Supplement for "Fast Cost-Constrained Regression"

Hyeong Jin Hyun and Xiao Wang

The proofs of Theorems 1, 2 and 3 are provided in Section S1. The sufficient conditions for QM* and several loss functions that satisfy QM* are presented in Section S2. The stopping criteria for the FCR and GFCR algorithms are discussed in S3. Additional details and results of numerical experiments and NHANES data analysis are given in Sections S5, and S6.

S1 Proofs of Theorems

For the references from the main article, we add the character 'm' before the number such as (1) for Equation (1) in the main article.

S1.1 Proof of Theorem 1

Proof. Define the loss function as

$$L(\beta_1, \ldots, \beta_p) := \frac{1}{2n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

Observe that $\sum_{k=1}^{p} c_k I(\beta_k^\dagger) \leq C$ is true by the definition of $\boldsymbol{\beta}^\dagger$.

Let $\beta_j^\dagger \in \mathcal{I}^\dagger$. Since $(\beta_1^\dagger, \cdots, \beta_{j-1}^\dagger, t, \beta_{j+1}^\dagger, \beta_p^\dagger) \in \mathcal{C}$ for all $t \in \mathbb{R}$ we have

$$\beta_j^\dagger = \arg\min_{t \in \mathbb{R}} L(\beta_1^\dagger, \ldots, \beta_{j-1}^\dagger, t, \beta_{j+1}^\dagger, \ldots, \beta_p^\dagger).$$

By simple algebra, we have

$$\begin{aligned}\beta_j^\dagger &= \arg\min_{t\in\mathbb{R}} \frac{1}{2n}\|\boldsymbol{y} - \mathbf{X}\beta^\dagger - (t-\beta_j^\dagger)X_j\|_2^2 \\ &= \arg\min_{t\in\mathbb{R}} \left[\frac{\|X_j\|_2^2}{2n}(t-\beta_j^\dagger)^2 - \frac{1}{n}(t-\beta_j^\dagger)X_j^\top(\boldsymbol{y}-\mathbf{X}\beta^\dagger)\right] \\ &= \arg\min_{t\in\mathbb{R}} \left[t - \left(\beta_j^\dagger + \frac{1}{\|X_j\|_2^2}X_j^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}^\dagger)\right)\right]^2. \end{aligned} \tag{S1.1}$$

Hence, we have $\beta_j^\dagger = \beta_j^\dagger + \frac{1}{\|X_j\|_2^2}X_j^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}^\dagger)$.

Now, let $\beta_j^\dagger \notin \mathcal{I}^\dagger$. If the perturbation t of $\beta_j^\dagger$ violates the constraints, i.e. $\sum_{k\neq j} c_k I(\beta_k^\dagger) \leq C$, but $\sum_{k\neq j} c_k I(\beta_k^\dagger) + c_j > C$, then the minimizer is $t = 0$. Otherwise, $(\beta_1^\dagger, \cdots, \beta_{j-1}^\dagger, t, \beta_{j+1}^\dagger, \beta_p^\dagger) \in \mathcal{C}$ for all $t \in \mathbb{R}$, and hence the following holds, by the definition of $\beta^\dagger$,

$$0 = \arg\min_{t\in\mathbb{R}} L(\beta_1^\dagger, \ldots, \beta_{j-1}^\dagger, t, \beta_{j+1}^\dagger, \ldots, \beta_p^\dagger).$$

This completes the first argument. For the second argument, we notice

$$\boldsymbol{\beta}^\dagger = \arg\min_{(t_1,\cdots,t_p)\in\mathcal{C}} \left[L(t_1, \beta_2^\dagger, \ldots, \beta_p^\dagger) + L(\beta_1^\dagger, t_2, \ldots, \beta_p^\dagger) + \cdots + L(\beta_1^\dagger, \beta_2^\dagger, \ldots, t_p)\right] \tag{S1.2}$$

$$= \arg\min_{(t_1,\cdots,t_p)\in\mathcal{C}} \sum_{j=1}^p \left[t_j - \left(\beta_j^\dagger + \frac{1}{\|X_j\|_2^2}X_j^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}^\dagger)\right)\right]^2,$$

which ends the proof. The first equality (S1.2) holds from proof of contradiction. Let $t^\dagger \neq \boldsymbol{\beta}^\dagger$ be solution that satisfies (S1.2). Then,

$$\sum_{j=1}^p L(\beta_1^\dagger, \ldots, t_k^\dagger, \ldots, \beta_p^\dagger) \leq pL(\beta_1^\dagger, \ldots, \beta_p^\dagger). \tag{S1.3}$$

However, for any $t_k^\dagger$ such that $L(\beta_1^\dagger, \ldots, t_k^\dagger, \ldots, \beta_p^\dagger) \leq L(\beta_1^\dagger, \ldots, \beta_k^\dagger, \ldots, \beta_p^\dagger)$, $t_k^\dagger = \beta_k^\dagger$ due to the definition of $\beta_k^\dagger$. Contrapositive of the statement implies that

$$\text{if } t_k^\dagger \neq \beta_k^\dagger, \text{ then } L(\beta_1^\dagger, \ldots, t_k^\dagger, \ldots, \beta_p^\dagger) > L(\beta_1^\dagger, \ldots, \beta_k^\dagger, \ldots, \beta_p^\dagger).$$

Since there exists at least one $k \in \{1, 2, \ldots, p\}$ such that $t_k^\dagger \neq \beta_k^\dagger$, it contradicts (S1.3). □

S1.2 Proof of Theorem 2

Proof. Define $\boldsymbol{\beta}_{-j}^\dagger(t) = (\beta_1^\dagger, \ldots, \beta_{j-1}^\dagger, t, \beta_{j+1}^\dagger, \ldots, \beta_p^\dagger)$ and observe that $\sum_{k=1}^p c_k I(\beta_k^\dagger) \leq C$ holds by the definition of $\boldsymbol{\beta}^\dagger$.

