PENALIZED REGRESSION WITH MULTIPLE LOSS FUNCTIONS AND VARIABLE SELECTION BY VOTING

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

[a] *Department of Statistics and Data Science, School of Management, Fudan University*
Shanghai 200433, China
[b] *Department of Statistics, Texas A&M University*
College Station, TX 77843, USA

Supplementary Material

We collect in the following all technical details that cannot be accommodated in the main article, including the regularity conditions and proofs of the theoretical results from Section 2. Also, we present additional numerical results and discuss a possible extension of our method.

S1 Assumptions

In this section we list seven regularity conditions that are needed in the proof of Theorems 1–2, followed by some explanations. We first introduce some notation. For a matrix M let $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ denote its maximum and minimum singular value. Further set $\|M\| = \lambda_{\max}(M)$ and $\|M\|_{2,\infty} = \sup_{\|v\|=1} \|Mv\|_\infty$, where the notation $\|M\|_\infty$ means the sup-norm of the matrix M, i.e. the maximum absolute value of its entries.

Assumption 1. Let $\psi_k(\cdot)$ be a subdifferential of $\ell_k(\cdot)$ and $\mathcal{N}_k$ be the set of not differentiable points of $\psi_k(\cdot)$. Then the equation $E\{\psi_k(\varepsilon_1 - x)\} = 0$ in terms of x has a unique solution $t_k \in \mathbb{R}$ and the distribution of ε_1 satisfies $\mathrm{pr}(\varepsilon_1 - t_k \in \mathcal{N}_k) = 0$ $(k = 1, \ldots, K^*)$.

Assumption 2. The function $\psi_k(\cdot)$ is such that $E\{\psi_k(\varepsilon_1 - t_k + x)\} = \eta_k x + o(|x|)$ as $|x| \to 0$ for some constant $\eta_k > 0$, and that $E\{|\psi_k(\varepsilon_1 - t_k)|^m\} \le c\, m! T^{m-2}$ for any $m \ge 2$ and some constant $T > 0$ $(k = 1, \ldots, K^*)$. For sufficiently small $|x|$, the expectation $E[\{\psi_k(\varepsilon_1 - t_k + x) - \psi_k(\varepsilon_1 - t_k)\}^2]$ exists and is continuous at $x = 0$ for $k = 1, \ldots, K^*$.

Assumption 3. For some constants $\kappa, \nu_0 \in (0, 1)$, the full model size $p = p_n$ and the number $q = q_n$ of non-zero parameters satisfy $\log p = O(n^{\kappa})$ and $q = O(n^{\nu_0})$.

Assumption 4. Let $X_{\mathcal{Q}} = (X_{1\mathcal{Q}}, \ldots, X_{n\mathcal{Q}})^{\mathrm{T}}$ and $X_{\mathcal{Q}^c} = (X_{1\mathcal{Q}^c}, \ldots, X_{n\mathcal{Q}^c})^{\mathrm{T}}$, where $X_{i\mathcal{Q}} = (1, X_{i1}, \ldots, X_{iq})^{\mathrm{T}}$ and $X_{i\mathcal{Q}^c} = (X_{i(q+1)}, \ldots, X_{ip})^{\mathrm{T}}$ for $i = 1, \ldots, n$. Then, for $k = 1, \ldots, K^*$,

$$\sup_{\theta \in \mathcal{B}_{n,k}} \|X_{\mathcal{Q}}^{\mathrm{T}} G_k(\theta) X_{\mathcal{Q}^c}\|_{2,\infty} = O(n^{1-\nu_1}) \text{ for some constant } \nu_1 \text{ and}$$

$$\inf_{\theta \in \mathcal{B}_{n,k}} \lambda_{\min}\{X_{\mathcal{Q}}^{\mathrm{T}} G_k(\theta) X_{\mathcal{Q}}\} \ge M_1 n \text{ for some positive constant } M_1,$$

where $\mathcal{B}_{n,k}$ is a $(q+1)$-dimensional ball centered at $\vartheta_{\mathcal{Q}}^{(k)} = (\vartheta_0 + t_k, \vartheta_1, \ldots, \vartheta_q)^{\mathrm{T}}$ with a radius ρ_n such that $\rho_n \gg n^{(\nu_0-1)/2}$, and $G_k(\theta)$ is a $n \times n$ diagonal matrix for any $\theta \in \mathbb{R}^{q+1}$, whose (i,i)th entry is $\partial E\{\psi_k(Y_i - X_{i\mathcal{Q}}^{\mathrm{T}}\theta + x)\}/\partial x|_{x=0}$.

