Statistica Sinica Preprint No: SS-2022-0391	
Title	Penalized Regression with Multiple Loss Functions and Variable Selection by Voting
Manuscript ID	SS-2022-0391
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0391
Complete List of Authors	Guorong Dai, Ursula U. Müller and Raymond J. Carroll
Corresponding Authors	Guorong Dai
E-mails	guorongdai@fudan.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

PENALIZED REGRESSION WITH MULTIPLE LOSS FUNCTIONS AND VARIABLE SELECTION BY VOTING

Guorong Dai [a], Ursula U. Müller [b] and Raymond J. Carroll [b]

[a] *Fudan University and* [b] *Texas A&M University*

Abstract: We consider a sparse linear model with a fixed design matrix in a high dimensional scenario. We introduce a new variable selection procedure called "voting", which combines the results from multiple regression models with different penalized loss functions to select the relevant predictors. A predictor is included in the final model if it receives enough votes, i.e. is selected by most of the individual models. By employing multiple different loss functions our method takes various properties of the error distribution into account. This is in contrast to the standard penalized regression approach, which typically relies on just one criterion. When that single criterion is not met the standard approach is likely to fail, whereas our method is still able to identify the underlying sparse model. Working with the voting procedure reduces the number of predictors that are incorrectly selected, which simplifies the structure and improves the interpretability of the fitted model. We prove model selection consistency and illustrate the advantages of our method numerically using simulated and real data sets.

Key words and phrases: High dimensional data; Linear model; Model selection consistency; Sparse estimators.

1. Introduction

In the past few decades variable selection has attracted much attention in the statistical community because of its usefulness in analyzing modern

data sets in which the number of features can be very high, often greatly exceeding the number of observations. Variable selection is used to determine the covariates that have an effect on an outcome and that should be included in the model as relevant predictors. This can substantially improve the simplicity and interpretability of a fitted statistical model. We refer interested readers to Desboulets (2018) for a detailed overview of this topic in a variety of regression settings. In this article we consider the variable selection problem in a high dimensional linear model with a deterministic design matrix, where the number of predictors can exceed the sample size. We further assume that the model is sparse, i.e. only a fraction of the predictors significantly affects the response. Our goal is to identify this fraction of important predictors and to exclude those with no influence.

One popular approach for variable selection in a linear model is penalized regression. It yields a sparse estimator for the parameter vector through minimizing an objective function, which typically consists of two parts: a loss function and a penalty term. Various types of penalties, such as the L_1 penalty (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), and the weighted L_1 penalty (Zou, 2006; Zou and Li, 2008), have been applied to the quadratic loss function. Also widely used has been the quantile loss function (check function),

which, because of its robustness, is the preferred choice when the error distribution is heavy-tailed; see Wu and Liu (2009), Wang et al. (2012) and Fan et al. (2014), among others.

In addition to these articles working with the quadratic or quantile loss function, Zou and Yuan (2008) introduced a composite quantile regression approach, which combines multiple quantile loss functions into one composite loss function by taking the sum. They use the adaptive Lasso penalty to detect sparsity. An extension of their results to an ultrahigh dimensional scenario is considered in Gu and Zou (2020). Bradic et al. (2011) developed a composite quasi-likelihood function, which approximates the log-likelihood function of the random error by a weighted linear combination of convex loss functions, and adopts a weighted L_1 penalty. The use of composite loss functions in Zou and Yuan (2008), Gu and Zou (2020) and Bradic et al. (2011) improves the quality of the estimation. However, all the penalized regression methods mentioned above share a common limitation in variable selection: they work well only if certain assumptions on the error distribution are satisfied. If not, the precision of the selection results can be considerably impaired. For example, one should not expect the quadratic loss function to perform satisfactorily when the error distribution has an infinite first moment. Section 2.1 includes a detailed discussion of the lim-

itations of loss functions for certain classes of error distributions. Since the error distribution is usually unknown in practice, one cannot always choose loss functions that are suitable for the underlying data generating mechanism.

To overcome the problems of standard penalized regression based on one single loss function and improve the quality of variable selection, we propose a selection process that combines *multiple* sparse estimators calculated with different loss functions by a "voting procedure". Only predictors that have received the most votes, i.e. are selected by the majority of these estimators, are included in the final model. By using different loss functions we take multiple properties of the error distribution into account, not just one feature such as the existence of a certain moment. When some of these properties are violated so that some of the estimators fail to select the correct model, our proposed process is still able to identify the important predictors; see Proposition 1 and the following remark for details. This is the key advantage of our method over existing approaches that rely on a single property. In Theorems 1 and 2 we establish model selection consistency of our proposed process and present a probabilistic bound for two popular finite-sample performance measures when weighted L_1 penalties are used. This provides the theoretical foundation for our method. Numerical studies

in Section 4 demonstrate a further advantage: even if the error distribution is well behaved and the standard penalized regression methods should work well, our method can considerably reduce the number of incorrectly selected unimportant predictors, yielding simpler and more interpretable fitted models.

The rest of this article is organized as follows. In Section 2 we present our model and some notation, followed by a discussion of the limitations of selection methods that use one single loss functions (Section 2.1). Then, in Section 2.2, we introduce our selection approach, which is based on multiple estimators derived from different loss functions with general penalty terms, and a voting step. In Section 2.3 we prove model selection consistency and discuss theoretical properties of our method for a special case, namely when weighted L_1 penalties are used. Section 3 details the implementation of our approach, including the algorithm, tuning parameter selection and post-selection estimation. In Section 4 we present the results from a number of comprehensive numerical studies. Section 5 consists of concluding remarks. All technical details, including the assumptions and the proofs of the theoretical results from Section 2, additional numerical results and a discussion on the extension of our method are deferred to the Supplementary Material.

2. Variable selection with multiple loss functions

In the following, for convenience of notation, we let the lower case letter c represent a generic positive constant. The symbol $I(\cdot)$ stands for the indicator function. For two positive sequences a_n and b_n, the expression $a_n \gg b_n$ means $\limsup_{n\to\infty}(b_n/a_n) = 0$. For a $(p+1)$-dimensional vector $v = (v_0, v_1, \ldots, v_p)^{\mathrm{T}}$ let $\mathrm{supp}(v)$ denote the index set of the non-zero components of v, $\mathrm{supp}(v) = \{j \in \{1, \ldots, p\} : v_j \neq 0\}$. The notation $|\mathcal{S}|$ denotes the cardinality of a set $\mathcal{S} \subset \{1, \ldots, p\}$ and $\mathcal{S}^c$ its complement, $\mathcal{S}^c = \{1, \ldots, p\} \backslash \mathcal{S}$. Lastly, we write $\mathrm{N}(\mu, \sigma^2)$ for a univariate normal distribution with mean μ and variance σ^2.

In order to introduce the linear model formally, we consider

$$Y = X\vartheta + \varepsilon, \tag{2.1}$$

where $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is an n-dimensional vector of responses, $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ is a deterministic $n \times (p+1)$ design matrix of predictors with rows $X_i = (1, X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ for $i = 1, \ldots, n$, $\vartheta = (\vartheta_0, \vartheta_1, \ldots, \vartheta_p)^{\mathrm{T}}$ is a $(p+1)$-dimensional vector of parameters, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is an n-dimensional vector of independent and identically distributed random errors. To guarantee identifiability of the intercept term ϑ_0, we set the median of the errors ε_i equal to zero. The number of predictors $p = p_n$

tends to infinity as n increases and is allowed to exceed the sample size n. We suppose that the parameter vector ϑ in (2.1) is sparse, i.e. there exists a set $\mathcal{Q} \subset \{1, \ldots, p\}$ such that $\vartheta_j \neq 0$ for $j \in \mathcal{Q}$ and $\vartheta_j = 0$ for $j \in \mathcal{Q}^c$. Without loss of generality let $\mathcal{Q} = \{1, 2, \ldots, q\}$ for some sequence $q = q_n$ of positive integers such that $q < p$.