Let $\beta_j^\dagger \in \mathcal{I}^\dagger$. Since $(\beta_1^\dagger, \cdots, \beta_{j-1}^\dagger, t, \beta_{j+1}^\dagger, \beta_p^\dagger) \in \mathcal{C}$ for all $t \in \mathbb{R}$ we have

$$\begin{aligned}
\beta_j^\dagger &= \arg\min_{t \in \mathbb{R}} L(\beta_1^\dagger, \ldots, \beta_{j-1}^\dagger, t, \beta_{j+1}^\dagger, \ldots, \beta_p^\dagger) \\
&= \arg\min_{t \in \mathbb{R}} Q(\boldsymbol{\beta}_{-j}^\dagger(t), \boldsymbol{\beta}^\dagger).
\end{aligned}$$

The second equality comes from (2.12). It leads us to get

$$\begin{aligned}
\beta_j^\dagger &= \arg\min_{t \in \mathbb{R}} Q(\boldsymbol{\beta}_{-j}^\dagger(t), \boldsymbol{\beta}^\dagger) \\
&= \arg\min_{t \in \mathbb{R}} \|\tilde{X}\boldsymbol{\beta}_{-j}^\dagger(t) - \tilde{X}\boldsymbol{\beta}^\dagger + \boldsymbol{g}\|_2^2 \\
&= \arg\min_{t \in \mathbb{R}} \|\tilde{X}_j^\top (t - \beta_j^\dagger) + \boldsymbol{g}\|_2^2 \\
&= \arg\min_{t \in \mathbb{R}} \left[t - (\beta_j^\dagger + \frac{1}{\|\tilde{X}_j\|_2^2} \tilde{X}_j^\top \boldsymbol{g}) \right]^2 .
\end{aligned}$$

Therefore, we have $\beta_j^\dagger = \beta_j^\dagger + \frac{1}{\|\tilde{X}_j\|_2^2}\tilde{X}_j^\top \boldsymbol{g}$. Now, let $\beta_j^\dagger \notin \mathcal{I}^\dagger$. If the perturbation t of $\beta_j^\dagger$ violates the constraints, that is, $\sum_{k\neq j} c_k I(\beta_k^\dagger) \leq C$, but $\sum_{k\neq j} c_k I(\beta_k^\dagger) + c_j > C$, then the minimizer is $t = 0$. Otherwise, $(\beta_1^\dagger, \cdots, \beta_{j-1}^\dagger, t, \beta_{j+1}^\dagger, \beta_p^\dagger) \in \mathcal{C}$ for all $t \in \mathbb{R}$, and hence the following holds, by the definition of $\beta^\dagger$,

$$\begin{aligned} 0 &= \arg\min_{t\in\mathbb{R}} L(\beta_1^\dagger, \ldots, \beta_{j-1}^\dagger, t, \beta_{j+1}^\dagger, \ldots, \beta_p^\dagger) \\ &= \arg\min_{t\in\mathbb{R}} Q(\boldsymbol{\beta}_{-j}^\dagger(t), \boldsymbol{\beta}^\dagger). \end{aligned}$$

This completes the first argument. For the second argument, as similar in the case in Theorem 1,

$$\begin{aligned} \boldsymbol{\beta}^\dagger &= \arg\min_{t_1,\cdots,t_p,\ \boldsymbol{t}\in\mathcal{C}} \Big[Q((t_1, \beta_2^\dagger, \ldots, \beta_p^\dagger), \boldsymbol{\beta}^\dagger) + Q((\beta_1^\dagger, t_2, \ldots, \beta_p^\dagger), \boldsymbol{\beta}^\dagger) + \cdots + Q((\beta_1, \beta_2^\dagger, \ldots, t_p), \boldsymbol{\beta}^\dagger)\Big] \\ &= \arg\min_{t_1,\cdots,t_p,\ \boldsymbol{t}\in\mathcal{C}} \sum_{j=1}^p \left[t_j - \left(\beta_j^\dagger + \frac{1}{\|X_j\|_2^2} X_j^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^\dagger)\right)\right]^2, \end{aligned}$$

which ends the proof. The first equality follows from the exactly same argument in the proof of Theorem 1. □

S1.3 Proof of Theorem 3

For the proof of Theorem 3, we introduce notation. We denote the submatrices with the index set $\mathcal{I}$ by $\mathbf{X}_\mathcal{I} \in \mathbb{R}^{n\times|\mathcal{I}|}$. The subvector of a vector $\boldsymbol{\beta}$ with index set $\mathcal{I}$ is denoted by $\boldsymbol{\beta}_\mathcal{I} \in \mathbb{R}^{|\mathcal{I}|}$. We denote a sparse vector by $\boldsymbol{\beta}|_\mathcal{I} \in \mathbb{R}^p$ where $(\boldsymbol{\beta}|_\mathcal{I})_j = \beta_j \mathbf{1}(j \in \mathcal{I})$.

We use technical lemmas from Huang et al. (2018) without any proof.

Lemma 1 (Lemmas 20, 21, and 22 (Huang et al., 2018)). *Let $\mathcal{I}, \mathcal{J}$ be disjoint subsets of $\{1, 2, \ldots, p\}$ with $|\mathcal{I}| = s_1$ and $|\mathcal{J}| = s_2$. Assume $\mathbf{X} \sim SRC(s_1 + s_2, b_-(s_1 + s_2), b_+(s_1 + s_2))$. Let θ_{s_1,s_2} be the sparse orthogonality constant. Then we have*

$$nb_-(s_1) \leq \|\mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}}\|_2 \leq nb_+(s_1), \tag{S1.4}$$

$$\frac{1}{nb_-(s_1)} \leq \|(\mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}})^{-1}\|_2 \leq \frac{1}{nb_+(s_1)}, \tag{S1.5}$$

$$\|\mathbf{X}_{\mathcal{I}}^\top\|_2 \leq \sqrt{nb_+(s_1)}, \tag{S1.6}$$

$$\theta_{s_1,s_2} \leq \max((b_+(s_1 + s_2) - 1), (1 - b_-(s_1 + s_2)), \tag{S1.7}$$

$$\mathbb{P}\left(\max_{|\mathcal{I}| \leq s} \|\mathbf{X}_{\mathcal{I}}^\top \boldsymbol{\zeta}\|_2 \leq \sigma\sqrt{s}\sqrt{2\log(p/\alpha)n}\right) \geq 1 - 2\alpha. \tag{S1.8}$$

We now define some notation that will be useful in proving Theorem 3. First,

$$h(q) := \max_{|\mathcal{I}| \leq q} \frac{\|\mathbf{X}_{\mathcal{I}}^\top \bar{\boldsymbol{\zeta}}\|_2}{n}, \tag{S1.9}$$

where $\bar{\boldsymbol{\zeta}} := \mathbf{X}(\boldsymbol{\beta}^{true} - \boldsymbol{\beta}^*) + \boldsymbol{\zeta}$. For simplicity, let $\mathcal{J}^* = (\mathcal{I}^*)^c$. Let $\mathcal{I}^k$ be the sequence of active sets generated by FCR (Algorithm 1). We define the two measures of interest to bound

$$D_2(\mathcal{I}^k) := \|\boldsymbol{\beta}^*|_{\mathcal{I}^* \setminus \mathcal{I}^k}\|_2, \quad \text{and} \quad \Delta^k := \boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^*|_{\mathcal{I}^k}.$$