Assumption 5. The design matrix X is such that

$$c_1\, n \le \lambda_{\min}(X_{\mathcal{Q}}^{\mathrm{T}} X_{\mathcal{Q}}) \le \lambda_{\max}(X_{\mathcal{Q}}^{\mathrm{T}} X_{\mathcal{Q}}) \le c_2\, n \text{ for some constants } 0 < c_1 \le c_2,$$

$$\|X\|_\infty = O(n^{1/2-(\nu_0-2\nu_1)_+/2-\nu_2}) \text{ and } \max_{1\le j\le p}\textstyle\sum_{i=1}^n X_{ij}^2 = O(n)$$

for some constant $\nu_2 \in [0, 1/2)$ that satisfies $\kappa < (\nu_0 - 2\nu_1)_+ + 2\nu_2 \le 1$, where $\{\kappa, \nu_0, \nu_1\}$ are the constants specified in Assumptions 3–4.

Assumption 6. The tuning parameter $\lambda_{n,k}$ satisfies

$$\lambda_{n,k} \;\gg\; n^{(\nu_0-2\nu_1)_+/2+\nu_2-1/2} \quad (k = 1, \ldots, K^*).$$

Assumption 7. The weight d_{kj} of the weighted L_1 penalty in (2.6) is such that

$$D_{n,k} = \max_{j\in\mathcal{Q}} d_{kj} = o(n^{\nu_1-\nu_0/2}),\ \lambda_{n,k} D_{n,k} = O(n^{-(1+\nu_0)/2}) \text{ and}$$

$$\liminf_{n\to\infty}(\min_{j\in\mathcal{Q}^c} d_{kj}) > 0 \text{ for } k = 1, \ldots, K^*.$$

Assumptions 1–2 regulate the error distribution and the loss functions $\{\ell_1(\cdot), \ldots, \ell_{K^*}(\cdot)\}$ used by the estimator in (2.6). The constant t_k introduced in Assumption 2 serves as an "offset" for the intercept term $(\vartheta_0 + t_k)$ of the kth regression model based on the loss function $\ell_k(\cdot)$. This ensures that $g(\theta) = \psi_k(Y_1 - X_1^{\mathrm{T}}\theta)$ is an *unbiased* estimating function for the parameter vector $(\vartheta_0 + t_k, \vartheta_1, \ldots, \vartheta_p)^{\mathrm{T}}$. It is easy to check that all the smoothness and moment conditions in Assumptions 1–2 are satisfied by the quantile loss functions used in the numerical study of Section 4 whenever for $k = 1, \ldots, K^*$, the distribution function of ε_1 is differentiable with a positive derivative value at point t_k, that is, at its $k/(K+1)$ quantile.

Assumption 3 is a mild condition on the growth rate of the model size in a linear model with a diverging number of parameters. The full model size $p = p_n$ is allowed to increase exponentially with n, while only the number $q = q_n$ of non-zero parameters is dominated by n. This is significantly weaker than its counterparts in many articles on high dimensional variable selection, e.g. $q = o(n^{1/2})$ in Wang et al. (2012) and Fan et al. (2014), and $q = o(n^{1/5})$ in Gao and Carroll (2017).

Assumptions 4–5 guarantee the good behavior of the design matrix. In particular, we allow the entries of the design matrix X to diverge, rather than requiring them to be bounded.

Assumptions 6–7 are imposed on the weighted L_1 penalty to ensure important predictors can be detected and irrelevant ones will be excluded. Some practical choices of d_{kj} and $\lambda_{n,k}$ have been provided and discussed in Section 3 for the implementation of our method.

All these assumptions are fairly mild and standard in the context of penalized regression for high dimensional linear models. Conditions similar to Assumptions 1-7 were required in Bradic et al. (2011) for penalized regression with a weighted linear combination of different loss functions. Let us point out again that we do not impose any conditions on the last $(K - K^*)$ of the K estimators $\{\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_K\}$ from (2.3): Only Assumptions 1, 2, 4, 6 and 7 involve the index k, but restrict k to the set $\{1, \ldots, K^*\}$, i.e. they concern only the first K^* estimators.

S2 Proofs

Proof of Proposition 1: Under the condition of Proposition 1, we have

$$\mathrm{pr}\{\cap_{j\in\mathcal{Q}}\{\textstyle\sum_{k=1}^K I(j \in \widehat{\mathcal{Q}}_k) \geq K^*\}\} \to 1 \text{ and}$$

$$\mathrm{pr}\{\cap_{j\in\mathcal{Q}^c}\{\textstyle\sum_{k=1}^K I(j \in \widehat{\mathcal{Q}}_k) \leq K - K^*\}\} \to 1.$$

This, combined with the fact that $K^* = \max\{\alpha, K - \alpha + 1\}$, implies

$$\mathrm{pr}\{\cap_{j\in\mathcal{Q}}\{\textstyle\sum_{k=1}^K I(j \in \widehat{\mathcal{Q}}_k) \geq \alpha\}\} \to 1 \text{ and}$$

$$\mathrm{pr}\{\cap_{j\in\mathcal{Q}^c}\{\textstyle\sum_{k=1}^K I(j \in \widehat{\mathcal{Q}}_k) < \alpha\}\} \to 1,$$

which give $\mathrm{pr}\{\widehat{\mathcal{Q}}(\alpha) = \mathcal{Q}\} \to 1$ according to the definition (2.4) of $\widehat{\mathcal{Q}}(\alpha)$.

