further, one can perceive this as a generalized score function, which in the robust statistics field is used to further study the properties of M-estimators, e.g. Fisher consistency, the influence function, etc. Examples of loss functions that satisfy (A3) include the Huber loss function (Huber, 2004) which is equivalent to Winsorizing the residuals, quantile loss, absolute value and squared loss functions, etc. Assumption (A4) is a technical assumption for proving almost sure convergence of a general function of $\widehat{\beta}$, see Theorem 2 in Bayati and Montanari (2011a). Assumption (A5) states that the components of $\boldsymbol{\varepsilon}$ are i.i.d. with cumulative distribution function $F_\varepsilon$ and density function $f_\varepsilon$, which is used to determine the intervals in Assumption (A3). Assumption (A4) has been used in Bayati and Montanari (2011a) and Zhou et al. (2020, Lemma 1) there taking $\kappa = 2$.

### 2.3 The estimator from the robust approximate message passing algorithm

The robust approximate message passing (RAMP) algorithm, see Huang (2020); Bradic (2016) and Zhou et al. (2020, Section 3.2), is an iterative procedure consisting of three steps. The iteration number is denoted by $t = 1, 2, \ldots$. The estimator $\widehat{\beta}_{(t)}$ is updated in each iteration $t$ and is denoted by $\widehat{\beta}$ at convergence. The difference between the estimator $\widehat{\beta}$ from the RAMP algorithm at convergence and the corresponding estimator from regularisation in (1) converges to zero in $\ell_2$-norm with probability one. This has been shown in Huang (2020, Theorem 2.2.) for the generalised AMP algorithm with nonnegative convex loss function, and in Bayati and Montanari (2011b, Theorem 1.8) for the LS loss. This convergence in $\ell_2$-norm ensures the validity of using the RAMP algorithm to approximate the minimizer of (1).

We restate here the part of the RAMP algorithm which is directly linked to the estimators to be used in this paper. At iteration step $t$ of the algorithm, the estimator $\widehat{\beta}_{(t+1)}$ corresponding to the $\ell_1$-regularised M-estimator is updated as follows,

$$\widehat{\beta}_{(t+1)} = \eta(\widetilde{\beta}_{(t)}; \theta_{(t)}), \text{ where } \widetilde{\beta}_{(t)} = \widehat{\beta}_{(t)} + X^\top G(z_{(t)}; b_{(t)}). \tag{4}$$

The soft-thresholding function

$$\eta(x; \theta) = \operatorname{sign}(x) \cdot \max(|x| - \theta, 0) \tag{5}$$

in (4) is applied componentwise to the estimator $\widetilde{\beta}_{(t)}$ of major interest. This estimator asymptotically follows a Gaussian distribution centering at the true regression coefficient vector $\beta$, see Section 3, and it is linked to the debiased/desparsified estimator, see (2) as an example using the LS loss function. The soft-thresholding function incorporates the $\ell_1$-regulariser $\|\beta\|_1$. The tuning parameter $\theta_{(t)}$ is updated in Step 2 in the RAMP algorithm, which is linked to the tuning parameter $\lambda$, for the details see Zhou et al. (2020, Eq. (2.21)). The rescaled effective score function $G$ is determined by the convex loss function $\rho$ and is defined in Zhou et al. (2020, Eq. (11)). Step 3 in the RAMP algorithm is given by (4), which describes the connection between the estimators $\widehat{\beta}_{(t)}$ and $\widetilde{\beta}_{(t)}$. The argument $z_{(t)}$ in the rescaled effective score function is updated in Step 1 in the RAMP algorithm, and the parameter $b_{(t)}$ is updated in Step 2. Details of the complete RAMP algorithm can be found in Zhou et al. (2020, Section 3.2) and for a brief overview, see Appendix A.

The properties of the RAMP estimator rely on the independent Gaussian design Assumption (A1) and violations may cause convergence problems. As a remedy, we incorporate a decorrelation step as proposed in Wang et al. (2016, Eq.(3)). We briefly state the idea of the decorrelation. We perform a singular value decomposition of the design matrix $X = UDV^\top$. By noticing that $\sqrt{p}UD^{-1}U^\top = \{(XX^\top/p)^-\}^{1/2}$ where $(\cdot)^-$ denotes the Moore-Penrose pseudo-inverse, the linear model can be rewritten as

$$\{(XX^\top/p)^-\}^{1/2}Y = \sqrt{p}UV^\top\beta + \{(XX^\top/p)^-\}^{1/2}\boldsymbol{\varepsilon}. \tag{6}$$

Then, the new response is $\widetilde{Y} = \{(XX^\top/p)^-\}^{1/2}Y$ and the new design matrix is $\widetilde{X} = \sqrt{p}UV^\top$. The RAMP algorithm is now applied to $\widetilde{Y}, \widetilde{X}$.

# 3 Confidence intervals and hypothesis testing

## 3.1 Componentwise inference

For a component $\beta_j$ with $j \in \{1, \dots, p\}$, we are interested in testing the null hypothesis

$$H_{0,j} : \beta_j = \beta_{0,j} \text{ versus } H_{a,j} : \beta_j \neq \beta_{0,j}.$$

Rather than considering the sequence $\widehat{\beta}_{(t)}, t = 1, 2, \dots$ as in the previous literature, the main focus of this project is the sequence $\widetilde{\beta}_{(t)}$ in (4), which was proven in Zhou et al. (2020, Section 5.2, before Corollary 2), to converge weakly to $B_0 + \bar{\zeta}_{(t)}Z_{(t)}$, $n, p \to \infty$, where $B_0$ is defined in Assumption (A2), $Z_{(t)}$ is a standard normally distributed variable independent of the data, $\bar{\zeta}_{(t)}$ is the square-root of the state evolution parameter (see Step 2 in Appendix A), and is estimated by $\bar{\zeta}^2_{\text{emp},(t)} = n^{-1}\sum_{i=1}^n G(z_{i,(t)}; b_{(t)})^2$, which is via (4) directly obtainable from the RAMP algorithm. Let $\widetilde{\beta}$ and $\bar{\zeta}$ denote the estimator $\widetilde{\beta}_{(t)}$ and the parameter $\bar{\zeta}_{(t)}$ from the RAMP algorithm at convergence. The estimator $\widetilde{\beta}_{(t),j}$ is an asymptotically unbiased estimator for the true coefficient $\beta_{0,j}$ with common variance $\bar{\zeta}_{(t)}$ for each component with $j = 1, \dots, p$. Under the null hypothesis, it holds that when $p \to \infty$,

$$T_j(\beta_{0,j}) = \frac{\widetilde{\beta}_j - \beta_{0,j}}{\bar{\zeta}} \xrightarrow{d} N(0, 1). \tag{7}$$

Note that the scaling term $\sqrt{n}$ is incorporated in the state evolution parameter $\bar{\zeta}$, see Bayati and Montanari (2011a, Proof of Lemma 1(b), e.g. the late almost sure convergence on p. 775). For fixed and large $n$ and $p$, the asymptotic standard normality holds in an approximate way, $T_j(\beta_j) \approx N(0, 1)$, $j = 1, \dots, p$. By replacing $\bar{\zeta}$ by $\zeta_{\text{emp}}$, the $p$-value of the test statistic can be calculated as

$$P_j = 2\{1 - \Phi(|\widetilde{\beta}_j - \beta_{0,j}|/\zeta_{\text{emp}})\}. \tag{8}$$

As usual, for a given significance level $\alpha$, the null hypothesis $H_{0,j}$ is rejected if $P_j \leq \alpha$. A confidence interval of $\beta_j$ with asymptotic confidence level $1-\alpha \in (0,1)$ can be constructed as

$$\widehat{\mathrm{CI}}_j(1-\alpha) = [\widetilde{\beta}_j - \Phi^{-1}(1-\alpha/2)\bar{\zeta}_{\mathrm{emp}}, \widetilde{\beta}_j + \Phi^{-1}(1-\alpha/2)\bar{\zeta}_{\mathrm{emp}}],$$

where $\Phi$ denotes the standard normal cumulative distribution function.

It is important to point out that the asymptotic normality holds for the full estimated vector $\widetilde{\beta}$, as opposed to only for the non-zero components under the 'perfect selection' ($\beta_{\min}$ assumption). By our construction or similar debiasing/desparsifying approaches, the selection uncertainty of the non-zero components is taken into account.

As a special case we take the least squares loss. In this case the estimator $\widetilde{\beta}_{(t)}$ is comparable to the desparsified estimator in (2) in the following aspects: 1. the estimator $\widehat{\beta}_U$ in (2) is obtained by adding to the Lasso estimator, a term containing a decorrelated design matrix $MX^\top/n$ and the residual $Y - X\widehat{\beta}_{\mathrm{LS}}(\lambda)$. The estimator of focus in this paper, $\widetilde{\beta}$ in (4), is constructed by adding to the estimator corresponding to the Lasso $\widehat{\beta}$, a term that contains a standard Gaussian design matrix $X^\top$ (Zhou et al., 2020, Assumption (A1)), and for least squares loss the adjusted residual $G$ is equivalent to $Y - X\widehat{\beta}_{\mathrm{LS}}(\lambda)$. 2. Both estimators $\widehat{\beta}_U$ and $\widetilde{\beta}$ are asymptotically Gaussian distributed.

While comparing our approach to the debiasing/desparsifying approaches, our construction does not require a numerical approximation of the inverse of the KKT characterisation of the Lasso estimator as in van de Geer et al. (2014) or the term that is proportional to the subgradient of the $\ell_1$-regulariser as in Javanmard and Montanari (2014), thus it is a convenient numerical construction. Further, unlike van de Geer et al. (2014); Javanmard and Montanari (2014) assuming i.i.d. Gaussian distributed errors $\boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2 I_n)$, the RAMP algorithm only imposes moment conditions on $\boldsymbol{\varepsilon}$ (Zhou et al., 2020, Assumption (A5)). The existing research investigating the asymptotic distribution of estimators in the debiasing/desparsifying framework, mostly either requires the loss functions to be differentiable, see van de Geer et al. (2014) requiring twice differentiability, or focuses on a prespecified loss function, see Zhao et al. (2014, 2019); Bradic and Kolar (2017) with quantile loss function; Javanmard and Montanari (2014); Zhang and Cheng (2017); Dezeure et al. (2017) with least squares loss function. On the contrary, our approach offers a general framework incorporating a large group of loss functions without imposing assumptions on differentiability. The construction in this paper is applicable to a broad class of $\ell_1$-regularised M-estimators.