We then define several sets of our interest at the k-th iteration

$$\mathcal{I}_1^k := \mathcal{I}^k \cap \mathcal{I}^*, \;\; \mathcal{I}_2^k := \mathcal{I}^* \setminus \mathcal{I}^k, \;\; \mathcal{J}_3^k := \mathcal{I}^k \cap \mathcal{J}^*, \;\; \mathcal{J}_4^k := \mathcal{J}^* \setminus \mathcal{J}_3^k,$$

and $(k+1)$-th iteration

$$\mathcal{I}_{11}^k := \mathcal{I}_1^k \setminus \mathcal{I}^{k+1}, \;\; \mathcal{I}_{22}^k := \mathcal{I}_2^k \setminus \mathcal{I}^{k+1}, \;\; \mathcal{J}_{33}^k := \mathcal{I}^{k+1} \cap \mathcal{J}_3^k, \;\; \mathcal{J}_{44}^k := \mathcal{I}^{k+1} \cap \mathcal{J}_4^k.$$

Lemma 2. *Suppose $|\mathcal{I}^k| \le q$. If $\mathbf{X} \sim SRC(q, b_-(q), b_+(q))$, we have*

$$\|\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^*\|_2 \le \Big(1 + \frac{\theta_{q,q}}{b_-(q)}\Big) D_2(\mathcal{I}^k) + \frac{h(q)}{b_-(q)}, \tag{S1.10}$$

$$\|\Delta^k\|_2 \le \frac{\theta_{q,q}}{b_-(q)} \|\boldsymbol{\beta}^*|_{\mathcal{I}_2^k}\|_2 + \frac{h(q)}{b_-(q)}. \tag{S1.11}$$

Proof of Lemma 2. We have

$$\begin{aligned}
\boldsymbol{\beta}_{\mathcal{I}^k}^{k+1} &= (\mathbf{X}_{\mathcal{I}^k}^\top \mathbf{X}_{\mathcal{I}^k})^{-1} \mathbf{X}_{\mathcal{I}^k}^\top \boldsymbol{y} \\
&= (\mathbf{X}_{\mathcal{I}^k}^\top \mathbf{X}_{\mathcal{I}^k})^{-1} \mathbf{X}_{\mathcal{I}^k}^\top (\mathbf{X}_{\mathcal{I}_1^k} \boldsymbol{\beta}_{\mathcal{I}_1^k}^* + \mathbf{X}_{\mathcal{I}_2^k} \boldsymbol{\beta}_{\mathcal{I}_2^k}^* + \bar{\boldsymbol{\zeta}}), \qquad \text{(S1.12)} \\
(\boldsymbol{\beta}^*|_{\mathcal{I}^k})_{\mathcal{I}^k} &= (\mathbf{X}_{\mathcal{I}^k}^\top \mathbf{X}_{\mathcal{I}^k})^{-1} \mathbf{X}_{\mathcal{I}^k}^\top \mathbf{X}_{\mathcal{I}^k} (\boldsymbol{\beta}^*|_{\mathcal{I}^k})_{\mathcal{I}^k} \\
&= (\mathbf{X}_{\mathcal{I}^k}^\top \mathbf{X}_{\mathcal{I}^k})^{-1} \mathbf{X}_{\mathcal{I}^k}^\top (\mathbf{X}_{\mathcal{I}_1^k} \boldsymbol{\beta}_{\mathcal{I}_1^k}^*), \qquad \text{(S1.13)}
\end{aligned}$$

where the first equality uses the definition of β^{k+1} and the third equality is

from simple algebra. Therefore,

$$\begin{aligned}
\|\Delta^k\|_2 &= \|\boldsymbol{\beta}^{k+1}_{\mathcal{I}^k} - (\boldsymbol{\beta}^*|_{\mathcal{I}^k})_{\mathcal{I}^k}\|_2 \\
&= \|(\mathbf{X}^\top_{\mathcal{I}^k}\mathbf{X}_{\mathcal{I}^k})^{-1}\mathbf{X}^\top_{\mathcal{I}^k}(\mathbf{X}_{\mathcal{I}^k_2}\boldsymbol{\beta}^*_{\mathcal{I}^k_2} + \bar{\boldsymbol{\zeta}})\|_2 \\
&\leq \frac{1}{nb_-(q)}\Big(\|\mathbf{X}^\top_{\mathcal{I}^k}\mathbf{X}_{\mathcal{I}^k_2}\boldsymbol{\beta}^*_{\mathcal{I}^k_2}\|_2 + \|\mathbf{X}^\top_{\mathcal{I}^k}\bar{\boldsymbol{\zeta}}\|_2\Big) \\
&\leq \frac{\theta_{q,q}}{b_-(q)}\|\boldsymbol{\beta}^*|_{\mathcal{I}^k_2}\|_2 + \frac{h(q)}{b_-(q)} \\
&= \frac{\theta_{q,q}}{b_-(q)}D_2(\mathcal{I}^k) + \frac{h(q)}{b_-(q)},
\end{aligned}$$

where the first equality uses $\text{supp}(\boldsymbol{\beta}^{k+1}) = \mathcal{I}^k$, the second inequality follows from (S1.12) and (S1.13). The first inequality comes from (S1.5) and the triangle inequality, and the second inequality follows from the definition of $\theta_{a,b}$, and $h_2(q)$, which completes the proof of (S1.11). The third equality follows from the definition of $D_2(\mathcal{I}^k)$. Then the triangle inequality $\|\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^*\|_2 \leq \|\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^*|_{\mathcal{I}^k}\|_2 + \|\boldsymbol{\beta}^*\ |_{\mathcal{I}^*\setminus\mathcal{I}^k}\ \|_2$ and (S1.11) implies (S1.10). □

Lemma 3.

$$D_2(\mathcal{I}^{k+1}) \leq \|\boldsymbol{\beta}^*_{\mathcal{I}^k_{11}}\|_2 + \|\boldsymbol{\beta}^*_{\mathcal{I}^k_{22}}\|_2, \tag{S1.14}$$

$$\|\boldsymbol{\beta}^*_{\mathcal{I}^k_{11}}\|_2 \leq \|\Delta^k_{\mathcal{I}^k_{11}}\|_2 + \|\boldsymbol{\beta}^{k+1}_{\mathcal{I}^k_{11}}\|_2, \tag{S1.15}$$

$$\|\boldsymbol{\beta}^*_{\mathcal{I}^k_{22}}\|_2 \leq \frac{\|\boldsymbol{d}^{k+1}_{\mathcal{I}^k_{22}}\|_2 + \theta_{q,q}\|\Delta^k_{\mathcal{I}^k}\|_2 + \theta_{q,q}D_2(\mathcal{I}^k) + h(q)}{b_-(q)}, \quad \textit{if } \mathbf{X} \sim SRC(q, b_-(q), b_+(q)). \tag{S1.16}$$