2.1 Limitations of methods based on single loss functions

To identify the set $\mathcal{Q}$ in model (1), we will use *multiple* loss functions. This is in contrast to many existing methods that are based on a *single* loss function $\ell(\cdot)$, which yields estimators of the form

$$\widehat{\vartheta}_{\text{SIN}} = \arg\min_{\theta}\{\textstyle\sum_{i=1}^{n} \ell(Y_i - X_i^{\mathrm{T}}\theta) + \phi(\theta)\} \tag{2.2}$$

with parameter vector $\theta = (\theta_0, \theta_1, \ldots, \theta_p)^{\mathrm{T}} \in \mathbb{R}^{p+1}$ and penalty function $\phi(\cdot)$, e.g. the L_1 penalty (Tibshirani, 1996) and the SCAD penalty (Fan and Li, 2001).

Ideally, if the error distribution in model (2.1) is known, one could let $\ell(\cdot)$ in (2.2) equal the negative log-likelihood of ε_1. Fan and Li (2001) have proven that a penalized maximum likelihood estimator based on the SCAD penalty ensures, under suitable conditions, model selection consistency, i.e. $\text{pr}\{\text{supp}(\widehat{\vartheta}_{\text{SIN}}) = \mathcal{Q}\} \to 1$. However, in practice the distribution of ε_1 is usually hard to specify, so one may just use a fixed loss function $\ell(\cdot)$, such

as $\ell(x) = x^2/2$. On the other hand, desirable properties of an estimator $\widehat{\vartheta}_{\text{SIN}}$, which is based on a certain loss function $\ell(\cdot)$, typically rely on the validity of some conditions concerning the error distribution. The quadratic loss function, for example, needs the existence of $E(\varepsilon_1^2)$. The quantile loss or check function (Koenker, 2005) requires that the density $f(\cdot)$ of ε_1 is positive at the considered quantile level. If the composite quantile loss function (Zou and Yuan, 2008; Gu and Zou, 2020) is used, the error density $f(\cdot)$ needs to be bounded and differentiable over its support. The performance of the composite loss function proposed by Bradic et al. (2011) depends on the quality of an initial estimator, which is calculated from penalized least squares regression in that article and requires $E(\varepsilon_1^2) < \infty$.

Violation of these assumptions can substantially impair the performance of $\widehat{\vartheta}_{\text{SIN}}$, leading to *inconsistency* in model selection. An example is provided in Section 3 of Fan et al. (2014) for a simplified version of model (2.1) with $\vartheta_1 = \ldots = \vartheta_q > 0$. When $\ell(\cdot)$ is the quadratic loss function and $\phi(\cdot)$ the L_1 penalty, those authors show that for heavy-tailed error distributions with infinite first moment, e.g. the Cauchy distribution, the estimator $\widehat{\vartheta}_{\text{SIN}}$ does not possess model selection consistency, that is, $\text{pr}\{\text{supp}(\widehat{\vartheta}_{\text{SIN}}) = \mathcal{Q}\} < \exp(-c)$ for some constant $c > 0$, unless $\vartheta_1 \gg n^{1/4}$. This is quite a stringent requirement that excludes cases where ϑ_1 is finite,

and is therefore unrealistic in practice; see Section 3 and the Supplementary Material of Fan et al. (2014) for the proof and further details.

2.2 Proposed method

Motivated by the above-mentioned limitations of variable selection strategies that use only a single loss function, we have developed a selection process that is based on multiple loss functions. This means that the selection result depends on multiple properties of the error distribution and not just one single feature, e.g. the value of density at the median. More importantly, when some of these properties fail to hold, the proposed process should still be able to recover the true model consistently. An intuitive and illustrative example is the linear model (2.1) with $\varepsilon_1 \sim 0.5\mathrm{N}(\text{-}5, 1) + 0.5\mathrm{N}(5, 1)$, a location mixture normal distribution. The median point of ε_1 has an extremely low density (less than $10^{\text{-}5}$) and an estimator that uses the absolute loss function $\ell(x) = |x|/2$ would therefore not work well. However, the values of the error density $f(\cdot)$ at most other quantiles are still significantly greater than zero: The density at the quantile level $\tau \in \{0.1, \ldots, 0.4, 0.6, \ldots, 0.9\}$ is above 0.1. Hence, for $k \in \{1, \ldots, 9\}\backslash\{5\}$, we can expect an estimator $\widehat{\vartheta}_k$ based on the check function $\ell_k(x) = x\{k/10 - I(x < 0)\}$ to perform satisfactorily in terms of variable selection. Then, if we include in our final

model only the predictors identified by most of the estimators $\{\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_9\}$ as influential, it is very likely that the undesirable influence of $\widehat{\vartheta}_5$, using the absolute loss function $\ell_5(x) = |x|/2$, will be reduced to an acceptable level. A numerical analysis of this example is provided in Section 4.1. Such a strategy that considers multiple estimators coming from different loss functions can be applied to a variety of scenarios.

We now formulate this idea and propose our variable selection method. For some fixed integer $K \geq 1$, we use K *different* convex loss functions $\{\ell_1(\cdot), \ldots, \ell_K(\cdot)\}$ to obtain estimators

$$\widehat{\vartheta}_k = (\widehat{\vartheta}_{k0}, \widehat{\vartheta}_{k1}, \ldots, \widehat{\vartheta}_{kp})^{\mathrm{T}} = \arg\min_{\theta}\{\textstyle\sum_{i=1}^{n} \ell_k(Y_i - X_i^{\mathrm{T}}\theta) + \phi_k(\theta)\} \quad (2.3)$$

for $k = 1, \ldots, K$, where $\theta = (\theta_0, \theta_1, \ldots, \theta_p)^{\mathrm{T}} \in \mathbb{R}^{p+1}$ and $\phi_k(\cdot)$ is a penalty function. We do not specify the form of $\ell_k(\cdot)$, allowing it to be any convex loss function such as the quadratic loss function or the check function. In Section 3 we study in detail the special case where $\ell_k(\cdot)$ is the quantile loss function.

Write $\widehat{\mathcal{Q}}_k = \mathrm{supp}(\widehat{\vartheta}_k)$ for the index set of the non-zero components of $\widehat{\vartheta}_k$ in (2.3), excluding the intercept term, which is usually not of interest in variable selection. Then, for some threshold $\alpha \in \{1, \ldots, K\}$, our final estimator $\widehat{\mathcal{Q}}(\alpha)$ for the index set $\mathcal{Q}$ of the relevant predictors is determined

by a *voting* procedure, i.e.

$$\widehat{\mathcal{Q}}(\alpha) = \{j \in \{1, \ldots, p\} : \textstyle\sum_{k=1}^{K} I(j \in \widehat{\mathcal{Q}}_k) \geq \alpha\}. \tag{2.4}$$

This means that the jth component ϑ_j of ϑ is included in the final estimator $\widehat{\mathcal{Q}}(\alpha)$, only if at least α of the K estimates $\{\widehat{\vartheta}_{1j}, \ldots, \widehat{\vartheta}_{Kj}\}$ are non-zero, i.e. it receives α or more *votes*.

In the above, the K estimated parameter vectors $\{\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_K\}$ result from minimizing different objective functions in (2.3) separately. Then they are combined, with the help of the voting step (2.4), to yield the final result $\widehat{\mathcal{Q}}(\alpha)$. This process may at first glance appear similar to the "stability selection" method proposed by Meinshausen and Bühlmann (2010), but it is substantially different: those authors use multiple random subsets of the data $\{(X_i, Y_i) : i = 1, \ldots, n\}$ to calculate a set of sparse estimators. If a predictor is identified as relevant by sufficiently many of these estimates, it is included in the model. In contrast to our article, the estimators in Meinshausen and Bühlmann (2010) were all computed via penalized least squares regression, i.e. using a *single* loss function, and the selection consistency was shown only for normal errors. The method is apparently likely to fail in scenarios with heavy-tailed error distributions, as discussed in Section 2.1. Illustrations in Section 4.1 will further demonstrate the advantage of our method over this type of resampling-based approach that depends

on just a single loss function.

In the voting step (2.4), even if some of the selection results $\{\widehat{\mathcal{Q}}_1, \ldots, \widehat{\mathcal{Q}}_K\}$ deviate significantly from the true model $\mathcal{Q}$, it is still possible that the non-zero parameters will receive α or more votes, while the zero ones will not, so that the final estimator $\widehat{\mathcal{Q}}(\alpha)$ performs well. In other words, our proposed method only requires a certain proportion of $\{\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_K\}$ to be model selection consistent. This property is straightforward to see. We state this formally in the following proposition and refer to Section S2 in the Supplementary Material for a short proof.