### 3.2 Simultaneous inference

In practice, we are often interested in the simultaneous inference on a subset of the regression coefficients $\{\beta_j, j \in S_0\}$, where $S_0 \subseteq \{1, \ldots, p\}$ with cardinality $s_0 = |S_0|$. By Zhou et al. (2020, Assumption (A2)), the components of $\beta$ are independent samples of a random variable $B_0$. Let the subvector $\beta_{S_0}$ have an estimator $\widetilde{\beta}_{S_0}$. Then the vector $\widetilde{\beta}_{S_0}$ is approximately $N(\beta_{S_0}, \bar{\zeta}^2 I_{s_0})$ distributed where $I_{s_0}$ is an $(s_0 \times s_0)$-dimensional identity matrix. The confidence region of $\beta_{S_0}$ with confidence level $1-\alpha$ can be constructed as follows,

$$\widehat{\mathrm{CR}}_{S_0}(1-\alpha) = \Big\{\beta_{S_0} \in \mathbb{R}^{s_0} : (\widetilde{\beta}_{S_0} - \beta_{S_0})^\top I_{s_0} (\widetilde{\beta}_{S_0} - \beta_{S_0}) \leq \bar{\zeta}^2_{\mathrm{emp}} \cdot q\chi^2_{s_0}(1-\alpha)\Big\},$$

where $q\chi^2_{s_0}(1-\alpha)$ denotes the $(1-\alpha)$th quantile of a chi-squared distribution with $s_0$ degrees of freedom.

To simultaneously test multiple hypotheses $H_{0,j}$ for the subset of coefficients $\{\beta_j, j \in S_0\}$, where $S_0 \subseteq \{1, \ldots, p\}$ the error rate can be controlled by adjusting the $p$-values (e.g., Holm (1979); Šidák (1967); Hochberg (1988) procedure, etc.). Since we want to numerically compare our proposed approach with the desparsifying approach in van de Geer et al. (2014) and the debiasing approach in Javanmard and Montanari (2014) while taking the loss function $\rho(\cdot)$ to be the least squares loss, we follow the choice of van de Geer et al. (2014) and adjust the $p$-values by the Holm-Bonferroni procedure. For a family of hypotheses $\{H_{0,j}, j \in S_0\}$, $S_0 \subseteq \{1, \ldots, p\}$ with a nominal probability of a type I error $\alpha$, the adjustment for the $\ell_1$-regularised M-estimators is as follows: (1) Obtain the $p$-values $P_j$ for testing the individual hypothesis $H_{0,j}$ by (8); (2) Sort the $p$-values in ascending order and denote the sorted $p$-values by $P_{[j]}$; (3) Adjust the significance levels for the individual tests to $\alpha/(s_0 - [j] + 1)$; (4) Reject the null hypothesis $H_{0,[j]}$ if $P_{[j]} \leq \alpha/(s_0 - [j] + 1)$.

## 4 Simulation study

In this section, we investigate the finite sample performance of the confidence intervals and hypothesis tests for the $\ell_1$-regularised M-estimators. Since our approach allows for non-differentiable loss functions which are less sensitive to non-Gaussian distributed errors, we tested our approach on the quantile loss at quantile level 0.5, the least squares loss and the Huber loss function. For the LS loss we consider the $\ell_1$-regularised least squares estimators. To show that the proposed approach has a stable and competitive performance, we compare the usage of the estimator from the RAMP algorithm with competing alternative approaches, namely the desparsifying and the debiasing approaches from Javanmard and Montanari (2014) and van de Geer et al. (2014).

### 4.1 Data generating procedure

The simulation settings and data generating procedure are described as follows. We randomly generate a matrix $X$ and a coefficient vector $\beta$, which are used in all replications for the same simulation setting. We consider a high-sparsity setting with $s = 5$ and a medium sparsity setting with $s = 50$. The response vector $Y = X\beta + \boldsymbol{\varepsilon}$. We consider $R = 500$ replications for each setting in the simulation. We showcase three $\ell_1$-regularised estimators for different purposes: (1) the quantile estimator at quantile level 0.5 exhibits that the proposed method can construct confidence intervals when the conditional quantile is of main interest. (2) The least squares estimator aims at comparing our construction with the competitors (the debiasing and deparsifying approaches) in small sample settings ($\delta = 0.2$) in which parameter estimation is more challenging. According to Tables 1 and 2, using a Dirac distribution or $N(0, 1)$ to generate the nonnull components of the true parameter vector $\beta$ leads to similar conclusions. Comparing the performance of our construction with alternative approaches is only conducted for $\ell_1$-regularised least squares estimators due to three main reasons (i) In practice, the $\ell_1$-regularised least squares estimator is the most commonly used estimator for sparse high-dimensional linear models. Hence, we designed (2) to show that our proposed method has competitive performance compared to alternatives for the most frequently encountered $\ell_1$-regularised least squares estimator. (ii) limitation of available codes of debiased quantile estimators, and (iii) to our best knowledge, there is no literature on debiasing the Huber's loss function. (3) Huber's estimator in small sample settings demonstrates the proposed method can incorporate robust loss functions. The performance is compared with the least squares estimator to show that switching to robust loss functions is helpful when outliers exist in regression errors, and further, the proposed method provides a reasonably convenient switch between the loss functions. In addition, all three estimators are compared in Table 4, see Section 4.5.

These are the settings for the simulations.

(1) *$\ell_1$-regularised quantile estimator*. The components of the matrix $X$ are independent and generated from $N(0, 1/n)$. The quantile level is 0.5, thus we estimate the median. The subvector of $\beta$ consisting of non-zero components is generated from a Dirac distribution with point mass equally distributed on $-1$ and 1, or from a $N(0, 1)$. We choose $p = 500$, $n = 250$ and $\delta = 0.5$, and $n = 100$ with $\delta = 0.2$. The considered distributions for $\boldsymbol{\varepsilon}$ are the standard normal $N(0, 1)$, student-$t$ with 3 degrees of freedom, and the mixture of normal distributions $0.5N(0, 1) + 0.5N(5, 9)$. The errors are centered and rescaled to have standard deviation 0.2 after sampling.

(2) *$\ell_1$-regularised least squares estimator.* We choose $p = 500$, $n = 100$ with $\delta = 0.2$, which is the more challenging setting with the smaller sample size among the two sample sizes in (1). The nonzero components of the subvector of $\beta$ are generated from a Dirac distribution with point mass equally distributed on $-1$ and 1. The same three error distributions as in (1) are used and are rescaled to have standard deviation 0.2 after sampling. Notice that the approaches in van de Geer et al. (2014); Javanmard and Montanari (2014) only require the design matrix $X$ to follow a multivariate Gaussian distribution $N(0, \Sigma_X)$ with arbitrary covariance matrix $\Sigma_X$, which is less strict than assumption (A1). For a fair comparison, we consider two covariance matrix structures. An independent Gaussian design with $(\Sigma_X)_{i,j} = n^{-1}I\{i = j\}$ and a correlated Gaussian design with Toeplitz structure $(\Sigma_X)_{i,j} = 0.9^{|i-j|}, i = 1, \ldots, n, j = 1, \ldots, p$.

(3) *$\ell_1$-regularised Huber estimator.* Here, we present results for the independent Gaussian design with $(\Sigma_X)_{i,j} = n^{-1}I\{i = j\}$ from the settings in (2) since according to Table 3, the general Gaussian design provides conclusions that are similar to the independent Gaussian design. However, to evaluate the robustness of the proposed method, we consider the following two mixed normal distributed errors reflecting outlying observations (Alfons et al., 2013; Khan et al., 2007): (i) *(Leverage point)* $0.1N(9, 0.2) + 0.9N(-1, 2)$; (ii) *(Clustering)* $0.1N(18, 0.01) + 0.9N(-2, 0.2)$. The parameter of the Huber loss function is chosen to be 1.5 for setting (i) and 3 for setting (ii).

## 4.2 Evaluation measures

For the subvector of the full vector $\beta$ consisting of true zero components and for the complementary subvector consisting of true nonzero components empirical coverage probabilities are computed by averaging the coverage of the individual intervals $\widehat{\text{CI}}_{r,j}(1-\alpha)$ over all simulation runs $r = 1, \ldots, R$ and over all components of the vector with length denoted by $p_{\text{vec}}$.

$$\widehat{\text{CP}}_{\text{vec}}(1-\alpha) = (p_{\text{vec}}R)^{-1} \sum_{j=1}^{p_{\text{vec}}} \sum_{r=1}^{R} I\{\beta_j \in \widehat{\text{CI}}_{r,j}(1-\alpha)\}.$$

Since the confidence intervals for each component of $\beta_j$ are constructed using a common variance $\bar{\zeta}^2_{\text{emp}}$, we obtain only one averaged length of the confidence intervals

$$\widehat{\mathcal{L}} = \frac{2\Phi(1-\alpha/2)}{R} \sum_{r=1}^{R} \bar{\zeta}_{\text{emp,r}}. \tag{9}$$

From (9), it is obvious that the averaged length of the confidence interval is determined by the asymptotic variance of the bias-corrected estimator $\widetilde{\beta}$. The well constructed confidence intervals are expected to be close to the nominal coverage probabilities, but with short lengths. Since this measure is closely related to the asymptotic variance of the estimators, we reuse it to evaluate the robustness of the Huber estimator in Section 4.5.

Two-sided hypothesis testing is done for $H_{0,j} : \beta_{0,j} = 0$ for $j$ in the sets $S$ and $S^c$. To evaluate the performance of the individual tests at significance level $\alpha$, we calculate the averaged false positive (FP) and true positive (TP) rates, with $P_{r,j}$ the P-value for simulation $r$, component $j$,

$$\text{FP}(\alpha) = (p-s)^{-1} \sum_{j \in S^c} \sum_{r=1}^{R} I\{P_{r,j} \leq \alpha\}/R, \text{ and } \text{TP}(\alpha) = s^{-1} \sum_{j \in S} \sum_{r=1}^{R} I\{P_{r,j} \leq \alpha\}/R.$$

Simultaneous testing is done for $\{H_{0,j}, j \in \{1, \ldots, p\}\}$ including all individual hypotheses. This test is evaluated by the empirical version of the familywise error rate (FWER) defined as

$$\text{FWER}(\alpha) = \frac{1}{R} \sum_{r=1}^{R} I\{\text{at least one } H_{0,j}^{(r)} \text{ is rejected at adjusted } \alpha, j \in S^c\},$$

and the rejection percentage (RP) observed in the simulation study is defined as

$$\mathrm{RP}(\alpha) = \frac{1}{s}\sum_{j\in S}\Big\{\sum_{r=1}^{r} I\{H_{0,j}^{(r)} \text{ is rejected at adjusted } \alpha\}/R\Big\}.$$

### 4.3 Simulation results for the $\ell_1$-regularised quantile estimator

The quantile loss function is used at quantile level 0.5, thus performing $\ell_1$-regularised regression for the median response. The confidence levels are 0.95 and 0.99. Example 95% componentwise confidence intervals of the subvector consisting of non-zero components of $\beta$ in different settings are included in Figure 2 in Appendix B using the random seed number 5. The same data are reused for the normal QQ-plots of the $T_j(\beta_j)$, $j = 1,\dots,p$ in Figure 3. All plots, as well as $p$-values of the Shapiro-Wilk tests (included in the captions of each plot), confirm normality.