Proof. By the definition of $D_2(\mathcal{I}^k)$, we have

$$D_2(\mathcal{I}^k) = \|\boldsymbol{\beta}^*|_{\mathcal{I}_2^k} = \|\boldsymbol{\beta}^*|_{\mathcal{I}_{11}^k \cup \mathcal{I}_{22}^k}\|_2 \leq \|\boldsymbol{\beta}^*_{\mathcal{I}_{11}^k}\|_2 + \|\boldsymbol{\beta}^*_{\mathcal{I}_{22}^k}\|_2,$$

which prove (S1.14). Also, using $\Delta^k = \boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^*|_{\mathcal{I}^k}$, we have

$$\|\boldsymbol{\beta}^{k+1}_{\mathcal{I}_{11}^k}\|_2 = \|\big(\boldsymbol{\beta}^*|_{\mathcal{I}^k}\big)_{\mathcal{I}_{11}^k} + \Delta^k_{A_{11}^k}\|_2 \geq \|\boldsymbol{\beta}^*_{\mathcal{I}_{11}^k}\|_2 - \|\Delta^k_{A_{11}^k}\|_2,$$

which prove (S1.15). For (S1.16), we consider

$$\begin{aligned}
\|\boldsymbol{d}^{k+1}_{\mathcal{I}_{22}^k}\|_2 &= \|\mathbf{X}^\top_{\mathcal{I}_{22}^k}\Big(\mathbf{X}_{\mathcal{I}^k}\boldsymbol{\beta}^{k+1}_{\mathcal{I}^k} - \boldsymbol{y}\Big)/n\|_2 \\
&= \|\mathbf{X}^\top_{\mathcal{I}_{22}^k}\Big(\mathbf{X}_{\mathcal{I}^k}\Delta^k_{\mathcal{I}^k} + \mathbf{X}_{\mathcal{I}^k}\boldsymbol{\beta}^*_{\mathcal{I}^k} - \mathbf{X}_{\mathcal{I}^*}\boldsymbol{\beta}^*_{\mathcal{I}^*} - \bar{\boldsymbol{\zeta}}\Big)/n\|_2 \\
&= \|\mathbf{X}^\top_{\mathcal{I}_{22}^k}\Big(\mathbf{X}_{\mathcal{I}^k}\Delta^k_{\mathcal{I}^k} - \mathbf{X}_{\mathcal{I}_{22}^k}\boldsymbol{\beta}^*_{\mathcal{I}_{22}^k} - \mathbf{X}_{\mathcal{I}_2^k\setminus\mathcal{I}_{22}^k}\boldsymbol{\beta}^*_{\mathcal{I}_2^k\setminus\mathcal{I}_{22}^k} - \bar{\boldsymbol{\zeta}}\Big)/n\|_2 \\
&\geq b_-(|\mathcal{I}_{22}^k|)\|\boldsymbol{\beta}^*_{\mathcal{I}_{22}^k}\|_2 - \theta_{|\mathcal{I}_{22}^k|,q}\|\Delta^k_{\mathcal{I}^k}\|_2 - \theta_{|\mathcal{I}_2^k|,|\mathcal{I}_2^k\setminus\mathcal{I}_{22}^k|}\|\boldsymbol{\beta}^*_{\mathcal{I}_2^k\setminus\mathcal{I}_{22}^k}\|_2 - \|\mathbf{X}^\top_{\mathcal{I}_{22}^k}\bar{\boldsymbol{\zeta}}/n\|_2 \\
&\geq b_-(q)\|\boldsymbol{\beta}^*_{\mathcal{I}_{22}^k}\|_2 - \theta_{q,q}\|\Delta^k_{\mathcal{I}^k}\|_2 - \theta_{q,q}D_2(\mathcal{I}^k) - h(q),
\end{aligned}$$

where the first equality uses the definition of $\boldsymbol{d}^{k+1}$, and the third equality follows from simple algebra. The first inequality uses (S1.4), and the definition of θ_{s_1,s_2}. The last inequality follows from the definition of $h(\cdot)$ and the monotonicity of $b_-(\cdot)$. This completes the proof. □

Lemma 4.

$$\|\boldsymbol{\beta}^{k+1}_{\mathcal{I}_{11}^k}\|_2 + \|\boldsymbol{d}^{k+1}_{\mathcal{I}_{22}^k}\|_2 \leq \sqrt{2}\left(\|\boldsymbol{\beta}^{k+1}_{\mathcal{J}_{33}^k}\|_2 + \|\boldsymbol{d}^{k+1}_{\mathcal{J}_{44}^k}\|_2\right). \tag{S1.17}$$

Proof of lemma 4. For inequality (S1.17), we first notice the relationship

between the true index set and the k-th index set from the knapsack algorithm such that

$$\|(\boldsymbol{\beta}^{k+1} + \boldsymbol{d}^{k+1})_{\mathcal{I}^{k+1}}\|_2^2 \geq \|(\boldsymbol{\beta}^{k+1} + \boldsymbol{d}^{k+1})_{\mathcal{I}^*}\|_2^2.$$

Reducing the contribution of $\mathcal{I}^{k+1} \cap \mathcal{I}^*$, we obtain

$$\|(\boldsymbol{\beta}^{k+1} + \boldsymbol{d}^{k+1})_{\mathcal{I}^{k+1}\setminus\mathcal{I}^*}\|_2^2 \geq \|(\boldsymbol{\beta}^{k+1} + \boldsymbol{d}^{k+1})_{\mathcal{I}^*\setminus\mathcal{I}^{k+1}}\|_2^2. \tag{S1.18}$$

Since $\beta_j^k d_j^k = 0$ for $j \in \{1, \ldots, p\}, \forall k \geq 1$, the both side of (S1.18) can be decomposed. We have

$$\|\boldsymbol{\beta}^{k+1}_{\mathcal{I}^{k+1}\setminus\mathcal{I}^*}\|_2^2 + \|\boldsymbol{d}^{k+1}_{\mathcal{I}^{k+1}\setminus\mathcal{I}^*}\|_2^2 \geq \|\boldsymbol{\beta}^{k+1}_{\mathcal{I}^*\setminus\mathcal{I}^{k+1}}\|_2^2 + \|\boldsymbol{d}^{k+1}_{\mathcal{I}^*\setminus\mathcal{I}^{k+1}}\|_2^2. \tag{S1.19}$$