Proof of Theorem 1: We will show the first conclusion of the theorem by proving the two conditions in Lemma 1 of Bradic et al. (2011) hold on an event with probability close to one.

Recall that $X_{i\mathcal{Q}} = (1, X_{i1}, \ldots, X_{iq})^{\mathrm{T}}$ for $i = 1, \ldots, n$ and $\vartheta_{\mathcal{Q}}^{(k)} = (t_k, \vartheta_1, \ldots, \vartheta_q)^{\mathrm{T}}$ for $k = 1, \ldots, K^*$, as defined in Assumption 4. Let

$$\begin{aligned}
&\Psi_k(\theta) = \{\psi_k(Y_1 - X_{1\mathcal{Q}}^{\mathrm{T}}\theta), \ldots, \psi_k(Y_n - X_{n\mathcal{Q}}^{\mathrm{T}}\theta)\}^{\mathrm{T}} \text{ for any } \theta \in \mathbb{R}^{q+1}, \\
&\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kp})^{\mathrm{T}} = X^{\mathrm{T}}\Psi_k(\vartheta_{\mathcal{Q}}^{(k)}) \text{ and} \\
&\Gamma_{n,k} = \{\max\nolimits_{j\in\mathcal{Q}^c}|\gamma_{kj}| \le n^{1/2}z_n\}. \qquad \text{(S2.1)}
\end{aligned}$$

Let $\widetilde{X}_j = (X_{ij}, \ldots, X_{nj})^{\mathrm{T}} \in \mathbb{R}^n$ for $j = 1, \ldots, p$. Then we have

$$\begin{aligned}
&\mathrm{pr}(|\gamma_{kj}| > n^{1/2}z_n) = \mathrm{pr}\{|\textstyle\sum_{i=1}^n X_{ij}\psi_k(\varepsilon_i - t_k)| > n^{1/2}z_n\} \\
&\le 2\exp\{-(2\|\widetilde{X}_j\|^2 + 2c\,n^{1/2}z_n\|\widetilde{X}_j\|_\infty)^{-1}nz_n^2\} \\
&= 2\exp[-\{2n^{-1}\|\widetilde{X}_j\|^2 + 2c\,n^{(\nu_0-2\nu_1)_+/2+\nu_2-1/2}\|\widetilde{X}_j\|_\infty\}^{-1}z_n^2] \\
&\le 2\exp(-c\,z_n^2) \quad (j \in \mathcal{Q}^c). \qquad \text{(S2.2)}
\end{aligned}$$

Here the second step uses Lemma 2.2.11 of van der Vaart and Wellner (1996) as well as the fact from Assumption 2 that $E\{\psi_k(\varepsilon_1 - t_k)\} = 0$ and $E\{|X_{ij}\psi_k(\varepsilon_i - t_k)|^m\} \le c\,m!X_{ij}^2(\|\widetilde{X}_j\|_\infty T)^{m-2}$ for any $m \ge 2$, while the last inequality uses Assumption 5. It follows that

$$\begin{aligned}
\mathrm{pr}(\Gamma_{n,k}) &\ge 1 - \textstyle\sum_{j\in\mathcal{Q}^c}\mathrm{pr}(|\gamma_{kj}| > n^{1/2}z_n) \\
&\ge 1 - 2(p-q)\exp(-c\,z_n^2). \qquad \text{(S2.3)}
\end{aligned}$$

Then, for $\widehat{\vartheta}_k^o$ defined in (2.7), with $\widehat{\vartheta}_{k\mathcal{Q}}^o = (\widehat{\vartheta}_{k0}^o, \widehat{\vartheta}_{k1}^o, \ldots, \widehat{\vartheta}_{kq}^o)^{\mathrm{T}}$ and $d_{k\mathcal{Q}} =$

$(0, d_{k1}, \ldots, d_{kq})^{\mathrm{T}}$, we have

$$\begin{aligned} \|\widehat{\vartheta}^o_{k\mathcal{Q}} - \vartheta^{(k)}_{\mathcal{Q}}\| &= O_p\{(q/n)^{1/2} + \lambda_{n,k}\|d_{k\mathcal{Q}}\|\} \\ &= O_p\{(q/n)^{1/2} + \lambda_{n,k} q^{1/2} D_{n,k}\} \\ &= O_p\{(q/n)^{1/2} + n^{-1/2}\} \\ &= O_p\{(q/n)^{1/2}\} = O_p(n^{(\nu_0-1)/2}), \qquad \text{(S2.4)} \end{aligned}$$

where the first step follows from Lemma 2 of Bradic et al. (2011), the third step uses Assumption 7 and the last step uses Assumption 3. The definition of $\widehat{\vartheta}^o_k$ in (2.7) implies that