The average coverage probabilities $\widehat{\mathrm{CP}}_{\mathrm{vec}}(1-\alpha)$ and averaged lengths of the confidence intervals, averaging over the components of $\beta$ are presented in Table 1. We observe that: (1) the subvector corresponding to zero components of $\beta$ and the full vector $\beta$ mostly have average coverage probabilities close to the nominal values for 0.95 and 0.99; (2) for the subvector consisting of the non-zero part of $\beta$ the observed coverage is closer to the nominal value in the high-sparsity setting where $s = 5$, as compared to the medium sparsity setting where $s = 50$; (3) $t_3$ distributed errors have slightly shorter averaged lengths of the confidence intervals and lower average coverage probabilities for both high and medium sparsity settings; (4) the average coverage probabilities are closer to the nominal values and the lengths of confidence intervals get shorter when increasing the sample size $n$ from 100 to 250. This improvement is especially visible for the medium sparsity settings where $s = 50$, whereas the coverage probabilities are already stable in the high sparsity settings where $s = 5$.

Averaged FP and TP rates for different settings are presented in Table 2. We see that for high-sparsity settings where $s = 5$, the FP rates are already close to the nominal significance levels in cases where the sample size is small, i.e., $n = 100$. Also, the TP rates remain stable and have no significant improvement when the sample size $n$ is increased to 250. However, in the medium sparsity settings where $s = 50$, the FP and TP rates are largely improved by increasing the sample size from 100 to 250. Another observation is that the TP rates in settings with $t_3$ distributed errors are mostly the highest among all three errors for the same sparsity $s$ and the same distribution of non-zero components of $\beta$.

For the values of FWER and RP we observe similar patterns as for testing the individual hypothesis. In the high-sparsity settings where $s = 5$, values of FWER and RP are already stable when the sample size is small ($n = 100$), and are not significantly improved when increasing the sample size to $n = 250$. However, the FWERs are larger than the nominal significance level. In the medium sparsity settings where $s = 50$, the values of FWER and RP can be improved by increasing the sample size.

### 4.4 Simulation results for the $\ell_1$-regularised least squares estimator

We compare the performance of the RAMP-based estimator with the desparsifying approach in van de Geer et al. (2014) and the debiasing approach in Javanmard and Montanari (2014). Based on the observations for the quantile estimator, we present comparisons only in settings where $n = 100$ and the nonzero components of $\beta$ follow a Dirac distribution.

As expected, Table 3 shows that the three approaches have superior performance in the independent Gaussian settings due to the following two reasons: (1) our approach relies on the RAMP algorithm assuming the independence between $X_{\cdot j}$'s in Assumption (A1); (2) the Lasso estimator works best when the designs do not contain groups of highly correlated variables. The proposed construction based on the RAMP algorithm has similar coverage probabilities and averaged length of the confidence intervals as the desparsifying approach. In contrast, the debiasing approach provides less accurate coverage probabilities, lower than the nominal probability 95% for the subvector $\beta_S$ and

Table 1: Quantile estimator. The averaged $1-\alpha = 95\%$ and $99\%$ coverage probabilities $\mathrm{CP}_{\mathrm{vec}}(1-\alpha)$ and average lengths $\mathcal{L}$ of confidence intervals of subvectors of $\beta$ consisting of true non-zero values, true zero values and the complete vector, for $n = 100, \delta = 0.2$ and $n = 250, \delta = 0.5$.

| $n = 100$ | $\delta = 0.2$ | $\mathrm{CP}_{\mathrm{vec}}(0.95)$ | | | $\mathcal{L}$ | $\mathrm{CP}_{\mathrm{vec}}(0.99)$ | | | $\mathcal{L}$ |
|---|---|---|---|---|---|---|---|---|---|
| $f_\varepsilon$ | $s$ | non-zero | zero | full vector | | non-zero | zero | full vector | |
| Subvector of $\beta$ of nonzeros: **Dirac distribution at -1 and 1** | | | | | | | | | |
| $N(0,1)$ | 5 | 0.96 | 0.95 | 0.95 | 0.91 | 0.99 | 0.99 | 0.99 | 1.20 |
| | 50 | 0.88 | 0.94 | 0.93 | 2.20 | 0.97 | 0.98 | 0.98 | 2.90 |
| $t_3$ | 5 | 0.94 | 0.95 | 0.95 | 0.86 | 0.98 | 0.99 | 0.99 | 1.13 |
| | 50 | 0.88 | 0.93 | 0.93 | 2.19 | 0.97 | 0.98 | 0.98 | 2.88 |
| $0.5N(0,1)$ | 5 | 0.96 | 0.95 | 0.95 | 0.91 | 0.99 | 0.99 | 0.99 | 1.20 |
| $+0.5N(5,9)$ | 50 | 0.89 | 0.95 | 0.94 | 2.22 | 0.97 | 0.99 | 0.98 | 2.92 |
| Subvector of $\beta$ of nonzeros: $\mathbf{N(0,1)}$ | | | | | | | | | |
| $N(0,1)$ | 5 | 0.92 | 0.95 | 0.95 | 0.88 | 0.98 | 0.99 | 0.99 | 1.16 |
| | 50 | 0.92 | 0.94 | 0.93 | 2.27 | 0.98 | 0.98 | 0.98 | 2.99 |
| $t_3$ | 5 | 0.94 | 0.95 | 0.95 | 0.83 | 0.98 | 0.99 | 0.99 | 1.16 |
| | 50 | 0.92 | 0.94 | 0.93 | 2.23 | 0.98 | 0.98 | 0.98 | 2.94 |
| $0.5N(0,1)$ | 5 | 0.92 | 0.95 | 0.95 | 0.88 | 0.98 | 0.99 | 0.99 | 1.16 |
| $+0.5N(5,9)$ | 50 | 0.92 | 0.94 | 0.93 | 2.24 | 0.98 | 0.98 | 0.98 | 2.95 |
| $n = 250$ | $\delta = 0.5$ | $\mathrm{CP}_{\mathrm{vec}}(0.95)$ | | | $\mathcal{L}$ | $\mathrm{CP}_{\mathrm{vec}}(0.99)$ | | | $\mathcal{L}$ |
| $f_\varepsilon$ | $s$ | non-zero | zero | full vector | | non-zero | zero | full vector | |
| Subvector of $\beta$ of nonzeros: **Dirac distribution at -1 and 1** | | | | | | | | | |
| $N(0,1)$ | 5 | 0.93 | 0.95 | 0.95 | 0.90 | 0.98 | 0.99 | 0.99 | 1.18 |
| | 50 | 0.91 | 0.95 | 0.95 | 1.45 | 0.98 | 0.99 | 0.99 | 1.91 |
| $t_3$ | 5 | 0.92 | 0.94 | 0.94 | 0.64 | 0.97 | 0.98 | 0.98 | 0.84 |
| | 50 | 0.91 | 0.95 | 0.94 | 1.36 | 0.98 | 0.99 | 0.99 | 1.79 |
| $0.5N(0,1)$ | 5 | 0.94 | 0.95 | 0.95 | 1.16 | 0.99 | 0.99 | 0.99 | 1.52 |
| $+0.5N(5,9)$ | 50 | 0.92 | 0.95 | 0.95 | 1.49 | 0.98 | 0.99 | 0.99 | 1.96 |
| Subvector of $\beta$ of nonzeros: $\mathbf{N(0,1)}$ | | | | | | | | | |
| $N(0,1)$ | 5 | 0.93 | 0.95 | 0.95 | 0.89 | 0.99 | 0.99 | 0.99 | 1.17 |
| | 50 | 0.95 | 0.95 | 0.95 | 1.23 | 0.99 | 0.99 | 0.99 | 1.62 |
| $t_3$ | 5 | 0.93 | 0.94 | 0.94 | 0.63 | 0.98 | 0.98 | 0.98 | 0.83 |
| | 50 | 0.95 | 0.95 | 0.95 | 1.04 | 0.99 | 0.99 | 0.99 | 1.36 |
| $0.5N(0,1)$ | 5 | 0.94 | 0.95 | 0.95 | 1.16 | 0.99 | 0.99 | 0.99 | 1.53 |
| $+0.5N(5,9)$ | 50 | 0.95 | 0.95 | 0.95 | 1.28 | 0.99 | 0.99 | 0.99 | 1.69 |

higher than 95% for the subvector $\beta_{S^c}$ related to the components that are zero and for the full vector $\beta$. Additionally, the proposed approach outperforms the desparsifying approach on hypothesis testing; however, the debiasing approach has the best performance.

For a correlated Gaussian design with a Toeplitz correlation matrix the most surprising observation is that all three approaches have more accurate coverage probabilities in the medium sparsity settings where $s = 50$ than in the high sparsity settings where $s = 5$. However, the three approaches remain to have superior performances in high sparsity settings on hypothesis testing, and the RAMP-based

Table 2: Quantile estimator. Average FP and TP rates for individual hypothesis testing; as well as FWER and RP for multiple testing. Rates are calculated for $n = 100$ ($\delta = 0.2$) and $n = 250$ ($\delta = 0.5$).