We obtain

$$\begin{aligned}
\|\boldsymbol{d}^{k+1}_{\mathcal{I}^k_{22}}\|_2^2 &= \|\boldsymbol{d}^{k+1}_{\mathcal{I}^*\setminus\mathcal{I}^k\setminus\mathcal{I}^{k+1}}\|_2^2 \\
&= \|\boldsymbol{d}^{k+1}_{\mathcal{I}^*\setminus\mathcal{I}^{k+1}}\|_2^2 \\
&\leq \|\boldsymbol{\beta}^{k+1}_{\mathcal{I}^{k+1}\setminus\mathcal{I}^*}\|_2^2 + \|\boldsymbol{d}^{k+1}_{\mathcal{I}^{k+1}\setminus\mathcal{I}^*}\|_2^2 - \|\boldsymbol{\beta}^{k+1}_{\mathcal{I}^*\setminus\mathcal{I}^{k+1}}\|_2^2 \\
&\leq \|\boldsymbol{\beta}^{k+1}_{\mathcal{J}^k_{44}}\|_2^2 + \|\boldsymbol{\beta}^{k+1}_{\mathcal{J}^k_{33}}\|_2^2 + \|\boldsymbol{d}^{k+1}_{\mathcal{J}^k_{44}}\|_2^2 + \|\boldsymbol{d}^{k+1}_{\mathcal{J}^k_{33}}\|_2^2 - \|\boldsymbol{\beta}^{k+1}_{\mathcal{I}^k_{11}}\|_2^2.
\end{aligned}$$

The first equality comes from the definition of $\mathcal{I}^k_{22}$, the second from the property of FCR $\boldsymbol{d}^{k+1}_{\mathcal{I}^k} = \mathbf{0}$, the third inequality from (S1.19). For the fourth inequality, we use $\mathcal{I}^{k+1} \setminus \mathcal{I}^* = \mathcal{J}^k_{44} \cup \mathcal{J}^k_{33}$ and $\mathcal{I}^k_{11} \subset \mathcal{I}^* \setminus \mathcal{I}^{k+1}$. We note that $\boldsymbol{\beta}^{k+1}_{\mathcal{J}^k_{44}} = \mathbf{0}$ and $\boldsymbol{d}^{k+1}_{\mathcal{J}^k_{33}} = \mathbf{0}$ since $\mathcal{J}^k_{44} \cup \mathcal{I}^k = \emptyset$ and $\mathcal{J}^k_{33} \subset \mathcal{I}^k$. It leads us

to obtain

$$\|\boldsymbol{\beta}^{k+1}_{\mathcal{I}^k_{11}}\|_2^2 + \|\boldsymbol{d}^{k+1}_{\mathcal{I}^k_{22}}\|_2^2 \leq \|\boldsymbol{\beta}^{k+1}_{\mathcal{J}^k_{33}}\|_2^2 + \|\boldsymbol{d}^{k+1}_{\mathcal{J}^k_{44}}\|_2^2.$$

Therefore,

$$\begin{aligned}\frac{1}{2}\left(\|\boldsymbol{\beta}^{k+1}_{\mathcal{I}^k_{11}}\|_2 + \|\boldsymbol{d}^{k+1}_{\mathcal{I}^k_{22}}\|_2\right)^2 &\leq \left(\|\boldsymbol{\beta}^{k+1}_{\mathcal{I}^k_{11}}\|_2^2 + \|\boldsymbol{d}^{k+1}_{\mathcal{I}^k_{22}}\|_2^2\right) \\ &\leq \|\boldsymbol{\beta}^{k+1}_{\mathcal{J}^k_{33}}\|_2^2 + \|\boldsymbol{d}^{k+1}_{\mathcal{J}^k_{44}}\|_2^2 \\ &\leq \left(\|\boldsymbol{\beta}^{k+1}_{\mathcal{J}^k_{33}}\|_2 + \|\boldsymbol{d}^{k+1}_{\mathcal{J}^k_{44}}\|_2\right)^2,\end{aligned}$$

which completes the proof. □

Lemma 5.

$$\|\beta^{k+1}_{\mathcal{J}^k_{33}}\|_2 \leq \|\Delta^k_{\mathcal{J}^k_{33}}\|_2. \tag{S1.20}$$

Furthermore, We have

$$\|\boldsymbol{d}^{k+1}_{\mathcal{J}^k_{44}}\| \leq \theta_{q,q}\|\Delta^k_{\mathcal{I}^k}\| + \theta_{q,q}D_2(\mathcal{I}^k) + h(q), \quad \textit{if } \mathbf{X} \sim SRC(q, b_-(q), b_+(q)). \tag{S1.21}$$

Proof. The definition of Δ^k gives

$$\|\boldsymbol{\beta}^{k+1}_{\mathcal{J}^k_{33}}\|_2 = \|\Delta^k_{\mathcal{J}^k_{33}} + \boldsymbol{\beta}^*_{\mathcal{J}^k_{33}}\|_2 = \|\Delta^k_{\mathcal{J}^k_{33}}\|_2,$$

since $\boldsymbol{\beta}^*$ vanishes on $\mathcal{J}^k_{33}$, which gives the inequality (S1.20). For inequality

(S1.21), $\boldsymbol{d}$ in the FCR algorithm gives

$$\begin{aligned}
\|\boldsymbol{d}^{k+1}_{\mathcal{J}^k_{44}}\| &= \|\mathbf{X}^\top_{\mathcal{J}^k_{44}}\Big(\mathbf{X}_{\mathcal{I}^k}\boldsymbol{\beta}^{k+1}_{\mathcal{I}^k} - \boldsymbol{y}\Big)\|_2 \\
&= \|\mathbf{X}^\top_{\mathcal{I}^k_{22}}\Big(\mathbf{X}_{\mathcal{I}^k}\Delta^k_{\mathcal{I}^k} + \mathbf{X}_{\mathcal{I}^k}\boldsymbol{\beta}^*_{\mathcal{I}^k} - \mathbf{X}_{\mathcal{I}^*}\boldsymbol{\beta}^*_{\mathcal{I}^*} - \bar{\boldsymbol{\zeta}}\Big)\|_2 \\
&= \|\mathbf{X}^\top_{\mathcal{J}^k_{44}}\Big(\mathbf{X}_{\mathcal{I}^k}\Delta^k_{\mathcal{I}^k} - \mathbf{X}_{\mathcal{I}^k_2}\boldsymbol{\beta}^*_{\mathcal{I}^k_2} - \bar{\boldsymbol{\zeta}}\Big)\|_2 \\
&\leq \theta_{|\mathcal{J}^k_{44}|,q}\|\Delta^k_{\mathcal{I}^k}\|_2 + \theta_{|\mathcal{J}^k_{44}|,|\mathcal{I}^k_2|}\|\boldsymbol{\beta}^*_{\mathcal{I}^k_2}\|_2 + \|\mathbf{X}^\top_{\mathcal{J}^k_{44}}\bar{\boldsymbol{\zeta}}\|_2 \\
&\leq \theta_{q,q}\|\Delta^k_{\mathcal{I}^k}\|_2 + \theta_{q,q}D_2(\mathcal{I}^k) + h(q).
\end{aligned}$$