$$X^{\mathrm{T}}_{\mathcal{Q}} \Psi_k(\widehat{\vartheta}^o_k) + n\lambda_{n,k} d_{k\mathcal{Q}} \circ \mathrm{Sign}(\widehat{\vartheta}^o_{k\mathcal{Q}}) = \mathbf{0}, \qquad \text{(S2.5)}$$

where symbol $\circ$ represents the Hadamard product, bold number $\mathbf{0}$ refers to the $(q+1)$-dimensional zero vector, and $\mathrm{Sign}(\cdot)$ is taken componentwise. Here $\mathrm{Sign}(x) = x/|x|$ for a scalar $x \neq 0$ and $\mathrm{Sign}(0) \in [-1, 1]$. With $\widetilde{d}_{k\mathcal{Q}^c} = (d^{-1}_{k(q+1)}, \ldots, d^{-1}_{kp})^{\mathrm{T}}$, we have that on the event $\Gamma_{n,k}$ defined in (S2.1),

$$\begin{aligned} &\|\widetilde{d}_{k\mathcal{Q}^c} \circ X^{\mathrm{T}}_{\mathcal{Q}^c} \Psi_k(\widehat{\vartheta}^o_{k\mathcal{Q}})\|_\infty \\ &\le \|\widetilde{d}_{k\mathcal{Q}^c} \circ X^{\mathrm{T}}_{\mathcal{Q}^c} \Psi_k(\vartheta^{(k)}_{\mathcal{Q}})\|_\infty + \|\widetilde{d}_{k\mathcal{Q}^c} \circ X^{\mathrm{T}}_{\mathcal{Q}^c} \{\Psi_k(\widehat{\vartheta}^o_{k\mathcal{Q}}) - \Psi_k(\vartheta^{(k)}_{\mathcal{Q}})\}\|_\infty \\ &\le c\,\{n^{1/2} z_n + \|X^{\mathrm{T}}_{\mathcal{Q}^c} G_k(\overline{\vartheta}_{k\mathcal{Q}}) X_{\mathcal{Q}} (\widehat{\vartheta}^o_{k\mathcal{Q}} - \vartheta^{(k)}_{\mathcal{Q}})\|_\infty\} \\ &\le c\,\{n^{1/2} z_n + \|X^{\mathrm{T}}_{\mathcal{Q}^c} G_k(\overline{\vartheta}_{k\mathcal{Q}}) X_{\mathcal{Q}}\|_{2,\infty} \|\widehat{\vartheta}^o_{k\mathcal{Q}} - \vartheta^{(k)}_{\mathcal{Q}}\|\} \\ &\le c\,(n^{1/2} z_n + n^{1-\nu_1} \|\widehat{\vartheta}^o_{k\mathcal{Q}} - \vartheta^{(k)}_{\mathcal{Q}}\|) \end{aligned}$$

$$= O(n^{(\nu_0-2\nu_1)_+/2+\nu_2+1/2}) + O_p(n^{1-\nu_1})O_p(n^{(\nu_0-1)/2})$$

$$= O(n^{(\nu_0-2\nu_1)_+/2+\nu_2+1/2}) + O_p(n^{\nu_0/2-\nu_1+1/2}) = o_p(n\lambda_{n,k}). \tag{S2.6}$$

In the above, the second inequality uses (S2.1), Assumption 7 and Taylor's expansion with $\overline{\vartheta}_{k\mathcal{Q}} = \vartheta^{(k)}_{\mathcal{Q}} + \mu(\widehat{\vartheta}^o_{k\mathcal{Q}} - \vartheta^{(k)}_{\mathcal{Q}})$ for some $\mu \in (0,1)$. The fourth step holds by Assumption 4 and the fact that $\|\overline{\vartheta}_{k\mathcal{Q}} - \vartheta^{(k)}_{\mathcal{Q}}\| \leq \|\widehat{\vartheta}^o_{k\mathcal{Q}} - \vartheta^{(k)}_{\mathcal{Q}}\| = O(n^{(\nu_0-1)/2})$ from (S2.4). The fifth step uses (S2.4) and the last step follows from Assumption 6.

Equations (S2.5) and (S2.6) guarantee that the conditions (27) and (28) of Lemma 1 in Bradic et al. (2011) are satisfied, which implies that $\widehat{\vartheta}^o_k$ is the unique global minimizer of the objective function in (2.6) on $\Gamma_{n,k}$. This combined with (S2.3), the definition of $\widehat{\vartheta}_k$ in (2.6) and Assumption 5 implies

$$\mathrm{pr}(\widehat{\vartheta}_k = \widehat{\vartheta}^o_k) \geq \mathrm{pr}(\Gamma_{n,k}) = 1 - 2(p-q)\exp(-c\, z_n^2) \to 1, \tag{S2.7}$$

which gives the first conclusion of the theorem. Equation (S2.7) and the definition of $\widehat{\vartheta}^o_k$ in (2.7) further yield

$$\mathrm{pr}(\cap_{j\in\mathcal{Q}^c}\{\widehat{\vartheta}_{kj} = \vartheta_j = 0\}) \to 1. \tag{S2.8}$$