| $n = 100$ | $\delta = 0.2$ | Significance $\alpha = 0.05$ | | | | Significance $\alpha = 0.01$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $f_\varepsilon$ | $s$ | FP | TP | FWER | RP | FP | TP | FWER | RP |
| Subvector of $\beta$ of nonzeros: **Dirac distribution at -1 and 1** | | | | | | | | | |
| $N(0, 1)$ | 5 | 0.05 | 0.99 | 0.10 | 0.63 | 0.01 | 0.96 | 0.05 | 0.49 |
| | 50 | 0.06 | 0.28 | 0.23 | 0.04 | 0.02 | 0.17 | 0.17 | 0.03 |
| $t_3$ | 5 | 0.05 | 0.99 | 0.10 | 0.70 | 0.01 | 0.95 | 0.03 | 0.57 |
| | 50 | 0.07 | 0.29 | 0.26 | 0.04 | 0.02 | 0.18 | 0.19 | 0.03 |
| $0.5N(0, 1)$ | 5 | 0.05 | 0.99 | 0.10 | 0.63 | 0.01 | 0.97 | 0.04 | 0.47 |
| $+0.5N(5, 9)$ | 50 | 0.05 | 0.27 | 0.17 | 0.03 | 0.02 | 0.16 | 0.10 | 0.02 |
| Subvector of $\beta$ of nonzeros: $\mathbf{N(0, 1)}$ | | | | | | | | | |
| $N(0, 1)$ | 5 | 0.05 | 0.80 | 0.10 | 0.67 | 0.01 | 0.76 | 0.03 | 0.64 |
| | 50 | 0.06 | 0.28 | 0.21 | 0.05 | 0.02 | 0.16 | 0.07 | 0.03 |
| $t_3$ | 5 | 0.05 | 0.81 | 0.11 | 0.69 | 0.01 | 0.78 | 0.02 | 0.66 |
| | 50 | 0.07 | 0.27 | 0.22 | 0.05 | 0.02 | 0.16 | 0.08 | 0.04 |
| $0.5N(0, 1)$ | 5 | 0.05 | 0.80 | 0.11 | 0.67 | 0.01 | 0.76 | 0.04 | 0.64 |
| $+0.5N(5, 9)$ | 50 | 0.07 | 0.28 | 0.22 | 0.05 | 0.02 | 0.16 | 0.08 | 0.04 |
| $n = 250$ | $\delta = 0.5$ | Significance $\alpha = 0.05$ | | | | Significance $\alpha = 0.01$ | | | |
| $f_\varepsilon$ | $s$ | FP | TP | FWER | RP | FP | TP | FWER | RP |
| Subvector of $\beta$ of nonzeros: **Dirac distribution at -1 and 1** | | | | | | | | | |
| $N(0, 1)$ | 5 | 0.05 | 0.99 | 0.09 | 0.66 | 0.01 | 0.95 | 0.02 | 0.53 |
| | 50 | 0.05 | 0.67 | 0.20 | 0.13 | 0.01 | 0.47 | 0.07 | 0.07 |
| $t_3$ | 5 | 0.06 | 1.00 | 0.20 | 0.96 | 0.02 | 1.00 | 0.10 | 0.93 |
| | 50 | 0.05 | 0.73 | 0.22 | 0.18 | 0.01 | 0.54 | 0.08 | 0.12 |
| $0.5N(0, 1)$ | 5 | 0.05 | 0.91 | 0.11 | 0.33 | 0.01 | 0.77 | 0.04 | 0.25 |
| $+0.5N(5, 9)$ | 50 | 0.05 | 0.65 | 0.16 | 0.12 | 0.01 | 0.46 | 0.06 | 0.07 |
| Subvector of $\beta$ of nonzeros: $\mathbf{N(0, 1)}$ | | | | | | | | | |
| $N(0, 1)$ | 5 | 0.05 | 0.81 | 0.10 | 0.70 | 0.01 | 0.79 | 0.04 | 0.66 |
| | 50 | 0.05 | 0.56 | 0.06 | 0.27 | 0.01 | 0.46 | 0.02 | 0.22 |
| $t_3$ | 5 | 0.07 | 0.82 | 0.21 | 0.79 | 0.02 | 0.81 | 0.12 | 0.77 |
| | 50 | 0.05 | 0.68 | 0.07 | 0.39 | 0.01 | 0.58 | 0.03 | 0.34 |
| $0.5N(0, 1)$ | 5 | 0.05 | 0.79 | 0.07 | 0.52 | 0.01 | 0.74 | 0.02 | 0.44 |
| $+0.5N(5, 9)$ | 50 | 0.05 | 0.60 | 0.10 | 0.28 | 0.01 | 0.49 | 0.03 | 0.23 |

approach has the best performance in almost all settings. It is worth noticing that all three approaches have very high FWER and low PR in the multiple testing scenario, which suggests high error rates and weakened power in multiple testing for datasets with highly correlated variables.

### 4.5 Simulation results for the $\ell_1$-regularised Huber estimator

Although the major focus in this section is the robustness of the RAMP-based Huber estimator in the presence of outliers, in Table 4 we present a numerical comparison between RAMP-based Huber,

Table 3: Lasso estimator. The top half table presents the average coverage probabilities $\mathrm{CP}_{\mathrm{vec},j}(1-\alpha), j = 1, \ldots, p_{\mathrm{vec}}$ of subvectors of $\beta$ and average length $\mathcal{L}(1-\alpha)$ of confidence intervals for $n = 100$ ($\delta = 0.2$) for $1-\alpha = 0.95$. The bottom half table presents the average FP and TP rates for individual hypothesis testing, and FWER and PR for multiple testing. Three approaches constructing the confidence intervals are: (i) (9) based on the RAMP algorithm (left); (ii) the desparsifying approach in van de Geer et al. (2014) (middle); (iii) the debiasing approach in Javanmard and Montanari (2014) (right).

| Independent Gaussian Design | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n = 100$ | $\delta = 0.2$ | RAMP algorithm | | | | Desparsifying | | | | Debiasing | | | |
| | | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ |
| $f_\varepsilon$ | $s$ | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | |
| Subvector of $\beta$ of nonzeros: **Dirac distribution at -1 and 1** | | | | | | | | | | | | | |
| $N(0,1)$ | 5 | 0.95 | 0.95 | 0.95 | 0.90 | 0.96 | 0.96 | 0.96 | 0.96 | 0.82 | 0.99 | 0.99 | 0.83 |
| | 50 | 0.86 | 0.94 | 0.93 | 2.10 | 0.87 | 0.97 | 0.96 | 2.16 | 0.86 | 0.99 | 0.98 | 2.00 |
| $t_3$ | 5 | 0.93 | 0.95 | 0.95 | 0.89 | 0.95 | 0.96 | 0.96 | 1.00 | 0.81 | 0.99 | 0.99 | 0.83 |
| | 50 | 0.86 | 0.94 | 0.93 | 2.12 | 0.88 | 0.97 | 0.96 | 2.18 | 0.86 | 0.99 | 0.98 | 1.99 |
| $0.5N(0,1)$ | 5 | 0.95 | 0.95 | 0.95 | 0.90 | 0.95 | 0.96 | 0.96 | 0.94 | 0.82 | 0.99 | 0.99 | 0.84 |
| $+0.5N(5,9)$ | 50 | 0.86 | 0.94 | 0.93 | 2.12 | 0.88 | 0.97 | 0.96 | 2.18 | 0.86 | 0.99 | 0.98 | 1.99 |
| Correlated Gaussian Design | | | | | | | | | | | | | |
| $n = 100$ | $\delta = 0.2$ | RAMP algorithm | | | | Desparsifying | | | | Debiasing | | | |
| | | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ |
| $f_\varepsilon$ | $s$ | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | |
| Subvector of $\beta$ of nonzeros: **Dirac distribution at -1 and 1** | | | | | | | | | | | | | |
| $N(0,1)$ | 5 | 0.84 | 0.94 | 0.94 | 1.69 | 0.91 | 0.90 | 0.90 | 2.11 | 0.86 | 0.99 | 0.99 | 1.89 |
| | 50 | 0.94 | 0.94 | 0.94 | 2.81 | 0.94 | 0.92 | 0.92 | 3.44 | 0.96 | 0.99 | 0.98 | 2.91 |
| $t_3$ | 5 | 0.84 | 0.95 | 0.94 | 1.69 | 0.91 | 0.90 | 0.90 | 2.13 | 0.86 | 0.99 | 0.99 | 1.87 |
| | 50 | 0.94 | 0.94 | 0.93 | 2.79 | 0.94 | 0.92 | 0.92 | 3.45 | 0.96 | 0.99 | 0.98 | 2.98 |
| $0.5N(0,1)$ | 5 | 0.85 | 0.95 | 0.94 | 1.69 | 0.91 | 0.90 | 0.90 | 2.10 | 0.86 | 0.99 | 0.99 | 1.87 |
| $+0.5N(5,9)$ | 50 | 0.95 | 0.95 | 0.95 | 2.85 | 0.94 | 0.92 | 0.92 | 3.45 | 0.96 | 0.99 | 0.99 | 2.94 |

| Independent Gaussian Design | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n = 100$ | $\delta = 0.2$ | RAMP algorithm | | | | Desparsifying | | | | Debiasing | | | |
| $f_\varepsilon$ | $s$ | FP | TP | FWER | PR | FP | TP | FWER | PR | FP | TP | FWER | PR |
| Subvector of $\beta$ of nonzeros: **Dirac distribution at -1 and 1** | | | | | | | | | | | | | |
| $N(0,1)$ | 5 | 0.05 | 0.99 | 0.12 | 0.65 | 0.04 | 0.98 | 0.07 | 0.48 | 0.06 | 0.99 | 0.79 | 0.90 |
| | 50 | 0.06 | 0.28 | 0.18 | 0.04 | 0.03 | 0.22 | 0.54 | 0.02 | 0.06 | 0.28 | 0.94 | 0.11 |
| $t_3$ | 5 | 0.05 | 0.98 | 0.12 | 0.66 | 0.04 | 0.97 | 0.06 | 0.46 | 0.06 | 0.98 | 0.81 | 0.88 |
| | 50 | 0.06 | 0.28 | 0.14 | 0.04 | 0.03 | 0.22 | 0.51 | 0.02 | 0.06 | 0.28 | 0.94 | 0.11 |
| $0.5N(0,1)$ | 5 | 0.05 | 0.99 | 0.11 | 0.64 | 0.04 | 0.99 | 0.06 | 0.51 | 0.06 | 0.99 | 0.77 | 0.90 |
| $+0.5N(5,9)$ | 50 | 0.05 | 0.27 | 0.20 | 0.04 | 0.03 | 0.22 | 0.53 | 0.02 | 0.06 | 0.28 | 0.95 | 0.12 |
| Correlated Gaussian Design | | | | | | | | | | | | | |
| $n = 100$ | $\delta = 0.2$ | RAMP algorithm | | | | Desparsifying | | | | Debiasing | | | |
| $f_\varepsilon$ | $s$ | FP | TP | FWER | PR | FP | TP | FWER | PR | FP | TP | FWER | PR |
| Subvector of $\beta$ of nonzeros: **Dirac distribution at -1 and 1** | | | | | | | | | | | | | |
| $N(0,1)$ | 5 | 0.06 | 0.40 | 0.62 | 0.09 | 0.10 | 0.40 | 0.86 | 0.03 | 0.06 | 0.37 | 0.85 | 0.09 |
| | 50 | 0.06 | 0.16 | 0.74 | 0.01 | 0.08 | 0.13 | 0.88 | 0.00 | 0.06 | 0.12 | 0.90 | 0.01 |
| $t_3$ | 5 | 0.05 | 0.40 | 0.87 | 0.08 | 0.10 | 0.38 | 0.83 | 0.04 | 0.06 | 0.37 | 0.85 | 0.10 |
| | 50 | 0.06 | 0.16 | 0.74 | 0.01 | 0.08 | 0.12 | 0.87 | 0.00 | 0.06 | 0.12 | 0.91 | 0.01 |
| $0.5N(0,1)$ | 5 | 0.05 | 0.40 | 0.63 | 0.08 | 0.10 | 0.40 | 0.83 | 0.04 | 0.06 | 0.38 | 0.83 | 0.09 |
| $+0.5N(5,9)$ | 50 | 0.05 | 0.15 | 0.72 | 0.01 | 0.08 | 0.13 | 0.88 | 0.00 | 0.06 | 0.12 | 0.90 | 0.01 |

RAMP-based quantile, and RAMP-based least squares estimators to complement Table 5. Table 4 considers the small sample size settings used before, where $p = 500$, $n = 100$ resulting in the ratio $\delta = 0.2$, the sparsity $s = 5, 50$, and a correlated Gaussian design as used for Table 3. Further, the same regression error distributions $N(0,1)$, $t_3$, $0.5N(0,1)+0.5N(5,9)$ are considered for Table 4. Since these settings are part of the settings for Table 3 comparing the RAMP-based least squares with the debiasing and desparsifying approaches, the debiasing and desparsifying approaches are not included in Table 4, which now focuses on the performance comparison between the different estimators. We make the follow observations from Table 4: (1) The values in Table 4 agree with the records for correlated design matrix in Table 3. (2) The three estimators have a similar performance for all three error distributions. But for $0.5N(0,1)+0.5N(5,9)$ distributed errors, when the nonzero components of $\beta$ follow a Dirac distribution at -1 and 1 with $s=5$, the Huber estimator remain stable whereas the other two have slightly worse performance. (3) The Huber estimator has a larger variance than the least-squares estimator, which is reflected in the averaged length of the confidence intervals via the value $\bar{\zeta}$.