Proposition 1. *Consider a threshold* $\alpha \in \{1, \ldots, K\}$ *and let* $K^* = \max\{\alpha, K - \alpha + 1\}$. *Suppose, without loss of generality, that the first* K^* *of the* K *estimators given in* (2.3) *are model selection consistent, i.e.* $\mathrm{pr}(\widehat{\mathcal{Q}}_k = \mathcal{Q}) \to 1$ *for* $k = 1, \ldots, K^*$. *Then the vote estimator* $\widehat{\mathcal{Q}}(\alpha)$ *from* (2.4) *is also consistent, i.e.* $\mathrm{pr}\{\widehat{\mathcal{Q}}(\alpha) = \mathcal{Q}\} \to 1$.

Remark 1. Proposition 1 provides a sufficient condition for $\widehat{\mathcal{Q}}(\alpha)$ to be selection consistent by requiring $K^* = \max\{\alpha, K-\alpha+1\}$ of the K estimators $\{\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_K\}$ are model selection consistent. It is clear that setting

$$\alpha = \lceil K/2 \rceil \text{ or } \lceil (K+1)/2 \rceil \tag{2.5}$$

in (2.4) is optimal in the sense of minimizing K^* and thus relaxing the

requirement in Proposition 1 to the greatest extent. It means, in particular, that for this choice of α at least half of the estimators $\{\widehat{\mathcal{Q}}_1, \ldots, \widehat{\mathcal{Q}}_K\}$ must be consistent. A smaller (or larger) value of α would necessitate a larger number of consistent estimators. Note that the optimal α is unique if K is an odd number. The numerical results in Section 4 show that α close to $\lceil K/2 \rceil$ or $\lceil (K+1)/2 \rceil$, e.g. $\alpha = 5$ or 6 when $K = 9$, works quite well in practice, so the optimal α in (2.5) can indeed be recommended.

The condition in Proposition 1 may be stronger than necessary. For example, in a model with $\mathcal{Q} = \{1, 2\}$, if we consider $K = 2$ estimators such that $\widehat{\mathcal{Q}}_1 = \{1\}$ and $\widehat{\mathcal{Q}}_2 = \{2\}$ with probability approaching one, the condition in Proposition 1 means that both of the two estimators $\{\widehat{\mathcal{Q}}_1, \widehat{\mathcal{Q}}_2\}$ should be consistent. This is obviously violated by the fact that $\widehat{\mathcal{Q}}_1$ and $\widehat{\mathcal{Q}}_2$ both miss an important predictor, but $\widehat{\mathcal{Q}}(\alpha)$ with threshold $\alpha = 1$ is still consistent since each of the non-zero parameter $\{\vartheta_1, \vartheta_2\}$ receives one vote, i.e. $\widehat{\mathcal{Q}}(1) = \{1, 2\} = \mathcal{Q}$. Nevertheless, Proposition 1 still demonstrates the key feature of our method, i.e. being consistent in model selection even if some of the estimators $\{\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_K\}$ are inconsistent.

2.3 Variable selection with a weighted L_1 penalty

In the following we will verify the condition in Proposition 1 for a special case of the penalized estimator in (2.3), namely when the penalty function $\phi_k(\theta)$ is a weighted L_1 penalty, $\phi_k(\theta) = n\lambda_{n,k}\sum_{j=1}^{p} d_{kj}|\theta_j|$, with weights $d_{kj} > 0$ and tuning parameter $\lambda_{n,k} > 0$ ($k = 1, \ldots, K$; $j = 1, \ldots, p$). Here no penalty is imposed on the intercept term θ_0, since it is usually not of interest in variable selection. This type of penalty function is frequently used in the literature (Zou and Li, 2008; Bradic et al., 2011; Fan et al., 2014). We focus on this penalty, since it is convenient in both theory and implementation due to its convex nature. There are other reasonable choices for $\phi_k(\cdot)$, e.g. the SCAD penalty and the minimax concave penalty (Zhang, 2010), which can be treated in a similar way.

With $\phi_k(\cdot)$ the weighted L_1 penalty, the estimators in (2.3) become

$$\widehat{\vartheta}_k = \arg\min_\theta \{\textstyle\sum_{i=1}^{n} \ell_k(Y_i - X_i^{\mathrm{T}}\theta) + n\lambda_{n,k}\sum_{j=1}^{p} d_{kj}|\theta_j|\} \tag{2.6}$$

for $k = 1, \ldots, K$. Recall that $\widehat{\mathcal{Q}}_k$ contains the indices of the non-zero components of $\widehat{\vartheta}_k$ (without the intercept term), i.e. $\widehat{\mathcal{Q}}_k = \mathrm{supp}(\widehat{\vartheta}_k)$ for $k = 1, \ldots, K$. As stated in Proposition 1, to ensure our final vote estimator $\widehat{\mathcal{Q}}(\alpha)$ given in (2.4) satisfies $\mathrm{pr}\{\widehat{\mathcal{Q}}(\alpha) = \mathcal{Q}\} \to 1$, we assume without loss of generality that the first $K^* = \max\{\alpha, K - \alpha + 1\}$ estimators $\{\widehat{\mathcal{Q}}_1, \ldots, \widehat{\mathcal{Q}}_{K^*}\}$

are consistent for $\mathcal{Q}$. To guarantee that this assumption holds true for $\{\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_{K^*}\}$ defined in (6), we need to assume some regularity conditions on (i) the properties of the loss functions $\{\ell_1(\cdot), \ldots, \ell_{K^*}(\cdot)\}$ including their unbiasedness and smoothness, (ii) the growth rates of the full model size p and the number q of non-zero parameters, (iii) the behavior of the design matrix X, and (iv) the rates of the tuning parameter $\lambda_{n,k}$ and weights d_{kj} in the weighted L_1 penalty ($j = 1, \ldots, p$; $k = 1, \ldots, K^*$). In particular we consider *ultrahigh* dimensional scenarios, where $p = p_n$ may increase exponentially with n, while $q = q_n$ is only required to be of lower order than n. Also, the entries of X are allowed to diverge. For reasons of clarity, we provide the complete list of regularity conditions, along with detailed explanations, in Section S1 of the Supplementary Material. Assumptions of this type are common in the literature on high dimensional regression; see, e.g. Bradic et al. (2011), Wang et al. (2012) and Fan et al. (2014). We impose no conditions on the remaining $(K - K^*)$ estimators $\{\widehat{\vartheta}_{K^*+1}, \ldots, \widehat{\vartheta}_K\}$.

In the following Theorem 1 we first establish that, with probability tending to one, the estimator $\widehat{\vartheta}_k$ equals the "biased oracle estimator", given by $\widehat{\vartheta}_k^o = (\widehat{\vartheta}_{k0}^o, \widehat{\vartheta}_{k1}^o, \ldots, \widehat{\vartheta}_{kp}^o)^{\mathrm{T}} =$

$$\arg\min_{\theta \in \Theta} \{\textstyle\sum_{i=1}^n \ell_k(Y_i - X_i^{\mathrm{T}}\theta) + n\lambda_{n,k} \sum_{j=1}^p d_{kj}|\theta_j|\} \tag{2.7}$$

for $k = 1, \ldots, K^*$, which is the minimizer of the penalized objective function

in the set $\Theta = \{\theta = (\theta_0, \theta_1, \ldots, \theta_p)^{\mathrm{T}} \in \mathbb{R}^{p+1} : \theta_j = 0 \text{ for } j \in \mathcal{Q}^c\}$. This indicates that $\widehat{\vartheta}_k$ $(k = 1, \ldots, K^*)$ can successfully exclude all the irrelevant predictors. In the second part of Theorem 1, we show under the assumption that the values of non-zero parameters do not decay too quickly to be detected, each of $\{\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_{K^*}\}$ recovers the index set $\mathcal{Q}$ of the non-zero parameters with probability tending to one. Hence the final vote estimator $\widehat{\mathcal{Q}}(\alpha)$ possesses model selection consistency, according to Proposition 1.