Moreover, for $\widehat{\vartheta}_{k\mathcal{Q}} = (\widehat{\vartheta}_{k0}, \widehat{\vartheta}_{k1}, \ldots, \widehat{\vartheta}_{kq})^{\mathrm{T}}$ we know that, with probability tending to one,

$$\|\widehat{\vartheta}_{k\mathcal{Q}} - \vartheta^{(k)}_{\mathcal{Q}}\| = \|\widehat{\vartheta}^o_{k\mathcal{Q}} - \vartheta^{(k)}_{\mathcal{Q}}\| = O_p\{(q/n)^{1/2}\}. \tag{S2.9}$$

In the above, the first step uses (S2.7) while the last step follows from (S2.4). Then we have

$$\begin{aligned} \mathrm{pr}(\cap_{j\in\mathcal{Q}}\{|\widehat{\vartheta}_{kj}|>0\}) &\geq \mathrm{pr}(\cap_{j\in\mathcal{Q}}\{|\widehat{\vartheta}_{kj}|>|\vartheta_j|-\min_{j\in\mathcal{Q}}|\vartheta_j|\}) \\ &\geq \mathrm{pr}(\cap_{j\in\mathcal{Q}}\{|\widehat{\vartheta}_{kj}-\vartheta_j|<\min_{j\in\mathcal{Q}}|\vartheta_j|\}) \\ &\geq \mathrm{pr}(\|\widehat{\vartheta}_{k\mathcal{Q}}-\vartheta_{\mathcal{Q}}^{(k)}\|<\min_{j\in\mathcal{Q}}|\vartheta_j|)\to 1, \text{(S2.10)} \end{aligned}$$

where the convergence follows from (S2.9) and the condition that $\min_{j\in\mathcal{Q}}|\vartheta_j| \gg (q/n)^{1/2}$.

Combining (S2.8) and (S2.10) yields $\mathrm{pr}(\widehat{\mathcal{Q}}_k=\mathcal{Q})\to 1$ $(k=1,\ldots,K^*)$. It follows that $\mathrm{pr}\{\widehat{\mathcal{Q}}(\alpha)=\mathcal{Q}\}\to 1$ according to Proposition 1.

Proof of Theorem 2: Similar to (S2.2), we have

$$\begin{aligned} &\mathrm{pr}\{|\textstyle\sum_{i=1}^n X_{ij}\psi_k(\varepsilon_i-t_k)|>nM_1(1-\xi)|\vartheta_s|/q^{1/2}\} \\ &\leq 2\exp[-\{2\|\widetilde{X}_j\|^2/n+2c\,M_1(1-\xi)|\vartheta_s|\|\widetilde{X}_j\|_\infty/q^{1/2}\}^{-1}nM_1^2(1-\xi)^2\vartheta_s^2/q] \\ &\leq 2\exp(-c\,n\vartheta_s^2/q) \quad (j\in\mathcal{Q}), \end{aligned}$$

which implies

$$\begin{aligned} &\mathrm{pr}\{\|X_{\mathcal{Q}}^{\mathrm{T}}\Psi_k(\vartheta_{\mathcal{Q}}^{(k)})\|/n\leq M_1(1-\xi)|\vartheta_s|\} \\ &\geq \mathrm{pr}\{\|X_{\mathcal{Q}}^{\mathrm{T}}\Psi_k(\vartheta_{\mathcal{Q}}^{(k)})\|_\infty\leq nM_1(1-\xi)|\vartheta_s|/q^{1/2}\} \\ &= \mathrm{pr}[\cap_{j\in\mathcal{Q}}\{|\textstyle\sum_{i=1}^n X_{ij}\psi_k(\varepsilon_i-t_k)|\leq nM_1(1-\xi)|\vartheta_s|/q^{1/2}\}] \\ &\geq 1-\textstyle\sum_{j\in\mathcal{Q}}\mathrm{pr}\{|\sum_{i=1}^n X_{ij}\psi_k(\varepsilon_i-t_k)|>nM_1(1-\xi)|\vartheta_s|/q^{1/2}\} \end{aligned}$$

$$\geq 1 - 2q\exp(-c\, n\vartheta_s^2/q). \tag{S2.11}$$

For any $\theta = (\theta_0, \ldots, \theta_q)^{\mathrm{T}} \in \mathbb{R}^{q+1}$, let

$$L_k(\theta) = \textstyle\sum_{i=1}^n \ell_k(Y_i - X_{i\mathcal{Q}}^{\mathrm{T}}\theta) + n\lambda_{n,k} \textstyle\sum_{j=1}^q d_{kj}|\theta_j|.$$