Table 4: Comparing performance of the Huber, the Lasso, and the quantile estimator. The top half table presents the average coverage probabilities $\mathrm{CP}_{\mathrm{vec},j}(1-\alpha), j=1,\ldots,p_{\mathrm{vec}}$ of subvectors of $\beta$ and the averaged length $\mathcal{L}(1-\alpha)$ for $n=100$ ($\delta=0.2$) for $1-\alpha=0.95$. The bottom half table presents the average FP and TP rates for individual hypothesis testing, and FWER and PR for multiple testing. Three approaches constructing the confidence intervals are: (i) RAMP algorithm with Huber loss (left); (ii) RAMP algorithm with least squares loss (middle); (iii) RAMP algorithm with quantile loss (right).

| Correlated Gaussian Design | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n=100$ | $\delta=0.2$ | Huber | | | | Least Squares | | | | Quantile | | | |
| | | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ |
| $f_\varepsilon$ | $s$ | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | |
| Subvector of $\beta$ of nonzeros: **Dirac distribution at -1 and 1** | | | | | | | | | | | | | |
| $N(0,1)$ | 5 | 0.85 | 0.95 | 0.94 | 1.72 | 0.85 | 0.94 | 0.94 | 1.72 | 0.84 | 0.94 | 0.94 | 1.69 |
| | 50 | 0.95 | 0.95 | 0.95 | 2.87 | 0.95 | 0.94 | 0.95 | 2.82 | 0.95 | 0.94 | 0.94 | 2.82 |
| $t_3$ | 5 | 0.85 | 0.95 | 0.95 | 1.78 | 0.83 | 0.95 | 0.94 | 1.69 | 0.83 | 0.95 | 0.94 | 1.68 |
| | 50 | 0.95 | 0.94 | 0.94 | 2.83 | 0.95 | 0.95 | 0.95 | 2.83 | 0.95 | 0.95 | 0.95 | 2.83 |
| $0.5N(0,1)$ | 5 | 0.85 | 0.95 | 0.94 | 1.72 | 0.82 | 0.92 | 0.92 | 1.69 | 0.83 | 0.92 | 0.92 | 1.71 |
| $+0.5N(5,9)$ | 50 | 0.95 | 0.95 | 0.95 | 2.89 | 0.96 | 0.95 | 0.95 | 2.89 | 0.95 | 0.95 | 0.95 | 2.83 |
| $n=100$ | $\delta=0.2$ | Huber | | | | Least Squares | | | | Quantile | | | |
| | | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ | $\mathrm{CP}_{\mathrm{vec}}$ | | | $\mathcal{L}$ |
| $f_\varepsilon$ | $s$ | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | |
| Subvector of $\beta$ of nonzeros: $N(0,1)$ | | | | | | | | | | | | | |
| $N(0,1)$ | 5 | 0.93 | 0.95 | 0.95 | 1.76 | 0.94 | 0.95 | 0.95 | 1.75 | 0.94 | 0.95 | 0.95 | 1.77 |
| | 50 | 0.89 | 0.96 | 0.95 | 2.92 | 0.90 | 0.96 | 0.95 | 2.92 | 0.89 | 0.95 | 0.94 | 2.85 |
| $t_3$ | 5 | 0.93 | 0.95 | 0.95 | 1.78 | 0.93 | 0.94 | 0.94 | 1.78 | 0.93 | 0.94 | 0.94 | 1.75 |
| | 50 | 0.89 | 0.96 | 0.95 | 2.89 | 0.88 | 0.96 | 0.95 | 2.83 | 0.89 | 0.95 | 0.95 | 2.85 |
| $0.5N(0,1)$ | 5 | 0.93 | 0.95 | 0.95 | 1.76 | 0.94 | 0.95 | 0.95 | 1.75 | 0.94 | 0.95 | 0.95 | 1.77 |
| $+0.5N(5,9)$ | 50 | 0.89 | 0.96 | 0.95 | 2.90 | 0.89 | 0.96 | 0.95 | 2.84 | 0.89 | 0.96 | 0.95 | 2.88 |
| $n=100$ | $\delta=0.2$ | Huber | | | | Least Squares | | | | Quantile | | | |
| $f_\varepsilon$ | $s$ | FP | TP | FWER | PR | FP | TP | FWER | PR | FP | TP | FWER | PR |
| Subvector of $\beta$ of nonzeros: **Dirac distribution at -1 and 1** | | | | | | | | | | | | | |
| $N(0,1)$ | 5 | 0.05 | 0.39 | 0.62 | 0.08 | 0.05 | 0.39 | 0.62 | 0.08 | 0.05 | 0.39 | 0.62 | 0.08 |
| | 50 | 0.05 | 0.15 | 0.72 | 0.01 | 0.06 | 0.16 | 0.73 | 0.01 | 0.06 | 0.16 | 0.74 | 0.01 |
| $t_3$ | 5 | 0.05 | 0.40 | 0.63 | 0.08 | 0.05 | 0.41 | 0.63 | 0.08 | 0.05 | 0.41 | 0.61 | 0.09 |
| | 50 | 0.06 | 0.17 | 0.75 | 0.01 | 0.05 | 0.15 | 0.74 | 0.01 | 0.05 | 0.16 | 0.75 | 0.01 |
| $0.5N(0,1)$ | 5 | 0.05 | 0.40 | 0.64 | 0.08 | 0.05 | 0.40 | 0.64 | 0.08 | 0.05 | 0.40 | 0.64 | 0.08 |
| $+0.5N(5,9)$ | 50 | 0.05 | 0.15 | 0.76 | 0.01 | 0.05 | 0.15 | 0.77 | 0.01 | 0.05 | 0.16 | 0.78 | 0.01 |
| $n=100$ | $\delta=0.2$ | Huber | | | | Least Squares | | | | Quantile | | | |
| $f_\varepsilon$ | $s$ | FP | TP | FWER | PR | FP | TP | FWER | PR | FP | TP | FWER | PR |
| Subvector of $\beta$ of nonzeros: $N(0,1)$ | | | | | | | | | | | | | |
| $N(0,1)$ | 5 | 0.05 | 0.64 | 0.27 | 0.21 | 0.05 | 0.63 | 0.28 | 0.21 | 0.05 | 0.63 | 0.28 | 0.21 |
| | 50 | 0.04 | 0.32 | 0.16 | 0.07 | 0.04 | 0.32 | 0.19 | 0.07 | 0.05 | 0.34 | 0.23 | 0.08 |
| $t_3$ | 5 | 0.05 | 0.63 | 0.27 | 0.21 | 0.06 | 0.62 | 0.28 | 0.22 | 0.06 | 0.63 | 0.27 | 0.24 |
| | 50 | 0.04 | 0.32 | 0.15 | 0.07 | 0.04 | 0.32 | 0.16 | 0.08 | 0.05 | 0.33 | 0.18 | 0.08 |
| $0.5N(0,1)$ | 5 | 0.05 | 0.64 | 0.27 | 0.21 | 0.05 | 0.63 | 0.28 | 0.21 | 0.05 | 0.63 | 0.28 | 0.21 |
| $+0.5N(5,9)$ | 50 | 0.04 | 0.32 | 0.17 | 0.07 | 0.04 | 0.32 | 0.18 | 0.08 | 0.04 | 0.32 | 0.17 | 0.07 |

Next, we compare the performance of the RAMP-based Huber estimator with the RAMP-based least squares estimator and the desparsifying approach in van de Geer et al. (2014). Except for the error distributions the simulation settings are those used for Table 4. We consider two mixed normal distributions $0.1N(9, 0.2) + 0.9N(-1, 2)$ and $0.1N(18, 0.01) + 0.9N(-2, 0.2)$ reflecting situations with outlying observations as often seen in robust literature (Alfons et al., 2013; Khan et al., 2007). By this table, we illustrate that using robust estimators instead of the least squares estimator could improve estimation accuracy when outliers exist in the regression errors. And further, without complicated derivation when switching between estimators, the proposed method provides a reasonably accurate and convenient construction of confidence intervals and hypothesis testing. Ideally, robust estimators should be both accurate (low bias) and efficient (low asymptotic variance) (Huber, 2004; Hampel et al., 2011). The efficiency of $\widetilde{\beta}$ can be evaluated using the averaged length of the confidence intervals $\mathcal{L}$ in (9). To evaluate the accuracy of the biased-corrected estimator $\widetilde{\beta}$, we report the simulated mean squared error and the mean squared prediction error over $R = 500$ replications, that is,

$$\mathrm{MSE}(\widetilde{\beta}) = \frac{1}{R}\sum_{r=1}^{R}\Big(\frac{1}{p}\sum_{j=1}^{p}(\widetilde{\beta}_{r,j} - \beta_{r,j})^2\Big); \quad \mathrm{MSPE}(\widetilde{\beta}) = \frac{1}{R}\sum_{r=1}^{R}\Big(\frac{1}{n}\sum_{i=1}^{n}\big(Y_i^{(r)} - (X_{i\cdot}^{(r)})^\top\widetilde{\beta}\big)^2\Big),$$

where $(Y^{(r)}, X^{(r)}), r = 1, \ldots, 500$ are independent copies of the original dataset $(Y, X)$. In addition, we report a bias related measure, which is often seen in the robust statistics literature (Huber, 2004; Hampel et al., 2011) with $\Psi(\widetilde{\beta}) \to 0$ suggesting Fisher consistency of the estimator. For the proposed method,

$$\Psi(\widetilde{\beta}) = \frac{1}{R}\sum_{r=1}^{R}\Big(\sum_{i=1}^{n}\partial\rho\Big(\mathrm{Prox}(Y_i^{(r)} - (X_{i\cdot}^{(r)})^\top\widetilde{\beta}_r; b)\Big)\Big), \tag{10}$$

where $\partial\rho(z) = 2z$ for the least squares loss function and $\partial\rho(z) = zI\{|z| \le u\} + (u \cdot \mathrm{sign}(z))I\{|z| > u\}$ for the Huber loss. For the desparsifying approach in van de Geer et al. (2014),

$$\Psi(\widetilde{\beta}) = \frac{1}{R}\sum_{r=1}^{R}\Big(\sum_{i=1}^{n}\partial\rho\Big(Y_i^{(r)} - (X_{i\cdot}^{(r)})^\top\widetilde{\beta}_r\Big)\Big) = \frac{1}{R}\sum_{r=1}^{R}\Big(2(Y_i^{(r)} - (X_{i\cdot}^{(r)})^\top\widetilde{\beta})\Big). \tag{11}$$