Theorem 1. *Suppose Assumptions 1-7 given in Section S1 of the Supplementary Material are satisfied. Then, for $k = 1, \ldots, K^*$, with $K^* = \max\{\alpha, K - \alpha + 1\}$ as defined in Proposition 1,*

$$\mathrm{pr}(\widehat{\vartheta}_k = \widehat{\vartheta}_k^o) \;\geq\; 1 - 2(p-q)\exp(-c\, z_n^2) \;\to\; 1 \quad (n \to \infty). \tag{2.8}$$

Here $\widehat{\vartheta}_k^o$ is the estimator defined in (2.7), $c > 0$ is some constant and $z_n = n^{(\nu_0 - 2\nu_1)_+/2 + \nu_2}$ with constants $\{\nu_0, \nu_1, \nu_2\}$ specified in Assumptions 3–5 of Section S1 in the Supplementary Material. Further, as long as the magnitudes of the non-zero parameters do not decay too fast, i.e. they satisfy $\min_{j \in \mathcal{Q}} |\vartheta_j| \gg (q/n)^{1/2}$, we have $\mathrm{pr}(\widehat{\mathcal{Q}}_k = \mathcal{Q}) \to 1$ for $k = 1, \ldots, K^*$, which implies the conclusion of Proposition 1, i.e. $\mathrm{pr}\{\widehat{\mathcal{Q}}(\alpha) = \mathcal{Q}\} \to 1$.

Remark 2. Theorem 1 provides the asymptotic justification of our variable selection approach. In the first conclusion (2.8), the constants ν_0 and

ν_1 in the term z_n respectively control the divergence rates of the non-zero parameter size in the model (i.e. $q = O(n^{\nu_0})$ with $\nu_0 \in (0, 1)$) and of the correlation coefficients between the relevant and irrelevant predictors, while ν_2 regulates the magnitude of the (possibly divergent) maximum absolute entry in the design matrix X. Details are provided in Section S1 of the Supplementary Material; see Assumptions 3–5 and the explanations thereafter. For the probability lower bound in (2.8) to be meaningful, i.e. tending to one, we need a restriction on the model size p. We have

$$1 - 2(p - q)\exp(-c\, z_n^2) \;\geq\; 1 - 2\exp(\log p - c\, z_n^2), \tag{2.9}$$

which tends to one if $z_n^2 = n^{(\nu_0 - 2\nu_1)_+ + 2\nu_2} \gg \log p$. This is satisfied by Assumptions 3 and 5 in Section S1 of the Supplementary Material, which guarantee $\log p = O(n^\kappa)$ for some $\kappa < (\nu_0 - 2\nu_1)_+ + 2\nu_2$, and therefore

$$z_n^2 \;=\; n^{(\nu_0 - 2\nu_1)_+ + 2\nu_2} \;\gg\; n^\kappa \;\geq\; c \log p \;\text{ for some constant } c > 0.$$

This shows $1 - 2\exp(\log p - c\, z_n^2) \to 1$, i.e. $\mathrm{pr}(\widehat{\vartheta}_k = \widehat{\vartheta}_k^o)$ approaches one exponentially fast ($k = 1, \ldots, K^*$).

To establish the consistency of $\widehat{\mathcal{Q}}_k$, we need, in addition to the assumptions in Section S1 of the Supplementary Material, the condition $\min_{j \in \mathcal{Q}} |\vartheta_j| \gg (q/n)^{1/2}$ on the size of the non-zero parameters. This type of "smallest signal condition" ensures that non-zero parameters can be de-

tected and is common in the literature on high dimensional variable selection; see Wang et al. (2012), Fan et al. (2014) and Gao and Carroll (2017), among others. Our assumption is weaker than many of its counterparts, e.g. Wang et al. (2012) ($\min_{j\in\mathcal{Q}}|\vartheta_j| \gg (q^2/n)^{1/2}$) and Gao and Carroll (2017) ($\min_{j\in\mathcal{Q}}|\vartheta_j| \gg (q^5/n)^{1/2}$), with a sequence $q = q_n$ that is allowed to diverge.

Remark 3. Properties of the weighted L_1 penalty in (2.6) have been widely investigated in the penalized regression literature. For example, Zou and Li (2008) prove its model selection consistency in the context of low-dimensional likelihood models. Bradic et al. (2011) use the weighted L_1 penalty to construct a penalized composite quasi-likelihood estimator, for which they obtain a probability lower bound analogous to (2.8). However, they do not show that their estimator is able to identify the relevant predictors, thus leaving the selection consistency unclear. In Fan et al. (2014), the weighted L_1 penalty is combined with the quantile loss function. Model selection consistency is attained when $q = o(n^{1/2})$ (Condition 2 in Fan et al. (2014)). This is stronger than our requirements on the model sparsity that are specified in Assumptions 3–5. In contrast to the above-mentioned articles, Theorem 1 establishes the model selection consistency of the weighted L_1 penalty under fairly weak and reasonable conditions, allowing for ultra-high dimensional data and a general loss function. Besides this asymptotic

result we also show that our method performs well in finite samples; see Theorem 2 and our remarks preceding it.

Remark 4. Theorem 1 guarantees that the voting procedure (2.4) can identify the set $\mathcal{Q}$ with probability approaching one, even if some (the last $(K - K^*)$) of the estimators $\{\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_K\}$ fail. We investigate the finite-sample properties of our estimator in Section 4, where we conduct a number of comprehensive numerical studies. The results in that section show that, even if ε_1 follows a common distribution, e.g. normal or double exponential distribution, our approach, based on multiple loss functions and the voting procedure, still outperforms the standard methods, such as least squares/absolute deviations regression, which also attain selection consistency when the error distribution is well-behaved.

In addition to the asymptotic result established in Theorem 1, we also study finite-sample properties of our selection method. We assess its performance by considering the F-measure and the G-measure, defined as

$$F(\mathcal{S}) = \frac{2|\mathcal{S} \cap \mathcal{Q}|}{|\mathcal{S}| + |\mathcal{Q}|} \quad \text{and} \quad G(\mathcal{S}) = \frac{|\mathcal{S} \cap \mathcal{Q}|}{(|\mathcal{S}| \times |\mathcal{Q}|)^{1/2}} \quad \text{for any } \mathcal{S} \subset \{1, \ldots, p\},$$

with the convention that $0/0 = 0$ (Nan and Yang, 2014; Yu et al., 2022). Obviously, these two measures take values in $[0, 1]$. Large values indicate a set $\mathcal{S}$ is close to the true index set $\mathcal{Q}$ of the relevant predictors. Theorem 2

below provides a probabilistic bound for $F\{\widehat{\mathcal{Q}}(\alpha)\}$ and $G\{\widehat{\mathcal{Q}}(\alpha)\}$ under the assumptions listed in the Supplementary Material.

Theorem 2. *Without loss of generality let* $|\vartheta_1| \geq |\vartheta_2| \geq \ldots \geq |\vartheta_q| > 0$. *Suppose Assumptions 1-7 in Section S1 of the Supplementary Material are satisfied. For* $s \in \{1, \ldots, q\}$, *if the penalty on* $\{\vartheta_1, \ldots, \vartheta_q\}$ *is sufficiently small, i.e.* $\lambda_{n,k}(\sum_{j=1}^{q} d_{kj}^2)^{1/2} < M_1 \xi |\vartheta_s|$ *for some* $\xi \in (0, 1)$, *with* M_1 *the constant specified in Assumption 4, then*

$$\begin{aligned} &\mathrm{pr}(\min[F\{\widehat{\mathcal{Q}}(\alpha)\}, G\{\widehat{\mathcal{Q}}(\alpha)\}] \geq s/q) \\ &\geq 1 - 2K^*\{(p-q)\exp(-c_1 z_n^2) + q\exp(-c_2 n\vartheta_s^2/q)\}, \end{aligned} \tag{2.10}$$

where $\{c_1, c_2\}$ *are some positive constants,* z_n *is the sequence specified in Theorem 1 and* $K^* = \max\{\alpha, K - \alpha + 1\}$ *as defined in Proposition 1.*