Denote $\mathcal{U} = \{u \in \mathbb{R}^{q+1} : \|u\| = 1\}$. According to the proof of Lemma 2 in Bradic et al. (2011), we have

$$\begin{aligned} &\inf_{u\in\mathcal{U}} L_k(\vartheta_{\mathcal{Q}}^{(k)} + |\vartheta_s|u) - L_k(\vartheta_{\mathcal{Q}}^{(k)}) \\ &\geq n|\vartheta_s|\{M_1|\vartheta_s| - \|X_{\mathcal{Q}}^{\mathrm{T}}\Psi_k(\vartheta_{\mathcal{Q}}^{(k)})\|/n - \lambda_{n,k}\|d_{k\mathcal{Q}}\|\} \\ &> n|\vartheta_s|\{M_1(1-\xi)|\vartheta_s| - \|X_{\mathcal{Q}}^{\mathrm{T}}\Psi_k(\vartheta_{\mathcal{Q}}^{(k)})\|/n\}, \end{aligned} \tag{S2.12}$$

where the last step uses the assumption that $\lambda_{n,k}\|d_{k\mathcal{Q}}\| < M_1\,\xi\,|\vartheta_s|$. Since $\widehat{\vartheta}_{k\mathcal{Q}}^o$ is the unique minimizer of $L_k(\theta)$, we know

$$\{\inf_{u\in\mathcal{U}} L_k(\vartheta_{\mathcal{Q}}^{(k)} + |\vartheta_s|u) - L_k(\vartheta_{\mathcal{Q}}^{(k)}) > 0\} \subset \{\|\widehat{\vartheta}_{k\mathcal{Q}}^o - \vartheta_{\mathcal{Q}}^{(k)}\| < |\vartheta_s|\},$$

which indicates

$$\begin{aligned} \mathrm{pr}\{\|\widehat{\vartheta}_{k\mathcal{Q}}^o - \vartheta_{\mathcal{Q}}^{(k)}\| < |\vartheta_s|\} &\geq \mathrm{pr}\{\inf_{u\in\mathcal{U}} L_k(\vartheta_{\mathcal{Q}}^{(k)} + |\vartheta_s|u) - L_k(\vartheta_{\mathcal{Q}}^{(k)}) > 0\} \\ &\geq \mathrm{pr}[n|\vartheta_s|\{M_1(1-\xi)|\vartheta_s| - \|X_{\mathcal{Q}}^{\mathrm{T}}\Psi_k(\vartheta_{\mathcal{Q}}^{(k)})\|/n\} \geq 0] \\ &= \mathrm{pr}\{\|X_{\mathcal{Q}}^{\mathrm{T}}\Psi_k(\vartheta_{\mathcal{Q}}^{(k)})\|/n \leq M_1(1-\xi)|\vartheta_s|\} \\ &\geq 1 - 2q\exp(-c\, n\vartheta_s^2/q). \end{aligned} \tag{S2.13}$$

In the above the second step uses (S2.12) and the last step is due to (S2.11).

Therefore, similar to (S2.10), we know that

$$\begin{aligned}\mathrm{pr}\{\{\widehat{\vartheta}_k = \widehat{\vartheta}_k^o\} \cap (\cap_{j=1}^s \{|\widehat{\vartheta}_{kj}| > 0\})\} &\geq \mathrm{pr}(\{\widehat{\vartheta}_k = \widehat{\vartheta}_k^o\} \cap \{\|\widehat{\vartheta}_{k\mathcal{Q}} - \vartheta_{\mathcal{Q}}^{(k)}\| < |\vartheta_s|\}) \\ &= \mathrm{pr}(\{\widehat{\vartheta}_k = \widehat{\vartheta}_k^o\} \cap \{\|\widehat{\vartheta}_{k\mathcal{Q}}^o - \vartheta_{\mathcal{Q}}^{(k)}\| < |\vartheta_s|\}) \\ &\geq 1 - 2\{(p-q)\exp(-c_1 z_n^2) + q\exp(-c_2 n\vartheta_s^2/q)\},\end{aligned}$$

where the last step uses (S2.7) and (S2.13). Then we have

$$\begin{aligned}&\mathrm{pr}[\{|\widehat{\mathcal{Q}}(\alpha)| \leq q\} \cap \{|\widehat{\mathcal{Q}}(\alpha) \cap \mathcal{Q}| \geq s\}] \\ \geq\ &\mathrm{pr}[\cap_{k=1}^{K^*}\{\{\widehat{\vartheta}_k = \widehat{\vartheta}_k^o\} \cap (\cap_{j=1}^s\{|\widehat{\vartheta}_{kj}| > 0\})\}] \\ \geq\ &1 - 2K^*\{(p-q)\exp(-c_1 z_n^2) + q\exp(-c_2 n\vartheta_s^2/q)\}.\end{aligned}$$

It follows that

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

S3 Additional numerical results

S3.1 Simulations

In Table S1 we display the numerical results in the simulation settings described in Section 4.2 when the full model size $p = 800$.

Table S1: We consider the same scenario as Table 1 but now the full model size $p = 800$.