Table 5: Huber and Lasso estimators. The top half table presents the average coverage probabilities $\mathrm{CP}_{\mathrm{vec},j}(1-\alpha), j = 1, \ldots, p_{\mathrm{vec}}$ of subvectors of $\beta$ and the averaged TP and FP rates for $n = 100$ ($\delta = 0.2$) for $1 - \alpha = 0.95$. The bottom half table presents three estimation measurements MSE, MSPE, $\Psi$, and the averaged length $\mathcal{L}(1 - \alpha)$ of confidence intervals. Three approaches constructing the confidence intervals are: (i) RAMP algorithm with Huber loss (left); (ii) RAMP algorithm with least squares loss (middle); (iii) the desparsifying approach in van de Geer et al. (2014) (right).

| Subvector of $\beta$ of nonzeros: **Dirac distribution at -1 and 1** | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n = 100$ | $\delta = 0.2$ | RAMP Huber | | | | | RAMP LS | | | | | Desparsifying | | | | |
| | | $\mathrm{CP}_{\mathrm{vec}}$ | | | FP | TP | $\mathrm{CP}_{\mathrm{vec}}$ | | | FP | TP | $\mathrm{CP}_{\mathrm{vec}}$ | | | FP | TP |
| $f_\varepsilon$ | $s$ | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | | | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | | | $\beta_S$ | $\beta_{S^c}$ | $\beta$ | | |
| $0.1N(-2, 0.01)$ | 5 | 0.98 | 0.95 | 0.95 | 0.05 | 0.05 | 0.95 | 0.95 | 0.95 | 0.05 | 0.05 | 0.96 | 0.96 | 0.95 | 0.05 | 0.05 |
| $+0.9N(18, 0.2)$ | 50 | 0.93 | 0.96 | 0.95 | 0.04 | 0.08 | 0.95 | 0.95 | 0.95 | 0.05 | 0.05 | 0.95 | 0.95 | 0.95 | 0.05 | 0.05 |
| $0.1N(-1, 0.2)$ | 5 | 0.97 | 0.95 | 0.95 | 0.05 | 0.11 | 0.95 | 0.95 | 0.95 | 0.05 | 0.06 | 0.96 | 0.96 | 0.95 | 0.05 | 0.06 |
| $+0.9N(9, 0.2)$ | 50 | 0.92 | 0.96 | 0.95 | 0.04 | 0.11 | 0.95 | 0.95 | 0.95 | 0.05 | 0.06 | 0.93 | 0.94 | 0.95 | 0.05 | 0.06 |
| $f_\varepsilon$ | $s$ | MSE | MSPE | $\Psi$ | $\mathcal{L}$ | | MSE | MSPE | $\Psi$ | $\mathcal{L}$ | | MSE | MSPE | $\Psi$ | $\mathcal{L}$ | |
| $0.1N(-2, 0.01)$ | 5 | 6.29 | 67.60 | 0.30 | 9.80 | | 35.94 | 216.17 | 0.22 | 23.36 | | 37.48 | 220.10 | 2.50 | 23.65 | |
| $+0.9N(18, 0.2)$ | 50 | 6.83 | 70.37 | 0.27 | 10.35 | | 36.21 | 216.05 | 0.10 | 23.43 | | 37.60 | 218.41 | 2.31 | 23.56 | |
| $0.1N(-1, 0.2)$ | 5 | 1.77 | 18.02 | 0.15 | 5.19 | | 9.15 | 54.89 | 0.11 | 11.79 | | 9.29 | 55.82 | 1.26 | 11.93 | |
| $+0.9N(9, 0.2)$ | 50 | 2.40 | 21.14 | 0.13 | 6.14 | | 9.39 | 55.45 | 0.10 | 11.94 | | 9.74 | 57.00 | 1.20 | 12.11 | |

The performance of the proposed construction using the Huber loss function is compared with the proposed construction using the least squares loss and the desparsifying approach in van de Geer et al. (2014). All three constructions have similar averaged coverage rates that are close to the nominal one.

The proposed construction using the Huber loss has slightly higher TP rates. However, the Huber estimator has a dominant superior performance in estimation. The MSEs and MSPEs are much lower than for the other two methods. The averaged length of the confidence intervals using the Huber estimator is also much shorter than that of the competitors.

# 5 Data Application

## 5.1 Sparse signal recovery

We consider the audio wave signals example used in Zhou et al. (2020, Section 7.2) which is available in the R (R Core Team, 2022) package `signal` (signal developers, 2014). The artificial compressed sensing process involves a wavelet transform of the original audio signal for obtaining a 'sparse' representation of $\beta \in \mathbb{R}^{2047}$. To avoid confusion, this data application is not a typical dataset in statistics. In statistical analysis, the predictive variables and the response variable are considered as data which are used to estimate the unknown parameter vector $\beta$. However, the main methodology used in this paper—the approximate message passing algorithm—was initially proposed with an application to compressed sensing which perceives the signal $\beta$ as data. We leave this atypical data example here for suggesting an alternative application in compressed sensing.

The artificial compressed sensing process is as follows. First, the sparse signal $\beta$ is compressed by a randomly generated compression matrix $X \in \mathbb{R}^{1024\times 2047}$; components of the matrix $X_{ij}$ are i.i.d. with a $N(0, 1/1024)$ distribution. Next, the compressed signal $X\beta$ is sent to a receiver; the received signal $Y$ from transmission is corrupted by error $\boldsymbol{\varepsilon}$. Components of the error vector $\boldsymbol{\varepsilon}$ are randomly generated from either $t_3$ or mixed normal distribution $0.5N(0,1) + 0.5N(5,9)$, and are rescaled to have standard deviation 0.03.

In practice, we are interested in recovering the sparse signal $\beta$ from the compression matrix $X$ and the received signal $Y$. Here, the wavelet audio signal $\beta$ is known, and is artificially compressed and corrupted.

We first construct componentwise confidence intervals. For a clearer presentation, we only plot confidence intervals of the last 20 entries of $\beta$, see Figure 1. Normal QQ-plots (not shown) of the statistics $T_j(\beta_j), j = 1, \ldots, p$ confirm normality with a Shapiro-Wilk $p$-value of 0.460 for $t_3$ distributed errors and 0.615 for $0.5N(0,1) + 0.5N(5,9)$ distributed errors.

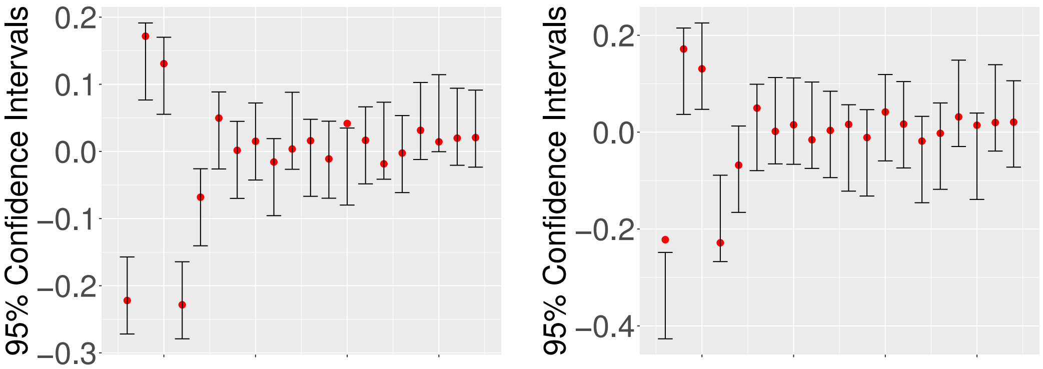


Figure 1: Audio signal data. The 95% confidence intervals of the last 20 entries of $\beta$. True values are depicted by red dots. Left: $t_3$ errors, right: $0.5N(0,1) + 0.5N(5,9)$ errors.

Next, we consider a multiple testing scenario including individual null hypotheses $H_{0,j} : \beta_j = 0$ versus two-sided alternative hypotheses $H_{a,j} : \beta_j \neq 0$ for all components of the wavelet coefficient $\beta$ of the audio signal fraction. Since most $\beta_j$'s are close to zero with countable non-negligible entries, we set cut-off values by taking the $(\tau/2)$th and $(1 - \tau/2)$th empirical quantile of the $\beta_j$'s with $\tau =$

$0.01, 0.05$. Our goal is to identify $\beta_j$'s with magnitude exceeding the two cut-off values by the multiple hypothesis test. The nominal significance levels of the test are $\alpha = 0.01$ and $0.05$. We replicate the artificial compressed sensing process $R = 200$ times, and evaluate the performance of the simultaneous hypothesis test by the familywise error rate (FWER) and rejection percentage (RP), see Table 6. We observe that when setting the level of the cut-off value $\tau = 0.05$ (i.e., select $\beta_j$'s whose magnitude greater than 97.5% or less than 2.5% of $\beta_j$'s magnitude), the FWER and RP of the test are both low for the both $\alpha = 0.01$ and $0.05$. On the contrary, the FWER and RP of the test are both high when the cut-off level $\tau = 0.01$ (i.e., select $\beta_j$'s whose magnitude exceed the range of 99% of $\beta_j$'s magnitude). This observation is not surprising: The cut-off level $\tau$ decides if $\beta_j$'s are counted as non-negligible entries; higher cut-off level $\tau$ results in less non-negligible entries. When an individual null hypothesis $H_{0,j}$ is rejected, it can be counted as a Type I error when $\tau = 0.01$, resulting in high FWER, but counted as a correct rejection when $\tau = 0.05$. This situation happens for $\beta_j$'s whose magnitudes are not high enough to be counted as non-negligible entries when $\tau = 0.01$, but are counted as non-negligible entries when $\tau = 0.05$. Similarly, when the magnitudes of the $\beta_j$'s are low enough for rejecting $H_{0,j}$ but high enough to exceed the cut-off level 0.05, the Type II error increases resulting in a low RP.