In the fourth inequality, we employ the definition of θ_{s_1,s_2}, and in the fifth inequality, we use its monotonicity property. □

Lemma 6. *If* $\mathbf{X} \sim SRC(q, b_-(q), b_+(q))$*, then*

$$D_2(\mathcal{I}^{k+1}) \leq rD_2(\mathcal{I}^k) + \frac{r}{\theta_{q,q}}h(q). \tag{S1.22}$$

Proof of Lemma 6.

$$\begin{aligned}
D_2(\mathcal{I}^{k+1}) &\leq \Big(\|\boldsymbol{\beta}^*_{\mathcal{I}^k_{11}}\|_2 + \|\boldsymbol{\beta}^*_{\mathcal{I}^k_{22}}\|_2\Big)/b_-(q) \\
&\leq \Big(\|\boldsymbol{\beta}^*_{\mathcal{I}^k_{11}}\|_2 + \|\boldsymbol{d}^*_{\mathcal{I}^k_{22}}\|_2 + \|\Delta^k_{\mathcal{I}^k_{11}}\|_2 + \theta_{q,q}\|\Delta^k_{\mathcal{I}^k}\|_2 + \theta_{q,q}D_2(\mathcal{I}^k) + h(q)\Big)/b_-(q) \\
&\leq \Big(\sqrt{2}(\|\boldsymbol{\beta}^{k+1}_{\mathcal{J}^k_{33}}\|_2 + \|\boldsymbol{d}^{k+1}_{\mathcal{J}^k_{44}}\|_2) + \|\Delta^k_{\mathcal{I}^k_{11}}\|_2 + \theta_{q,q}\|\Delta^k_{\mathcal{I}^k}\| + \theta_{q,q}D_2(\mathcal{I}^k) + h(q)\Big)/b_-(q) \\
&\leq \Big((\sqrt{2} + (1+\sqrt{2})\theta_{q,q})\|\Delta^k\|_2 + (1+\sqrt{2})\theta_{q,q}D_2(\mathcal{I}^k) + (1+\sqrt{2})h(q)\Big)/b_-(q) \\
&\leq \Big(\frac{\sqrt{2}\theta_{q,q} + (1+\sqrt{2})\theta^2_{q,q}}{b_-(q)^2} + \frac{(1+\sqrt{2})\theta_{q,q}}{b_-(q)}\Big)D_2(\mathcal{I}^k) \\
&\qquad + \Big(\frac{\sqrt{2} + (1+\sqrt{2})\theta_{q,q}}{b_-(q)^2} + \frac{1+\sqrt{2}}{b_-(q)}\Big)h(q).
\end{aligned}$$

The initial inequality is derived from equation (S1.14), the subsequent inequality is a result of referencing equations (S1.15) and (S1.16), the third inequality is obtained by referring to equation (S1.17), the fourth inequality is established by combining the sums of equations (S1.20) and (S1.21), and the final inequality is derived from equation (S1.11). □

Lemma 7. *Let $\boldsymbol{\beta}^*$ be the solution for the optimization problem (1.1). Then, we have*

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{true}\|_2 \leq \sqrt{\frac{b_+(p_0)}{b_-(p_0+q)}}\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^{true}\|_2.$$

Proof. First, we note that when the loss function is l_2, the optimization problem (1.1) is equivalent to

$$\min_{\boldsymbol{\beta}}(\boldsymbol{\beta} - \boldsymbol{\beta}^{true})^\top \Sigma(\boldsymbol{\beta} - \boldsymbol{\beta}^{true}), \tag{S1.23}$$

where $\Sigma = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top]$ is a $p \times p$ matrix by equation (3) in Yu et al. (2022). Let us define the nonzero index set of $\boldsymbol{\beta}^{true}$

$$\mathcal{I}_0 := \{1, 2, \ldots, p_0\}.$$

Then, from

$$b_-(p_0+q)\cdot I \leq \mathbf{X}^\top_{\mathcal{I}^*\cup\mathcal{I}_0}\mathbf{X}_{\mathcal{I}^*\cup\mathcal{I}_0} \leq b_+(p_0+q)\cdot I$$

$$b_-(p_0)\cdot I \leq \mathbf{X}^\top_{\mathcal{I}_0}\mathbf{X}_{\mathcal{I}_0} \leq b_+(p_0)\cdot I,$$

taking the expectation at both side will give

$$\frac{1}{n-1}b_-(p_0+q)\cdot I \le \Sigma_{\mathcal{I}^*\cup\mathcal{I}_0} \le \frac{1}{n-1}b_+(p_0+q)\cdot I, \tag{S1.24}$$

$$\frac{1}{n-1}b_-(p_0)\cdot I \le \Sigma_{\mathcal{I}_0} \le \frac{1}{n-1}b_+(p_0)\cdot I. \tag{S1.25}$$

where $\Sigma_{\mathcal{I}}$ is the $|\mathcal{I}|\times|\mathcal{I}|$ sub-matrix. Then we have

$$\begin{aligned}
b_-(p_0+q)\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^{true}\|_2^2 &\le (\boldsymbol{\beta}^*-\boldsymbol{\beta}^{true})^\top_{\mathcal{I}^*\cup\mathcal{I}_0}\Sigma_{\mathcal{I}^*\cup\mathcal{I}_0}(\boldsymbol{\beta}^*-\boldsymbol{\beta}^{true})_{\mathcal{I}^*\cup\mathcal{I}_0}\\
&\le (\boldsymbol{\beta}^*-\boldsymbol{\beta}^{true})^\top\Sigma(\boldsymbol{\beta}^*-\boldsymbol{\beta}^{true})\\
&\le (\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}^{true})^\top\Sigma(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}^{true})\\
&\le (\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}^{true})^\top_{\mathcal{I}_0}\Sigma_{\mathcal{I}_0}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}^{true})_{\mathcal{I}_0}\\
&\le b_+(p_0)\|\boldsymbol{\beta}^{true}-\bar{\boldsymbol{\beta}}\|_2^2.
\end{aligned}$$

The first inequality follows from (S1.24), the second and fourth from simple algebra, the third from (S1.23), and the final from (S1.25). This completes the proof. □

Lemma 8. *With probability* $1-2\alpha$*, we have*

$$h(q) \le \sqrt{\frac{b_+(q)b_+(p_0)b_+(p_0+q)}{b_-(p_0+q)}}\|\boldsymbol{\beta}^{true}-\bar{\boldsymbol{\beta}}\|_2 + \sigma\sqrt{\frac{2q\log(p/\alpha)}{n}}. \tag{S1.26}$$