	$\mathrm{N}(0,3)$			T_2			DE			LMN			SMN		
	FN	FP	MSE	FN	FP	MSE	FN	FP	MSE	FN	FP	MSE	FN	FP	MSE
$\mathrm{VQ_C}(5)$	0.70	0.32	0.56	0.38	0.11	0.41	0.19	0.10	0.30	0.74	0.57	0.56	0.20	0.14	0.32
$\mathrm{VQ_V}(5)$	0.35	1.10	0.53	0.12	1.14	0.45	0.04	1.07	0.36	0.44	1.31	0.58	0.03	1.34	0.38
RVQ(5)	0.32	1.83	—	1.30	1.91	—	0.10	1.94	—	0.35	1.89	—	0.51	2.03	—
RVA(5)	0.43	6.21	—	0.23	5.90	—	0.09	6.37	—	0.46	7.86	—	0.08	6.24	—
$\mathrm{VQ_C}(6)$	0.93	0.11	0.59	0.70	0.05	0.51	0.27	0.03	0.32	1.01	0.19	0.60	0.31	0.06	0.34
$\mathrm{VQ_V}(6)$	0.46	0.37	0.59	0.22	0.38	0.40	0.08	0.35	0.30	0.59	0.52	0.62	0.06	0.55	0.32
RVQ(6)	0.42	0.65	—	1.67	0.63	—	0.16	0.74	—	0.53	0.76	—	0.66	0.69	—
RVA(6)	0.56	2.68	—	0.34	2.49	—	0.14	2.74	—	0.66	3.48	—	0.13	2.66	—
LSR	0.08	22.80	0.57	0.60	25.70	1.06	0.01	19.54	0.42	0.08	23.42	0.61	0.13	23.53	0.66
LADR	0.29	8.91	0.76	0.03	10.23	0.54	0.01	8.23	0.32	0.69	8.35	1.02	0.02	9.40	0.45
CQR	0.38	7.27	0.78	0.94	13.49	0.74	0.30	9.18	0.40	0.32	9.20	0.83	0.46	13.84	0.57

S3.2 A real data example

We now perform the same analysis as in Section 4.2, but instead of simulated predictors we now use real data, namely $p = 46$ indices of major international equities, North American bonds and major commodities. The transformation $\log(V_t/V_y) \times 100$ is applied to each index, where V_t and V_y denote today's and yesterday's closing values. The `R` package `FusionLearning` provides $n = 232$ records of three years' market performances of these indices with three-day spacing between the values. As shown in Gao and Carroll (2017), the values are not autocorrelated at a 5% significance level.

We generate the response vector Y through the model $Y = X\vartheta + \varepsilon$, where ϑ is the parameter vector with seven non-zero components given in equation (4.3) at the beginning of Section 4.2. We consider the same five error distributions listed after (4.3). Here we generate Y via the simulated model instead of using the original responses, because we need to know the true index set $\mathcal{Q}$ of important predictors. We also want to evaluate the performance of various methods under different error distributions, so that the analysis is more illustrative and informative. The strategy using real predictors and simulated responses was also used by Meinshausen and Bühlmann (2010) for their numerical study of variable selection methods. Since the observations are real data, the tuning parameters in all methods are chosen by criterion (3.3). The loss function $\ell_k(\cdot)$ in (3.3) is the same as the one used for estimation.

The results displayed in Table S2 present a similar picture to those in Tables 1 and S1. Our method still gives the lowest number of false positives in most cases, while the false negatives and L_2 errors are close to or better than the other approaches. We again notice that the performance is rather poor when the absolute or the quadratic loss functions are used and the errors have the location mixture normal or the T_2 distribution. Generally speaking, our method outperforms the other approaches and recovers the

Table S2: We consider the same scenario as Table 1, but now the design matrix is from a real data set of financial indices. The tuning parameters of all methods are determined by criterion (3.3).

	$N(0,3)$			T_2			DE			LMN			SMN		
	FN	FP	Err	FN	FP	Err	FN	FP	Err	FN	FP	Err	FN	FP	Err
$\mathrm{VQ_C}(5)$	0.28	0.72	0.61	0.18	0.32	0.47	0.03	0.37	0.35	0.45	0.76	0.72	0.07	0.27	0.37
RVQ(5)	0.15	11.53	—	0.55	10.12	—	0.06	11.01	—	0.18	11.34	—	0.18	10.86	—
RVA(5)	0.23	2.90	—	0.05	2.53	—	0.01	2.72	—	0.53	2.78	—	0.02	2.70	—
$\mathrm{VQ_C}(6)$	0.40	0.38	0.63	0.30	0.20	0.50	0.06	0.24	0.34	0.70	0.37	0.77	0.14	0.15	0.38
RVQ(6)	0.29	6.77	—	0.86	5.90	—	0.11	6.44	—	0.34	6.58	—	0.32	6.23	—
RVA(6)	0.36	1.53	—	0.11	1.46	—	0.02	1.55	—	0.81	1.56	—	0.04	1.55	—
LSR	0.09	9.66	0.62	0.48	9.74	1.87	0.05	9.81	0.98	0.15	9.88	1.24	0.17	9.32	1.40
LADR	0.09	5.65	0.83	0.02	4.81	0.59	0.00	5.43	0.38	0.32	6.14	1.15	0.01	5.25	0.55
CQR	0.17	7.79	0.74	0.07	6.16	0.59	0.03	4.89	0.43	0.17	9.07	0.99	0.04	4.32	0.49

true models precisely across all settings. This real data example again confirms the advantage of aggregating selection results from multiple loss functions via the voting procedure.