Remark 5. The two terms $2(p-q)\exp(-c_1 z_n^2)$ and $2q\exp(-c_2 n\vartheta_s^2/q)$ in (2.10) correspond to the probabilities that the kth estimator $\widehat{\vartheta}_k$ successfully excludes all unimportant predictors and identifies the first s relevant ones: For the events $A_k = \cap_{j=q+1}^{p}\{\widehat{\vartheta}_{kj} = 0\}$ and $B_k(s) = \cap_{j=1}^{s}\{\widehat{\vartheta}_{kj} \neq 0\}$, Theorem 1 states that $\mathrm{pr}(A_k) \geq \mathrm{pr}(\widehat{\vartheta}_k = \widehat{\vartheta}_k^o) \geq 1 - 2(p-q)\exp(-c_1 z_n^2)$. In the proof we show that $\mathrm{pr}\{B_k(s)\} \geq 1 - 2q\exp(-c_2 n\vartheta_s^2/q)$. Result (2.10) immediately follows from the fact that

$$\{\min[F\{\widehat{\mathcal{Q}}(\alpha)\}, G\{\widehat{\mathcal{Q}}(\alpha)\}] \geq s/q\} \quad \supset \quad \{|\widehat{\mathcal{Q}}(\alpha)| \leq q\} \cap \{|\widehat{\mathcal{Q}}(\alpha) \cap \mathcal{Q}| \geq s\}$$

$$\supset \quad \cap_{k=1}^{K^*}\{A_k \cap B_k(s)\}.$$

To ensure the probability lower bound in (2.10) approaches one, we need that $(p-q)\exp(-c_1 z_n^2) \to 0$ and $q\exp(-c_2 n\vartheta_s^2/q) \to 0$ as $n \to \infty$. The first requirement has been verified in the discussion after (2.9). The second one holds true whenever $|\vartheta_s| \gg (q\log q/n)^{1/2}$. This is a condition on the signal strength which is similar to the assumption on $\min_{j\in\mathcal{Q}}|\vartheta_j|$ in Theorem 1.

3. Implementation

This section details the implementation of our method. In the following numerical study of Section 4, we will use the check function (Koenker, 2005), given by

$$\ell_k(x) = x\{k/(K+1) - I(x<0)\} \quad (k=1,\ldots,K), \tag{3.1}$$

for calculating the estimator in (2.6). Other choices of $\ell_k(\cdot)$, such as the expectile loss function (Newey and Powell, 1987) that includes $\ell(x) = x^2/2$ as a special example, would work as well. The optimization problem (2.6) with loss function (3.1) can be efficiently solved by the `R` package `rqPen`, which implements the algorithm proposed by Yi and Huang (2017). It is straightforward to show that the conditions on the loss functions $\{\ell_1,\ldots,\ell_{K^*}\}$ that are required by Theorems 1 and 2 (see Assumptions 1 and 2 in the Supple-

mentary Material) are satisfied by the check function (3.1), as long as the distribution function of ε_1 is differentiable with a positive derivative value at its $k/(K+1)$ quantile ($k = 1, \ldots, K^*$).

A reasonable choice of the weights d_{kj} for the weighted L_1 penalty in (2.6) is the derivative of the SCAD function divided by the tuning parameter, i.e.

$$d_{kj} = I(|\widehat{\vartheta}_{kj}^{(0)}| \le \lambda_{n,k}) + I(|\widehat{\vartheta}_{kj}^{(0)}| > \lambda_{n,k})(a\lambda_{n,k} - |\widehat{\vartheta}_{kj}^{(0)}|)_+/\{(a-1)\lambda_{n,k}\}, \quad (3.2)$$

where a is a constant usually set to 3.7 (Fan and Li, 2001) and the initial estimator $\widehat{\vartheta}_k^{(0)} = \{\widehat{\vartheta}_{k0}^{(0)}, \widehat{\vartheta}_{k1}^{(0)}, \ldots, \widehat{\vartheta}_{kp}^{(0)}\}^{\mathrm{T}} = \arg\min_\theta\{\sum_{i=1}^n \ell_k(Y_i - X_i^{\mathrm{T}}\theta) + n\lambda_{n,k}\sum_{j=1}^p |\theta_j|\}$. The weighted L_1 penalty with such weights was also used, along with various loss functions, by Zou and Li (2008), Bradic et al. (2011) and Fan et al. (2014), among others. Bradic et al. (2011) showed that the data-driven weights d_{kj} given by (3.2) satisfy (under suitable conditions) the requirements in Assumption 7 of Section S1 in the Supplementary Material with probability tending to one. A detailed discussion on properties of this class of weighted L_1 penalty functions in the context of quantile regression can be found in Section 5 of Fan et al. (2014).

Regarding the tuning parameter $\lambda_{n,k}$ of the penalty in (2.6), we recom-

mend taking

$$\widehat{\lambda}_k = \arg\min_{\lambda\in\mathcal{L}} \big[\log\{\textstyle\sum_{i=1}^n \ell_k(Y_i - X_i^{\mathrm{T}}\widehat{\vartheta}_{k,\lambda})\} + C_n\,|\mathrm{supp}(\widehat{\vartheta}_{k,\lambda})|\big], \qquad (3.3)$$

where C_n is of order $\log p \times \log(\log n)/n$, $\mathcal{L}$ is a candidate grid and $\widehat{\vartheta}_{k,\lambda}$ represents the solution of (2.6) with $\lambda_{n,k} = \lambda$. This type of Bayesian information criteria is a fairly popular tool for selecting the tuning parameter in penalized regression (Chen and Chen, 2008; Wang et al., 2009; Kim et al., 2012; Wang et al., 2013; Lee et al., 2014; Peng and Wang, 2015; Sherwood and Wang, 2016; Dai et al., 2023). Alternatively one could use cross validation to determine $\lambda_{n,k}$. However, that approach would be considerably more time-consuming than employing the criterion (3.3), in particular for our case with multiple optimization tasks in high dimensional scenarios. To facilitate computation further we suggest solving the K minimization problems (2.6) *in parallel*, whenever possible. The processing time of our method on simulated data will be reported at the end of Section 4.2.

3.1 Post-selection estimation

After identifying the index set $\widehat{\mathcal{Q}}(\alpha)$ of the relevant predictors, we recommend estimating the slope vector $\vartheta_{-1} = (\vartheta_1, \ldots, \vartheta_p)^{\mathrm{T}} \in \mathbb{R}^p$ by composite quantile regression (Zou and Yuan, 2008; Gu and Zou, 2020): Our estimator

$\widehat{\theta}_{-1} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)^{\mathrm{T}} \in \mathbb{R}^p$ is obtained by solving

$$(\widehat{\theta}_{01}, \ldots, \widehat{\theta}_{0K}, \widehat{\theta}_1, \ldots, \widehat{\theta}_p) =$$

$$\arg\min_{(\theta_{01},\ldots,\theta_{0K})\in\mathbb{R}^K,\,(\theta_1,\ldots,\theta_p)\in\widehat{\Theta}} \sum_{k=1}^{K}\sum_{i=1}^{n} \ell_k(Y_i - \theta_{0k} - \sum_{j=1}^{p}\theta_j X_{ij}) \quad (3.4)$$

with $\widehat{\Theta} = \{(\theta_1, \ldots, \theta_p)^{\mathrm{T}} \in \mathbb{R}^p : \theta_j = 0 \text{ for } j \notin \widehat{\mathcal{Q}}(\alpha)\}$ and $\ell_k(\cdot)$ in (3.1).