Proof. By the definition of $h(q)$ in (S1.9)

$$h(q) \le \max_{|\mathcal{I}|\le q}\frac{\|\mathbf{X}_{\mathcal{I}}^\top\mathbf{X}(\boldsymbol{\beta}^{true}-\boldsymbol{\beta}^*)\|_2}{n} + \max_{|\mathcal{I}|\le q}\frac{\|\mathbf{X}_{\mathcal{I}}^\top\boldsymbol{\zeta}\|_2}{n}. \tag{S1.27}$$

As for the first term of (S1.27), we have

$$\begin{aligned}\max_{|\mathcal{I}|\leq q}\frac{\|\mathbf{X}_{\mathcal{I}}^{\top}\mathbf{X}(\boldsymbol{\beta}^{true}-\boldsymbol{\beta}^{*})\|_2}{n} &\leq \max_{|\mathcal{I}|\leq q}\frac{\|\mathbf{X}_{\mathcal{I}}^{\top}\|_2\|\mathbf{X}(\boldsymbol{\beta}^{true}-\boldsymbol{\beta}^{*})\|_2}{n}\\ &\leq \sqrt{b_+(q)}\frac{\|\mathbf{X}_{\mathcal{I}^*\cup\mathcal{I}_0}(\boldsymbol{\beta}^{true}-\boldsymbol{\beta}^{*})_{\mathcal{I}^*\cup\mathcal{I}_0}\|_2}{\sqrt{n}}\\ &\leq \sqrt{b_+(q)b_+(p_0+q)}\|\boldsymbol{\beta}^{true}-\boldsymbol{\beta}^{*}\|_2\\ &\leq \sqrt{\frac{b_+(q)b_+(p_0)b_+(p_0+q)}{b_-(p_0+q)}}\|\boldsymbol{\beta}^{true}-\bar{\boldsymbol{\beta}}\|_2.\end{aligned}$$

The fourth inequality comes from Lemma 7. The second term of (S1.27) gives

$$\max_{|\mathcal{I}|\leq q}\frac{\|\mathbf{X}_{\mathcal{I}}^{\top}\boldsymbol{\zeta}\|_2}{n}\leq\sigma\sqrt{\frac{2q\log(p/\alpha)}{n}}.$$

with probability at least $1-2\alpha$ by (S1.8), which completes the proof.

□

Proof of Theorem 3. Suppose $r<1$. Using (S1.22) recursively,

$$\begin{aligned}\|\boldsymbol{\beta}^{*}|_{\mathcal{I}^*\setminus\mathcal{I}^{k+1}}\|_2 = D_2(\mathcal{I}^{k+1}) &\leq rD_2(\mathcal{I}^{k})+\frac{r}{\theta_{q,q}}h(q)\\ &\leq\cdots\\ &\leq r^{k+1}D_2(\mathcal{I}^{0})+\frac{r}{(1-r)\theta_{q,q}}h(q)\\ &< r^{k+1}\|\boldsymbol{\beta}^{*}\|_2+\frac{r}{(1-r)\theta_{q,q}}h(q). \qquad \text{(S1.28)}\end{aligned}$$

Now, we also have

$$\begin{aligned}\|\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^*\|_2 &\leq \Big(1+\frac{\theta_{q,q}}{b_-(q)}\Big)D_2(\mathcal{I}^k)+\frac{h(q)}{b_-(q)}\\ &\leq \Big(1+\frac{\theta_{q,q}}{b_-(q)}\Big)\Big(r^k\|\boldsymbol{\beta}^*\|_2+\frac{r}{(1-r)\theta_{q,q}}h(q)\Big)+\frac{h(q)}{b_-(q)}\\ &= \Big(1+\frac{\theta_{q,q}}{b_-(q)}\Big)r^k\|\boldsymbol{\beta}^*\|_2+\left(\frac{r}{(1-r)\theta_{q,q}}\left(1+\frac{\theta_{q,q}}{b_-(q)}\right)+\frac{1}{b_-(q)}\right)h(q).\end{aligned} \tag{S1.29}$$

The initial inequality is a result of referencing equation (S1.10), the subsequent inequality is derived from the definition of $D_2(\mathcal{I}^k)$, and the third line is the outcome of some algebraic calculation.

Hence, from equations (S1.28) and (S1.29), with probability $1-2\alpha$,

$$\begin{aligned}\|\boldsymbol{\beta}^*|_{\mathcal{I}^*\setminus\mathcal{I}^{k+1}}\|_2 \leq r^{k+1}\|\boldsymbol{\beta}^*\|_2+\frac{r}{(1-r)\theta_{q,q}}\sqrt{\frac{b_+(q)b_+(p_0)b_+(p_0+q)}{b_-(p_0+q)}}\|\boldsymbol{\beta}^{true}-\bar{\boldsymbol{\beta}}\|_2\\ +\frac{r}{(1-r)\theta_{q,q}}\epsilon,\end{aligned}$$

and

$$\begin{aligned}\|\boldsymbol{\beta}^{k+1}-\boldsymbol{\beta}^*\|_2 &\leq \Big(1+\frac{\theta_{q,q}}{b_-(q)}\Big)r^k\|\boldsymbol{\beta}^*\|_2+\left(\frac{r}{(1-r)\theta_{q,q}}\left(1+\frac{\theta_{q,q}}{b_-(q)}\right)+\frac{1}{b_-(q)}\right)h(q)\\ &\leq \Big(1+\frac{\theta_{q,q}}{b_-(q)}\Big)r^k\|\boldsymbol{\beta}^*\|_2\\ &+\left(\frac{r}{(1-r)\theta_{q,q}}\left(1+\frac{\theta_{q,q}}{b_-(q)}\right)+\frac{1}{b_-(q)}\right)\sqrt{\frac{b_+(q)b_+(p_0)b_+(p_0+q)}{b_-(p_0+q)}}\|\boldsymbol{\beta}^{true}-\bar{\boldsymbol{\beta}}\|_2\\ &+\left(\frac{r}{(1-r)\theta_{q,q}}\left(1+\frac{\theta_{q,q}}{b_-(q)}\right)+\frac{1}{b_-(q)}\right)\epsilon.\end{aligned}$$

□

S1.4 Proof of Corollary 1

From the assumption that $\sum_{j=1}^{p_0} c_j I(\beta_j^{true}) \leq C$, we have, due to $\boldsymbol{\beta}^{true} \in \mathcal{C}$,

$$R_{\mathcal{C}} := \|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^{true}\|_2 = 0.$$

Also, from (S1.23), $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{true}$ because $(\boldsymbol{\beta}^* - \boldsymbol{\beta}^{true})^\top \Sigma (\boldsymbol{\beta}^* - \boldsymbol{\beta}^{true}) \geq 0$ and attains 0 when $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{true}$. Equation (3.16) in the main article gives us

$$\|\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^*\|_2 \leq b_3 r^k \|\boldsymbol{\beta}^*\|_2 + b_4 \epsilon.$$

If $k \geq log_{1/r} \frac{b_3 \|\boldsymbol{\beta}^*\|_2}{b_4 \epsilon}$, we have $b_3 r^k \|\boldsymbol{\beta}^*\|_2 \leq b_4 \epsilon$, which completes the first argument.