Table 6: Audio signal data. FWER and RP of multiple hypothesis tests at levels 0.05 and 0.01 to detect variables with magnitude exceeding certain cut-off values which are determined by the $(\tau/2)$th and $(1 - \tau/2)$th quantile of $\beta_j$'s.

|  | Cut-off level $\tau$ | 0.05 |  | 0.01 |  |
|---|---|---|---|---|---|
|  | Significance $\alpha$ | 0.05 | 0.01 | 0.05 | 0.01 |
| $t_3$ | FWER | 0.11 | 0.11 | 0.93 | 0.81 |
|  | RP | 0.22 | 0.20 | 0.88 | 0.85 |
| $0.5N(0,1)$ | FWER | 0.06 | 0.00 | 0.38 | 0.14 |
| $+0.5N(5,9)$ | RP | 0.13 | 0.11 | 0.58 | 0.51 |

## 5.2 Toxicity dataset

We consider the `toxicity` dataset available in the R package `robustbase` (Maechler et al., 2022). The original experiment aims at investigating the toxicity of carboxylic acids of 38 samples using 9 molecular descriptors. In the stage of data exploratory, nonlinear correlation between `toxicity` and most predictive variables, as well as outliers in the predictive variables are visible in the scatter plots of the response `toxicity` against the 9 predictive variables. In addition, we first fit a linear model by regressing `toxicity` on the 9 predictive variables. The residual plots including the Residual vs Fitted, Normal Q-Q plot, Scale-Location, Residual vs Leverage plots suggest that at least observations 28, 34, 38 potentially have outlying regression errors. We construct 36 pairwise interactions of the 9 predictive variables resulting in an expanded dataset with 45 predictive variables and 38 samples. Since this dataset potentially contains outliers, we consider the proposed testing procedure using both the Huber and the least squares loss functions. The dataset incorporates the decorrelation procedure in (6). Further, we compare our construction with the desparsifying method in van de Geer et al. (2014).

Similarly to Section 5, we consider a multiple testing scenario including individual hypotheses for all components to be equal to zero at nominal significance level $\alpha = 0.05$. The significant variables are included in Table 7. Our method using the Huber loss function with parameter 1.8 identifies 10 predictive variables including 5 main terms `pKa, ELUMO, Ecarb, Emet, IR`. With the least squares loss function, our method selects 9 predictive variables including `pKa, ELUMO, RM`. The method by van de Geer et al. (2014) identifies only an interaction term `Ecarb:IR` on the original expanded dataset; however, 4 main terms `pKa, ELUMO, IR, P` and 9 in total are significant by van de Geer et al. (2014) on the decorrelated dataset. Further, we investigate the bias of the estimators measured by the Fisher consistency for which a close-to-zero value suggests low bias. By using (11), the Fisher consistency of the RAMP-based least squares estimator is approximately equal to $-3.66$ and that of the deparsifying estimator is 5.35. Similarly, by (10), the Fisher consistency measure is around 0.09. In addition, we report the mean absolute prediction error (MAPE) using both the biased regularised estimator $\widehat{\beta}$ in (1) and $\widetilde{\beta}$ in (2) and (4) using the following expressions,

Table 7: Toxicity data. Significant variables by multiple hypothesis tests at nominal significance level 0.05.

| | Significant variables | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| RAMP Huber | pKa | ELUMO | Ecarb | Emet | IR | logKow:ELUMO | ELUMO:Ts | Ecarb:Ts | Emet:P | RM:IR |
| RAMP LS | pKa | ELUMO | RM | logKow:pKa | logKow:P | ELUMO:RM | ELUMO:IR | ELUMO:Ts | Emet:Ts | - |
| Desparsify | pKa | ELUMO | IR | P | logKow:RM | pKa:ELUMO | ELUMO:Ecarb | ELUMO:P | Ecarb:Ts | - |

$$\mathrm{MAPE}(\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - X_{i\cdot}^\top\widehat{\beta}\right|, \quad \mathrm{MAPE}(\widetilde{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - X_{i\cdot}^\top\widetilde{\beta}\right|.$$

The $\mathrm{MAPE}(\widehat{\beta})$ for the RAMP-based Huber estimator is 57.63, which outperforms the RAMP-based least squares estimator for which the MAPE value is 62.73 and the Lasso estimator with MAPE value 62.44. However, the $\mathrm{MAPE}(\widetilde{\beta})$ for the RAMP-based Huber estimator has the least-favorable value (107.59) compared to the RAMP-based least squares estimator (MAPE value 46.33) and the desparsified Lasso (MAPE value 43.98).

## 6 Discussion

The robust approximate message passing algorithm is interesting not only because it yields a mean squared error value for all components of the $\ell_1$-regularised estimators and as such makes valid post-selection inference possible. As we explored in this paper, without additional computational effort the algorithm also provides an estimator $\widetilde{\beta}$ comparable to the debiased and desparsified lasso estimators which can be used for hypothesis testing and for the construction of confidence intervals. An additional bonus is the flexibility in the choice of the loss function. The decorrelation step circumvents the theoretical requirement of having an uncorrelated design.

While this study has focused on the high-dimensional linear model, it would be interesting to expand the robust approximate message algorithms and their use for simultaneous inference using an $\ell_1$-regularized estimators to other types of distributions and models. In particular, generalised linear models (GLM) including logistic regression and models for functional data would be of interest for further theoretical development.

Attempts to exploit the asymptotic prediction ability of the AMP includes Barbier et al. (2019) for high dimensional GLM with general i.i.d. design in the Bayesian and machine learning framework; Emami et al. (2020) investigating generalised errors of the GLM with general Gaussian design in the neutral network framework using the multi-layer AMP algorithm (also see the relevant references mentioned in this paper in the 'Approximate Message Passing' section.) The booming research in the electrical engineering field mostly focuses on advancing the AMP algorithm itself and discussing the estimation accuracy. The subsequent investigations including hypothesis testing, information criterion, goodness-of-fit tests, etc., are of interest for statistical research.

Another promising development relying on the asymptotic characterisation of the AMP algorithm, focuses on high-dimensional logistic regression modeling which is widely used for classification problems.

Working with the analytical expression of the logistic link function, the papers Sur et al. (2019); Sur and Candès (2019); Candès et al. (2020); Zhao et al. (2020) have thorough investigations on (1) the conditions for the existence of the maximum likelihood estimator; (2) the asymptotic distribution of the likelihood ratio test statistic. In classical statistical theories, various properties such as consistency, asymptotic normality, etc., are based on the maximum likelihood estimator. The literature mentioned above on high dimensional logistic regression ignited a possibility of unifying the asymptotic theory in low- and high-dimensional regression models. Future research could generalise the work in Sur et al. (2019); Sur and Candès (2019); Zhao et al. (2020) to the exponential family.

## Acknowledgements

The authors than the Editor, Associate Editor and reviewer for their constructive remarks.
This work was supported by a Postdoc Fellowship of the Research Foundation Flanders, the Research Foundation Flanders and KU Leuven internal fund C16/20/002. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation-Flanders (FWO) and the Flemish Government.

## References

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

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460.

Bayati, M. and Montanari, A. (2011a). The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing. *IEEE Transactions on Information Theory*, 57(2):764–785.

Bayati, M. and Montanari, A. (2011b). The LASSO Risk for Gaussian Matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.

Belloni, A. and Chernozhukov, V. (2011). $l_1$-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.

Bradic, J. (2016). Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing. *Electronic Journal of Statistics*, 10(2):3894–3944.

Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):325–349.

Bradic, J. and Kolar, M. (2017). Uniform inference for high-dimensional quantile regression: linear functionals and regression rank scores. *0.32 preprint.* arXiv:1702.06209.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.

Candès, E. J., Sur, P., et al. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42.

Caner, M. and Kock, A. B. (2018). Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *Journal of Econometrics*, 203(1):143–168.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719.

Donoho, D., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.

Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint.* arXiv:1311.2445.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562.

Emami, M., Sahraee-Ardakan, M., Pandit, P., Rangan, S., and Fletcher, A. (2020). Generalization error of generalized linear models in high dimensions. In *International Conference on Machine Learning*, pages 2892–2901. PMLR.

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

Gordon, Y. (1985). Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289.

Gordon, Y. (1988). On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric aspects of functional analysis*, pages 84–106. Springer.

Gueuning, T. and Claeskens, G. (2018). A high-dimensional focused information criterion. *Scandinavian Journal of Statistics*, 45(1):34–61.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Huang, H. (2020). Asymptotic risk and phase transition of $l_1$-penalized robust estimator. *The Annals of Statistics*, 48(5):3090–3111.

Huber, P. J. (2004). *Robust statistics*, volume 523. John Wiley & Sons.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909.

Khan, J. A., Van Aelst, S., and Zamar, R. H. (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, 102(480):1289–1299.

Koenker, R. (2005). *Quantile Regression.* Econometric Society Monographs. Cambridge University Press.

Koenker, R. and Bassett, J. G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Lee, J., Sun, D., Sun, Y., and Taylor, J. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927.

Lei, L., Bickel, P. J., and El Karoui, N. (2018). Asymptotics for high dimensional regression M-estimates: fixed design results. *Probability Theory and Related Fields*, 172(3-4):983–1079.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., and Anna di Palma, M. (2022). *robustbase: Basic Robust Statistics.* R package version 0.95-0.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

R Core Team (2022). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

signal developers (2014). *signal: Signal processing.*

Sun, Q., Zhou, W., and Fan, J. (2020). Adaptive Huber Regression. *Journal of the American Statistical Association*, 115(529):254–265.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Sur, P., Chen, Y., and Candès, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1):487–558.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628.

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

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

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

Wang, X., Dunson, D., and Leng, C. (2016). Decorrelated feature space partitioning for distributed sparse regression. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 802–810.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768.

Zhao, Q., Sur, P., and Candes, E. J. (2020). The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. *arXiv preprint.* arXiv:2001.09351.

Zhao, T., Kolar, M., and Liu, H. (2014). A general framework for robust testing and confidence regions in high-dimensional quantile regression. *arXiv eprint.* arXiv: 0401062.

Zhao, W., Zhang, F., and Lian, H. (2019). Debiasing and distributed estimation for high-dimensional quantile regression. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2569–2577.

Zhou, J., Claeskens, G., and Bradic, J. (2020). Detangling robustness in high-dimensions: composite versus model-averaged estimation. *Electronic Journal of Statistics*, 14:2551–2599.

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

# Appendix

## A Robust approximate message passing algorithm

For the paper to be self-contained, we briefly revise the main ingredients of this algorithm. See Donoho and Montanari (2016) for more details. To incorporate non-differentiable loss functions, the proximal mapping operator with parameter $b > 0$ is used to adjust the residuals in the algorithm,

$$\text{Prox}(z, b) = \arg\min_{x \in \mathbb{R}} \{b\rho(x) + \frac{1}{2}(x - z)^2\}, \ b > 0;$$

and the effective score function $\widetilde{G}(z; b) = b \cdot \partial\rho(x)|_{x=\text{Prox}(z;b)}$, where $\partial\rho(x) = \{y : \rho(u) \geq \rho(x) + y(u - x), \forall u\}$ is the subgradient at non-differentiable points $x$ and the gradient at differentiable points. To incorporate the sparsity $s$, the rescaled effective score function is defined as $G(z;b) = \delta\omega^{-1}\widetilde{G}(z;b)$.