Under the conditions in Theorem 1, our method satisfies $\mathrm{pr}\{\widehat{\mathcal{Q}}(\alpha) = \mathcal{Q}\} \to 1$, so that with probability tending to one, the number of predictors in the selected model is less than the sample size, i.e. $|\widehat{\mathcal{Q}}(\alpha)| = |\mathcal{Q}| < n$. Equation (3.4) therefore depicts a low-dimensional regression problem, which can be solved by the `R` package `cqrReg` without adopting a penalty term. Property $\mathrm{pr}\{\widehat{\mathcal{Q}}(\alpha) = \mathcal{Q}\} \to 1$ also implies $\widehat{\theta}_{-1}$ in (3.4) is asymptotically equivalent to the oracle estimator $\widehat{\theta}^o_{-1}$ which solves (3.4) with $\widehat{\Theta}$ replaced by $\{(\theta_1, \ldots, \theta_p)^{\mathrm{T}} \in \mathbb{R}^p : \theta_j = 0 \text{ for } j \notin \mathcal{Q}\}$. Zou and Yuan (2008) derive the limiting distribution of $\widehat{\theta}^o_{-1}$ and show that it is more efficient than other methods such as least squares regression. These properties hold under rather weak conditions on the error distribution, allowing for heavy tails and low density values at some points. In Section 4 we also implement penalized composite quantile regression, comparing its selection and estimation performance with that of our method.

4. Numerical studies

In this section we study the numerical performance of our method. The response vector is $Y = X\vartheta + \varepsilon$ throughout, with various choices of ϑ and ε. The first component of X_i is set to be constant one, while the last p components are drawn independently from a p-variate normal distribution with mean zero and a covariance matrix whose (i, j)th entry equals $0.5^{|i-j|}$. The sample size is always $n = 200$, whereas p will be set to different quantities. In the optimization problem (2.6) we set $K = 9$ and use the quantile loss function (3.1). The weighted L_1 penalty with weights (3.2) is applied to our method as well as to the other approaches. All the results are summarized over 200 iterations. Our method is also applied to analyze a data set of financial market indices. In the interest of space, we present the detailed descriptions and results of this real data example in Section S3.2 of the Supplementary Material.

4.1 Simple illustrations to demonstrate the key feature

We first illustrate the advantage of using multiple loss functions by means of two simple studies. This provides an explicit comparison between our method and a reference selection process, which is based on resampling and

the estimator

$$\widetilde{\vartheta}_k = \arg\min_\theta \{\sum_{i\in\mathcal{I}_k} \ell(Y_i - X_i^{\mathrm{T}}\theta) + n\lambda_{n,k}\sum_{j=1}^p d_{kj}|\theta_j|\} \tag{4.1}$$

for $k = 1, \ldots, K$. In contrast to our estimator $\widehat{\vartheta}_k$ in (2.6), the estimator $\widetilde{\vartheta}_k$ involves only data with indices in $\mathcal{I}_k$, which is a subset of size $\lfloor n/2 \rfloor$ randomly selected from $\{1, \ldots, n\}$. Also, all the K estimators in (4.1) are calculated using *just one* loss function $\ell(\cdot)$. This is an adaption of the "stability selection" approach developed by Meinshausen and Bühlmann (2010). That method fixes the tuning parameter $\lambda_{n,k} = \lambda_n$ for all the K estimators and yields a "stability path" along the possible values of λ_n, consisting of the relative frequencies for each predictor to be selected with a randomly chosen $\mathcal{I}_k$. Our approach is different: We allow the tuning parameters for the K estimators to vary, while aiming to obtain $\{\widetilde{\vartheta}_1, \ldots, \widetilde{\vartheta}_K\}$ for a voting procedure similar to ours, see equation (2.4) in Section 2.2, so that the two methods are comparable. The size $\lfloor n/2 \rfloor$ of $\mathcal{I}_k$ was recommended by Meinshausen and Bühlmann (2010) to resemble the bootstrap process (Freedman, 1977; Bühlmann and Yu, 2002) and facilitate computation.

We set $p = 6$ and $\vartheta = (0, 1, 1, 0, 0, 1, 0)^{\mathrm{T}} \in \mathbb{R}^{p+1}$ while letting $\ell(x) = x^2/2$ or $|x|/2$ in (4.1). The distribution of ε_1 in model (2.1) is either standard Cauchy or mixture normal 0.5N(-5, 1) + 0.5N(5, 1). Under these two non-standard error distributions (with an infinite first moment or extremely low

density value at the median), the results of the study, visualized by box plots, will demonstrate the superiority of our proposed method over the resampling-based approach, which relies on a single loss function.

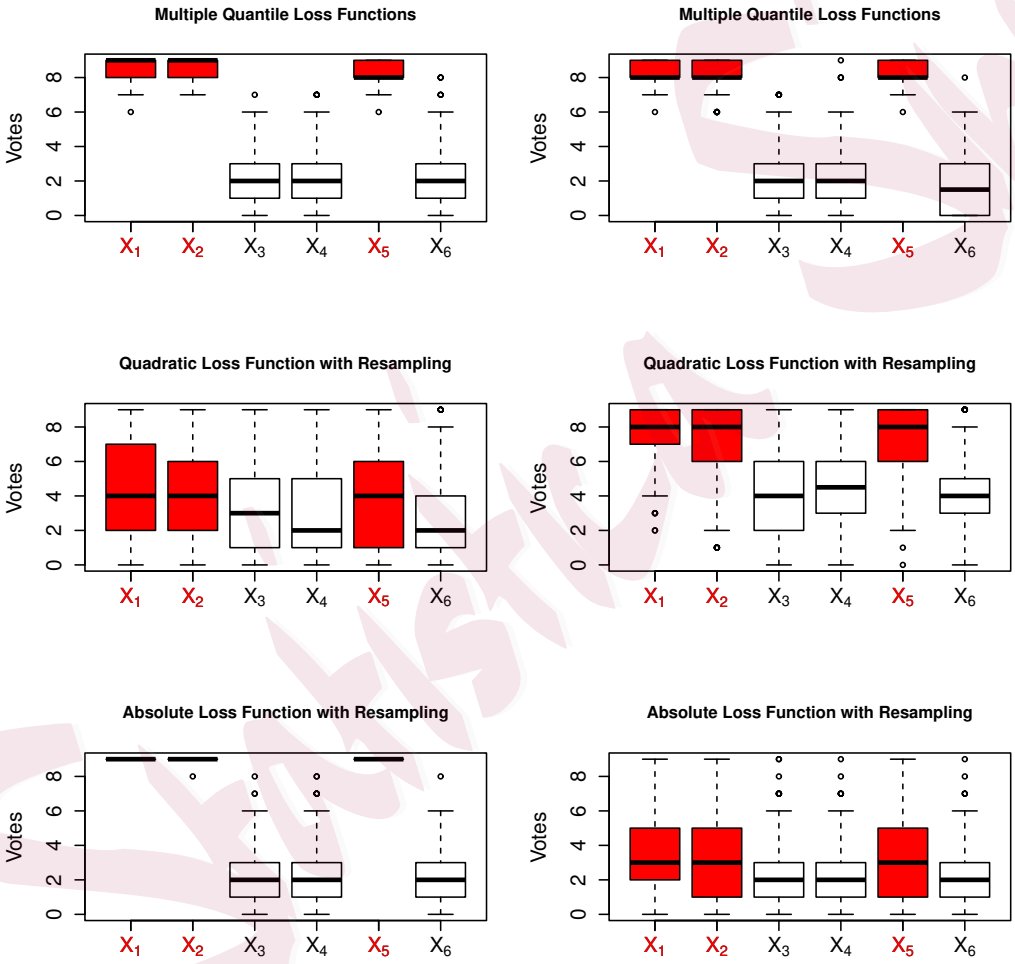
To reduce the impact of tuning parameter selection, we draw from the distribution of (X_1, Y_1) a validation set

$$\mathcal{V} = \{(X_i', Y_i') : i = 1, \ldots, 2000\} \tag{4.2}$$

independent of $\{(X_i, Y_i) : i = 1, \ldots, n\}$. The parameter $\lambda_{n,k}$ in (2.6) and in (4.1) is optimally determined by minimizing the prediction error (based on the corresponding loss function used for estimation) that is calculated from the data set $\mathcal{V}$, analogously to Wang et al. (2012), Mazumder et al. (2011), etc. Up to nine votes can be received by each predictor in the 200 iterations. They are presented in the six plots of Figure 1. When referring to "the (i, j)th plot" in the following, we mean the plot in the ith row and the jth column ($i \in \{1, 2, 3\}$; $j \in \{1, 2\}$).