S4 Extension to nonparametric additive models

In this section we consider the extension of our selection method to the following nonparametric additive model:

$$Y_i \;=\; g(X_i) + \varepsilon_i \;=\; \textstyle\sum_{j=1}^{p} g_j(X_{ij}) + \varepsilon_i \quad (i = 1, \ldots, n) \tag{S4.1}$$

with unknown smooth functions $\{g_1(\cdot), \ldots, g_p(\cdot)\}$. This model is useful for analyzing data with many covariates when the parametric regression model is too restrictive. In particular, because of its additive structure, it does not suffer from the curse of dimensionality. The nonparametric functions $\{g_1, g_2, \ldots, g_p\}$ can be estimated by orthogonal series estimators as follows: Let $\xi = (\xi_1, \ldots, \xi_M)^{\mathrm{T}}$ denote the first M elements of an orthonormal basis for approximating g_j, e.g. a cosine basis or a B-spline basis; see, for example, Section 2 in Huang et al. (2010) and Section 4 in Müller et al. (2012). Then $g_j(X_{ij})$ can be well approximated by $\xi(X_{ij})^{\mathrm{T}}\beta_j$ ($i = 1, \ldots, n$; $j = 1, \ldots, p$) with parameter vector $\beta_j \in \mathbb{R}^M$ and M sufficiently large, so that model

(S4.1) has the following approximate representation:

$$Y_i \;=\; \textstyle\sum_{j=1}^{p} g_j(X_{ij}) + \varepsilon_i \;\approx\; \textstyle\sum_{j=1}^{p} \xi(X_{ij})^{\mathrm{T}}\beta_j + \varepsilon_i \quad (i = 1, \ldots, n). \tag{S4.2}$$

We now write Z_i for the $(p \times M + 1)$-dimensional vector that consists of all basis functions, i.e. for $i = 1, ..., n$, let $Z_i = \{1, \xi(X_{i1})^{\mathrm{T}}, \ldots, \xi(X_{ip})^{\mathrm{T}}\}^{\mathrm{T}}$ with the constant one included to capture an intercept term such as the mean or median of ε_1 when regressing $\{Y_1, \ldots, Y_n\}$ on $\{Z_1, \ldots, Z_n\}$. Adapting the K estimators in (2.3) to the additive model yields the following sparse estimators $(\widehat{\beta}_{k1}^{\mathrm{T}}, \ldots, \widehat{\beta}_{kp}^{\mathrm{T}})^{\mathrm{T}}$ for $(\beta_1^{\mathrm{T}}, \ldots, \beta_p^{\mathrm{T}})^{\mathrm{T}}$:

$$(\widehat{\beta}_{k0}, \widehat{\beta}_{k1}^{\mathrm{T}}, \ldots, \widehat{\beta}_{kp}^{\mathrm{T}})^{\mathrm{T}} = \arg\min{}_b \{\textstyle\sum_{i=1}^{n} \ell_k(Y_i - Z_i^{\mathrm{T}} b) + \textstyle\sum_{j=1}^{p} \phi_k(\|b_j\|)\}$$

for $k = 1, \ldots, K$, where $b = (b_0, b_1^{\mathrm{T}}, \ldots, b_p^{\mathrm{T}})^{\mathrm{T}}$ with $b_0 \in \mathbb{R}$ and $b_j \in \mathbb{R}^M$ for $j = 1, \ldots, p$. As before, $\phi_k(\cdot)$ denotes a penalty function and $\|\cdot\|$ represents the L_2 norm of a vector. Then, analogously to (2.4), a voting procedure with a threshold α is given by $\widehat{\mathcal{T}}(\alpha) = \{j \in \{1, \ldots, p\} : \sum_{k=1}^{K} I\{\|\widehat{\beta}_{kj}\| > 0\} \geq \alpha\}$. This yields our estimator $\widehat{\mathcal{T}}(\alpha)$ for the index set $\{j \in \{1, \ldots, p\} : \|\beta_j\| > 0\}$ of the influential components among $\{g_1, \ldots, g_p\}$ in model (S4.2). We leave a rigorous analysis of the method sketched above and the treatment of more general semiparametric regression models for future work.

Bibliography

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.

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

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

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

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

Müller, U. U., A. Schick, and W. Wefelmeyer (2012). Estimating the error distribution function in semiparametric additive regression models. *Journal of Statistical Planning and Inference 142*(2), 552–566.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Series in Statistics. Springer.

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.