The RAMP algorithm, with a fixed tuning parameter $\alpha$ and iterations indexed by $t$, starts from $\widehat{\beta}_{(0)} = 0$ and updates iteratively using the following three steps:

**Step 1 Adjust the residuals:**

$$z_{(t)} = Y - X\widehat{\beta}_{(t)} + n^{-1}G(z_{(t-1)}; b_{(t-1)}) \sum_{j=1}^{p} I\left\{\eta\big(\widehat{\beta}_{(t-1),j} + X_j^\top G(z_{(t-1)}; b_{(t-1)}); \theta_{t-1}\big) \neq 0\right\};$$

$I$ denotes the indicator function, and the soft-thresholding function $\eta$ is defined in (5).

**Step 2 Effective score:** choose the scalar $b_{(t)}$ such that the empirical average of the rescaled effective score function $G(z;b)$ has slope 1; update the tuning parameter $\theta_{(t)} = \alpha\bar{\zeta}_{\text{emp},(t)}$ with a limit version, when $n, p \to \infty$, denoted by $\bar{\zeta}^2_{(t)} = E[G(\varepsilon + \bar{\sigma}_{(t)}Z; b_{(t)})]$, with $\bar{\sigma}^2_{(t)} = \delta^{-1}E[(\eta(B_0 + \bar{\zeta}_{(t-1)}Z; \theta_{(t-1)}) - B_0)^2]$, where $Z \sim N(0,1)$, for $B_0$ see Assumption (A2), and for $\varepsilon$, see (A5).

**Step 3 Estimation:** update the estimator of $\beta$

$$\widehat{\beta}_{(t+1)} = \eta(\widetilde{\beta}_{(t)}; \theta_{(t)}), \text{ where } \widetilde{\beta}_{(t)} = \widehat{\beta}_{(t)} + X^\top G(z_{(t)}; b_{(t)}).$$

Since a bias is introduced by applying the soft-thresholding function $\eta$ in Step 3, the estimator $\widetilde{\beta}_{(t)}$ which is obtained before applying the thresholding can be interpreted as a debiased estimator. This estimator is of main interest in this paper.

# B Additional simulation results for the $\ell_1$-regularised quantile estimator

See Figures 2 and 3.

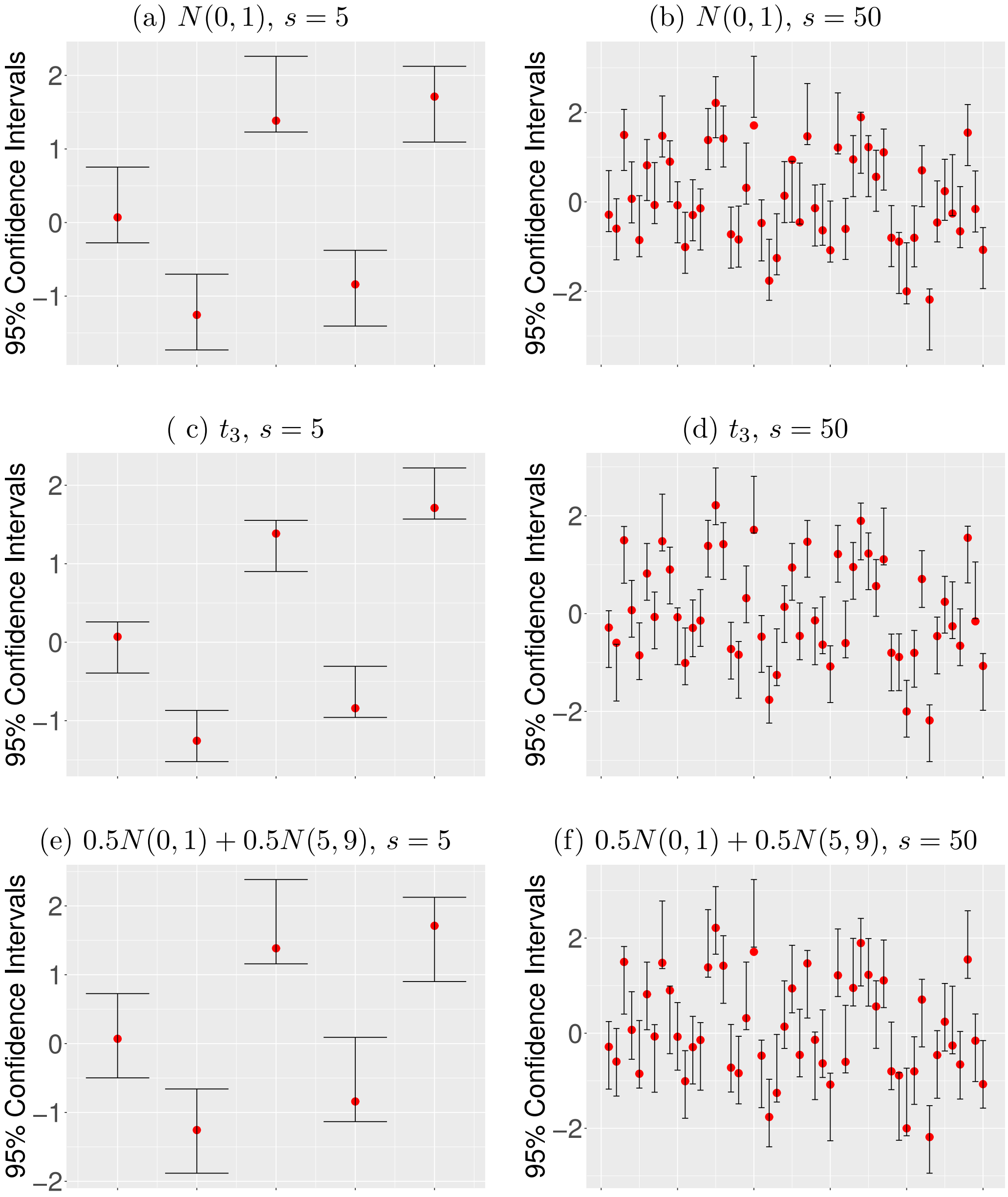
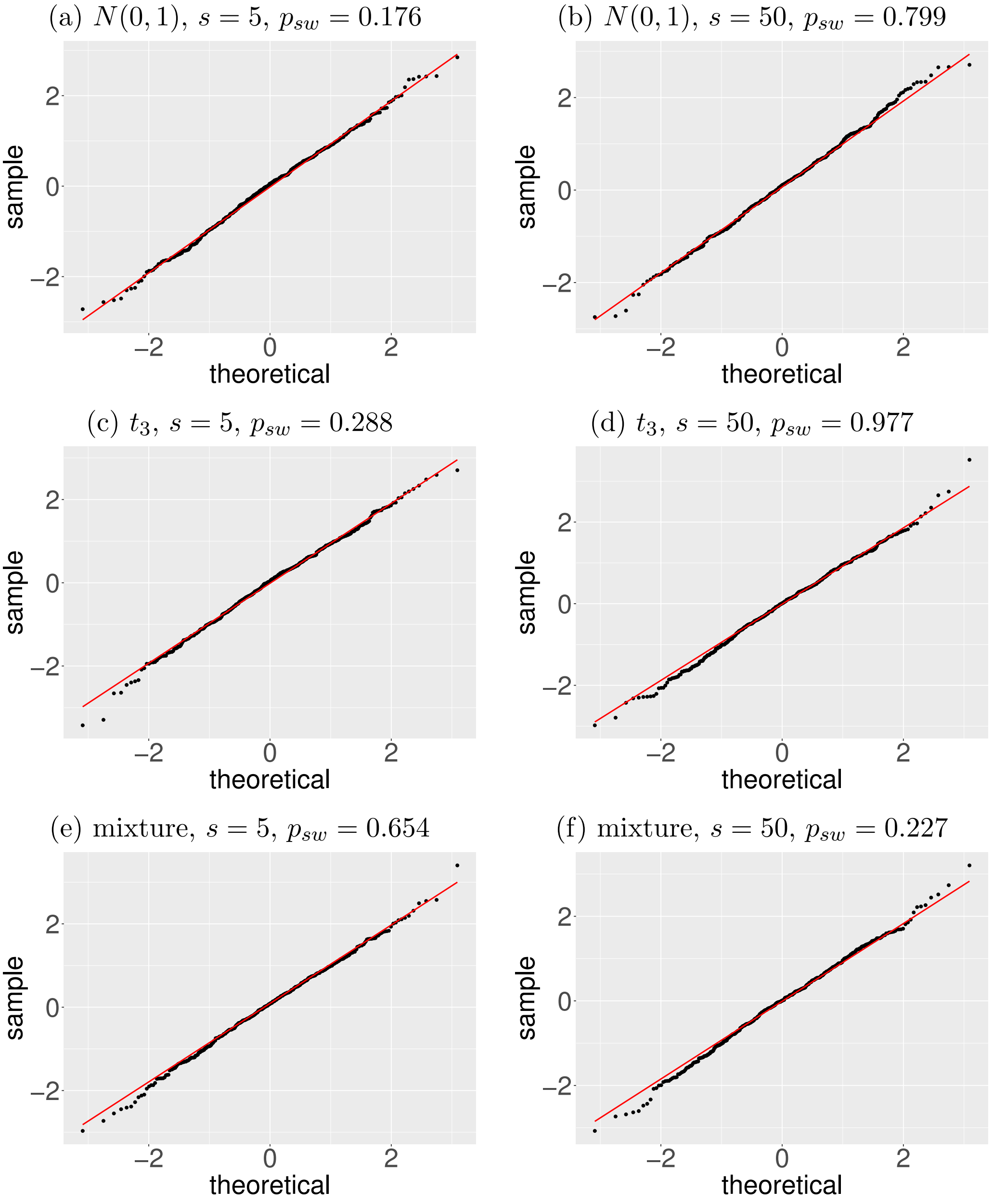


Figure 2: Quantile estimator. Example 95% confidence intervals plots of non-zero components of $\beta$; the non-zero values are randomly generated from $N(0,1)$ using the seed number 5. Plots for $s = 5$ are in the left column and for $s = 50$ are in the right column. Each row corresponds to plots for one error distribution, i.e., (a,b) $N(0,1)$ (c,d) $t_3$, (e,f) $0.5N(0,1) + 0.5N(5,9)$.



(a) $N(0,1)$, $s = 5$, $p_{sw} = 0.176$

(b) $N(0,1)$, $s = 50$, $p_{sw} = 0.799$

(c) $t_3$, $s = 5$, $p_{sw} = 0.288$

(d) $t_3$, $s = 50$, $p_{sw} = 0.977$

(e) mixture, $s = 5$, $p_{sw} = 0.654$

(f) mixture, $s = 50$, $p_{sw} = 0.227$

Figure 3: Quantile estimator. Example QQ-plots of $T_j(\beta_j)$, $j = 1, \ldots, p$; non-zero components of $\beta$ are randomly generated from $N(0,1)$ using the seed number 5. Plots for $s = 5$ are on the left column and for $s = 50$ are on the right column. Each row corresponds to plots for one error distribution, i.e., (a,b) $N(0,1)$, (c,d) $t_3$, (e,f) $0.5N(0,1) + 0.5N(5,9)$. The $p$-values of the Shapiro-Wilk test ($p_{sw}$) for each setting are presented at the caption of each plot.