Inspecting Figure 1, we observe in the $(1, 1)$th and $(1, 2)$th plots that our vote method with multiple quantile loss functions works very well for standard Cauchy errors (left panel) and for location mixture normal errors (right panel), clearly separating important (highlighted in red) and unimportant predictors. Although the check function $\ell_5(x) = |x|/2$ in (3.1) is not supposed to perform well when the error distribution is $0.5\mathrm{N}(-5, 1) + 0.5\mathrm{N}(5, 1)$,

Figure 1: Votes received by each predictor in 200 iterations. The error distribution in the panels on the left-hand side is standard Cauchy and on the right-hand side $0.5\mathrm{N}(\text{-}5, 1) + 0.5\mathrm{N}(5, 1)$. The first row displays the results from our vote method with multiple quantile loss functions, while the second and third rows present the results from the resampling-based vote method with the quadratic and absolute loss function. The important predictors are highlighted in red.



due to its extremely low density value at the median, our vote method still yields satisfactory selection results in this scenario. It indicates that the undesirable effect of $\widehat{\vartheta}_5$ calculated from $\ell_5(\cdot)$ has been reduced to an acceptable level, underlining the major advantage of our method. In contrast, the $(2,1)$th plot (Cauchy errors and quadratic loss) and the $(3,2)$th plot (location mixture normal errors and absolute loss) reveal the apparent failure of the resampling-based selection process when the random error follows the Cauchy or the location mixture normal distribution – all the predictors receive almost the same amount of votes and the important ones (highlighted in red) cannot be distinguished from the others. These results are in agreement with the statements in Sections 2.1 and 2.2: Variable selection based on the quadratic or the absolute loss function requires the error distribution to have a finite first moment or, respectively, the density value at the median to be positive. The quadratic loss function does not work well in the $(2,2)$th plot either. This may be attributed to the instability caused by the high variance of the mixture normal distribution $0.5\mathrm{N}(\text{-}5,1) + 0.5\mathrm{N}(5,1)$, which is 26. Under the Cauchy error distribution with a high density value at the median, the (3,1)th plot shows, as expected, that the absolute loss function gives results similar to those of our method. In general, the outcomes of these two simple experiments corroborate the

advantage of our vote method with multiple loss function, i.e. the ability to attain good selection results even if some of the loss functions fail.

4.2 Simulations

In the following we conduct simulations for high dimensional scenarios with $p = 500$ or 800. The non-zero components of ϑ are

$$(\vartheta_1, \vartheta_3, \vartheta_5, \vartheta_8, \vartheta_{10}, \vartheta_{13}, \vartheta_{16}) = (2.00, 1.50, 0.80, 1.00, 1.75, 0.75, 0.50). \quad (4.3)$$

Similar regression models were used by Fan et al. (2014) for simulations. We consider five different error distributions: (i) a normal distribution, $\mathrm{N}(0, 3)$; (ii) a t-distribution with two degrees of freedom, T_2; (iii) a double exponential distribution with mean 0 and variance 2; (iv) a location mixture normal distribution, $0.5\mathrm{N}(\text{-}3/2, 1) + 0.5\mathrm{N}(3/2, 1)$; and (v) a scale mixture normal distribution, $0.1\mathrm{N}(0, 25) + 0.9\mathrm{N}(0, 1)$. These choices cover a wide range of error distributions with various features.

As well as our vote method with multiple quantile loss functions we also implement a resampling-based version of our vote method with just one loss function $\ell(x) = x^2/2$ or $\ell(x) = |x|/2$, which recovers $\mathcal{Q}$ by a procedure similar to (2.4) but with $\widehat{\mathcal{Q}}_k$ replaced by $\mathrm{supp}(\widetilde{\vartheta}_k)$ from (4.1). The resampling-based approach serves as a reference to demonstrate the benefit of using multiple loss functions in our method. In addition, we consider three

standard approaches: least squares regression (LSR), least absolute deviations regression (LADR) and composite quantile regression (CQR) (Zou and Yuan, 2008; Gu and Zou, 2020) with quantile levels $\{0.1, 0.2, \ldots, 0.9\}$. The threshold for all voting procedures is $\alpha = 5$ or 6 (see Remark 1). For fair comparison, we apply the same weighted L_1 penalty with the weights given in (3.2) to our method and to all the competitors. The tuning parameters $\lambda_{n,k}$ are determined by minimizing prediction errors calculated over the validation data (4.2), but we also use $\widehat{\lambda}_k$ from (3.3) for the construction of our estimator. For each of these approaches, the means of false negatives (number of missed important predictors) and false positives (number of selected unimportant predictors), given by $|\widehat{\mathcal{Q}}^c \cap \mathcal{Q}|$ and $|\widehat{\mathcal{Q}} \cap \mathcal{Q}^c|$ for an estimator $\widehat{\mathcal{Q}}$ of $\mathcal{Q}$, are recorded in Table 1. The lower these two indices, the better a method performs. Based on the index set $\widehat{\mathcal{Q}}(\alpha)$ given by our selection procedure, we also construct the post-selection estimator $\widehat{\theta}_{-1}$ for the slope vector $\vartheta_{-1} = (\vartheta_1, \ldots, \vartheta_p)^{\mathrm{T}}$ as described in Section 3.1. We compare the L_2 error $E(\|\widehat{\theta}_{-1} - \vartheta_{-1}\|)$ of our estimator $\widehat{\theta}_{-1}$ with those of LSR, LADR and CQR which conduct selection and estimation simultaneously. Here $\|\cdot\|$ means the L_2 norm of a vector. The resampling-based selection approach (RVQ and RVA in the tables) does not yield an estimator for ϑ_{-1} so we do not consider its estimation error. The results of cases with $p = 800$ are

Table 1: Means of false negatives (FN), false positives (FP) and L_2 errors (Err) of our vote method using multiple quantile loss functions with tuning parameters selected by criterion (3.3) (VQ_C) or validation data (VQ_V), of the resampling-based vote method using the quadratic (RVQ) or the absolute loss function (RVA), of least squares regression (LSR), of least absolute deviations regression (LADR) and of composite quantile regression (CQR) for various error distributions. The lower FN, FP and Err, the better a method performs. The numbers (in parentheses) after the four vote methods refer to the values of the threshold α in the voting procedure (2.4). Here the sample size $n = 200$ and the full model size $p = 500$. The abbreviation DE denotes the double exponential distribution with mean 0 and variance 2; LMN a location mixture normal distribution, $0.5\text{N}(\text{-}3/2, 1) + 0.5\text{N}(3/2, 1)$; SMN a scale mixture normal distribution, $0.1\text{N}(0, 25) + 0.9\text{N}(0, 1)$.

	N(0, 3)			T_2			DE			LMN			SMN		
	FN	FP	Err	FN	FP	Err	FN	FP	Err	FN	FP	Err	FN	FP	Err
$\text{VQ}_\text{C}(5)$	0.59	0.41	0.55	0.36	0.15	0.40	0.10	0.13	0.28	0.69	0.92	0.57	0.12	0.12	0.28
$\text{VQ}_\text{V}(5)$	0.28	1.11	0.53	0.10	1.18	0.43	0.01	0.99	0.34	0.40	1.22	0.56	0.01	1.08	0.34
RVQ(5)	0.26	2.07	—	1.17	2.23	—	0.08	2.05	—	0.30	2.20	—	0.38	2.26	—
RVA(5)	0.36	6.88	—	0.15	6.33	—	0.04	6.73	—	0.53	7.03	—	0.03	6.92	—
$\text{VQ}_\text{C}(6)$	0.77	0.17	0.58	0.50	0.09	0.44	0.21	0.05	0.30	0.92	0.30	0.59	0.21	0.04	0.31
$\text{VQ}_\text{V}(6)$	0.42	0.42	0.58	0.17	0.44	0.39	0.06	0.33	0.29	0.56	0.48	0.61	0.02	0.45	0.29
RVQ(6)	0.39	0.77	—	1.43	0.85	—	0.14	0.80	—	0.49	0.87	—	0.55	0.89	—
RVA(6)	0.50	3.08	—	0.25	2.77	—	0.09	3.14	—	0.74	3.12	—	0.06	3.22	—
LSR	0.04	18.39	0.55	0.50	21.53	0.97	0.01	14.67	0.40	0.06	21.17	0.59	0.08	20.02	0.62
LADR	0.26	7.68	0.78	0.03	7.48	0.52	0.01	7.14	0.30	0.67	8.24	1.01	0.00	7.29	0.45
CQR	0.64	3.51	0.73	0.63	12.82	0.75	0.67	4.40	0.39	0.82	4.17	1.00	0.80	6.66	0.54

presented in Table S1 of the Supplementary Material.

In Tables 1 and S1 (of the Supplementary Material), our proposed method (VQ_C and VQ_V) produces in all the cases the lowest false positives, along with false negatives and L_2 errors that are fairly close to (if not lower than) those from the other approaches. These results indicate that our method achieves a good balance between the two aspects of variable

selection: it excludes unimportant predictors effectively, while the number of missed important predictors is kept at a reasonable level. Analogously to the results in Section 4.1, the approaches relying on the absolute loss function (RVA and LADR) and the quadratic loss function (RVQ and LSR) again break down when ε_1 follows the location mixture normal (with a low density value at the median) or the T_2 (with an infinite variance) distribution, giving high false positives or false negatives along with exceedingly large estimation errors. In contrast, our method recovers the underlying models in these two non-standard cases much more precisely. On the other hand, if the error distribution is normal or double exponential, i.e. of the more well-behaved type, our method still performs about as well as the resampling-based approach using the optimal loss function, i.e. the quadratic or absolute one. In contrast to our approach that uses the voting procedure, the approaches without voting, LSR, LADR and CQR, incorrectly identify considerably more unimportant predictors across all cases. This substantially impairs the *simplicity* and *interpretability* of the fitted models, which are the main goal of variable selection. In summary, the simulations confirm the usefulness of the voting strategy in reducing false positives. The numbers in Tables 1 and S1 demonstrate the advantage of our method over the other approaches, which is achieved by the combination of

multiple loss functions and the voting strategy. The simulation results also justify the recommended choices for the threshold α and tuning parameter $\lambda_{n,k}$ in equations (2.5) and (3.3). For α in the voting step (2.4), the numbers in Tables 1 and S1 suggest the choice $\alpha = 6$ tends to underfit models slightly (as do higher values of α), so we prefer to take $\alpha = \lceil K/2 \rceil = 5$ as recommended in (2.5). The Bayesian-type information criterion (3.3) for selecting $\lambda_{n,k}$ has a decent performance (see the results of $\mathrm{VQ_C}$ in Tables 1 and S1): It yields, in general, smaller models than the approach using the validation data (4.2) ($\mathrm{VQ_V}$). The tendency of the Bayesian information criteria to underfit models was also noticed by Lee et al. (2014) in their numerical study. Last but not least, criterion (3.3) can be recommended for practical applications because of its superior computational efficiency relative to cross validation.

We measured the computing speed on a 2.6 GHz processor when $p = 800$, $\varepsilon_1 \sim \mathrm{N}(0, 3)$ and $2,000$ validation data were used to select the tuning parameters from a candidate grid of 100 different quantities. In this setup the mean and the median processing times of least absolute deviations regression were 16 seconds in both cases, while those of our method were 63 and 62 seconds. The 9 minimization tasks in (2.6) were conducted in parallel using the `R` package `foreach`. Our method needed only 3 times more

time than a single penalized quantile regression process, although 9 such problems had to be solved. We expect that this should be almost the same if we determine the tuning parameters using criterion (3.3), because such a change does not increase the computational complexity. The use of parallel computing therefore keeps the computational burden of our method at a reasonable level.

5. Concluding remarks

In order to identify the relevant predictors in a high dimensional linear model, we have developed a selection process that combines the results of multiple sparse estimators. Our method was derived for linear regression, which, thanks to its simple structure, is the most fundamental regression model for the analysis of high dimensional data. Yet our method can be naturally extended to variable selection in more sophisticated regression models, e.g. nonparametric additive models (Huang et al., 2010) and semi-parametric partially linear models (Wang et al., 2011; Sherwood and Wang, 2016). By modifying the estimation procedure (2.3) according to the specific model, we expect that in these settings a selection process based on multiple sparse estimators and a voting procedure will be able to effectively identify the predictors that affect the response in a linear or nonlinear way.

As a concrete example, we elaborate the extension of our method to non-parametric additive models in Section S4 of the Supplementary Material.

Supplementary Materials

The assumptions and proofs of the theoretical results from Section 2, as well as additional numerical results and a discussion on the extension of our method, are provided in the online Supplementary Material. All the programs in Section 4 are available at `https://github.com/guorongdai/Variable-Selection-through-Vote`.

Acknowledgement

We thank the editors and reviewers for their constructive comments. Carroll's research was supported in part by the National Cancer Institute (Grant No. U01-CA057030). Dai's research was supported in part by the National Natural Science Foundation of China (Grant No. 72271060).

References

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

REFERENCES

Bühlmann, P. and B. Yu (2002). Analyzing bagging. *Annals of Statistics 30*(4), 927–961.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Dai, G., U. U. Müller, and R. J. Carroll (2023). Data integration in high dimension with multiple quantiles. *Statistica Sinica 33*, 169–191.

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. *Econometrics 6*(4), 45.

Fan, J., Y. Fan, and E. Barut (2014). Adaptive robust variable selection. *Annals of Statistics 42*(1), 324–351.

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

Freedman, D. (1977). A remark on the difference between sampling with and without replacement. *Journal of the American Statistical Association 72*(359), 681–681.

Gao, X. and R. J. Carroll (2017). Data integration with high dimensionality. *Biometrika 104*, 251–272.

Gu, Y. and H. Zou (2020). Sparse composite quantile regression in ultrahigh dimensions with tuning parameter calibration. *IEEE Transactions on Information Theory 66*, 7132–7154.

Huang, J., J. L. Horowitz, and F. Wei (2010). Variable selection in nonparametric additive models. *Annals of Statistics 38*(4), 2282.

REFERENCES

Kim, Y., S. Kwon, and H. Choi (2012). Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research 13*(1), 1037–1057.

Koenker, R. (2005). *Quantile Regression*. Cambridge, UK: Cambridge University Press.

Lee, E. R., H. Noh, and B. U. Park (2014). Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association 109*(505), 216–229.

Mazumder, R., J. H. Friedman, and T. Hastie (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association 106*(495), 1125–1138.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B 72*(4), 417–473.

Nan, Y. and Y. Yang (2014). Variable selection diagnostics measures for high-dimensional regression. *Journal of Computational and Graphical Statistics 23*(3), 636–656.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society 55*(4), 819–847.

Peng, B. and L. Wang (2015). An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics 24*(3), 676–694.

Sherwood, B. and L. Wang (2016). Partially linear additive quantile regression in ultra-high dimension. *Annals of Statistics 44*(1), 288–317.

REFERENCES

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

Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B 71*(3), 671–683.

Wang, L., Y. Kim, and R. Li (2013). Calibrating non-convex penalized regression in ultra-high dimension. *Annals of Statistics 41*(5), 2505.

Wang, L., X. Liu, H. Liang, and R. J. Carroll (2011). Estimation and variable selection for generalized additive partial linear models. *Annals of Statistics 39*(4), 1827.

Wang, L., Y. Wu, and R. Li (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association 107*(497), 214–222.

Wu, Y. and Y. Liu (2009). Variable selection in quantile regression. *Statistica Sinica 19*(2), 801–817.

Yi, C. and J. Huang (2017). Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics 26*(3), 547–557.

Yu, Y., Y. Yang, and Y. Yang (2022). Performance assessment of high-dimensional variable identification. *Statistica Sinica 32*, 695–718.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics 38*(2), 894–942.

REFERENCES

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

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics 36*(4), 1509–1533.

Zou, H. and M. Yuan (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics 36*(3), 1108–1126.

Guorong Dai (corresponding author), Department of Statistics and Data Science, School of Management, Fudan University, Shanghai 200433, China

E-mail: guorongdai@fudan.edu.cn

Ursula U. Müller, Department of Statistics, Texas A&M University, College Station, TX 77843, USA

E-mail: uschi@stat.tamu.edu

Raymond J. Carroll, Department of Statistics, Texas A&M University, College Station, TX 77843, USA

E-mail: carroll@stat.tamu.edu