Now, from (3.15) in the main article, we have

$$\begin{aligned}
\|\boldsymbol{\beta}^*|_{\mathcal{I}^* \setminus \mathcal{I}^k}\|_2 &\leq r^k \|\boldsymbol{\beta}^*\|_2 + b_1 \epsilon \\
&\leq (1-\xi)\bar{m} + b_1 \epsilon \quad \text{if } k \geq \log_{1/r} \frac{\|\boldsymbol{\beta}^*\|_2}{(1-\xi)\bar{m}} \\
&< (1-\xi)\bar{m} + \xi \bar{m} \\
&\leq \bar{m}.
\end{aligned}$$

Then, from $\|\boldsymbol{\beta}^*|_{\mathcal{I}^* \setminus \mathcal{I}^k}\|_2 < \bar{m}$, we have $\mathcal{I}^k \subset \mathcal{I}^*$.

S2 The QM* condition

S2.1 Sufficient conditions

The following lemma characterizes a class of loss functions that satisfies the QM* condition.

Lemma 9. *Assume $\Phi(y, \mu)$ is differentiable with respect to μ and write $\Phi'_\mu = \frac{\partial \Phi(y,\mu)}{\partial \mu}$. Then*

$$\nabla L(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \Phi'_\mu(y_i, \boldsymbol{x}_i^\top \boldsymbol{\beta}) \boldsymbol{x}_i.$$

(1) If Φ'_μ is Lipschitz continuous with constant C_1 such that

$$|\Phi'_\mu(y, \mu_1) - \Phi'_\mu(y, \mu_2)| \leq C_1 |\mu_1 - \mu_2| \quad \forall\ y, \mu_1, \mu_2 \in \mathbb{R},$$

then the QM condition holds for $\mathbf{H} = \frac{2C_1}{n} I_p$.*

(2) If $\Phi''_\mu = \frac{\partial \Phi^2(y,\mu)}{\partial \mu^2}$ exists and

$$\Phi''_\mu \leq C_2 \quad \forall\ y, \mu \in \mathbb{R},$$

then the QM condition holds for Φ and $\mathbf{H} = \frac{C_2}{n}$.*

S2.2 Examples

The function hsvm_δ is defined in Yang and Zou (2015).

Loss	$L(\boldsymbol{\beta})$	$\boldsymbol{g}_i(\boldsymbol{\beta})$	$\mathbf{H}$
Logistic (-1, 1 loss)	$\frac{1}{n}\sum \log(1+\exp(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}))$	$\frac{1}{n} \cdot \frac{1}{1+\exp(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta})}$	$\frac{1}{4n} I_p$
Logistic (0, 1 loss)	$\frac{1}{n}\sum -y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log\left(1+\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i)\right)$	$\frac{1}{n}\left(-y_i + \frac{\exp(\boldsymbol{x}_i \boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i \boldsymbol{\beta})}\right)$	$\frac{1}{4n} I_p$
Squared hinge loss	$\frac{1}{n}\sum \left((1-y_i \boldsymbol{x}_i^\top \boldsymbol{\beta})_+\right)^2$	$\frac{1}{n}\left(y_i(1-y_i \boldsymbol{x}_i^\top \boldsymbol{\beta})_+\right)$	$\frac{4}{n} I_p$
Huberized hinge loss	$\frac{1}{n}\sum \text{hsvm}_\delta(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta})$	$\frac{1}{n} y_i \text{hsvm}'_\delta(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta})$	$\frac{2}{n\delta} I_p$

S3 Stopping criteria

The FCR algorithm raises an important point about the stopping criteria for this algorithm. Specifically, we observe that given a finite set of subsets defined over integers $\{1, 2, \ldots, p\}$ that satisfy the cost restrictions $\sum_{j=1}^{p} c_j I(\beta_j) \leq C$, there exists only a limited number of active sets $\mathcal{I}'s$. As a result, the algorithm has *eventual periodicity*, that is, there exist integers m and $\tilde{m}$, which satisfy $\mathcal{I}^{m+\tilde{m}} = \mathcal{I}^m$ (Foucart, 2011). This periodicity in the set of the active set $\mathcal{I}'s$ leads to the corresponding outcome of $\boldsymbol{\beta}^{m+\tilde{m}} = \boldsymbol{\beta}^m$ within the algorithm framework.

Using periodicity, we establish stopping criteria for the algorithm. We terminate the algorithm when the difference in loss from the previous it-

eration falls below a predetermined tolerance, denoted δ. However, even with this criterion in place, the FCR algorithm tends to exhibit fluctuations around local minima and does not necessarily converge to a specific point. In the context of the coordinate descent algorithm, FCR updates all coordinates simultaneously, as opposed to one-by-one, which generally does not guarantee stable convergence. To address this, we introduce a criterion to detect periodic behavior while the algorithm is running. If periodicity is detected, we terminate the algorithm and select the active set $\mathcal{I}$ and its corresponding $\boldsymbol{\beta}$ where the objective function is minimized during the course of the algorithm. The FCR algorithm, when applied with these termination conditions, demonstrates superior performance compared to other existing methods, as detailed in Section 4.

S4 Grouped cost

It is a typical scenario in which purchasing a single component is not possible without acquiring all the variables. Buying a variable from a survey costs the same amount as buying all survey questionnaires. In a clinical data set, demographic information about the patient is usually collected all at once, rather than separately, to minimize communication expenses. When dealing with categorical variables, the generated dummy variables are

bundled together for purchase, even if some may not be needed. Suppose that we have G groups of variables and each group has size p_g variables, where we denote the g-th group cost as $\tilde{c}_g$ for $g = 1, \ldots, G$. This cost constraint is formulated as

$$\tilde{\mathcal{C}} := \{\boldsymbol{\beta} : \tilde{c}_g \tilde{I}(\beta_{g,1}, \cdots, \beta_{g,p_g}) \leq C\}, \tag{S4.30}$$

where $\tilde{I}(\beta_{g,1}, \cdots, \beta_{g,p_g})$ equals to one if one of $\beta_{g,j}$s is nonzero, and zero otherwise.

We apply Algorithm 2 in this scenario. We maintain the procedure for calculating $\boldsymbol{z}^k = \boldsymbol{\beta}^k + \boldsymbol{d}^k$ and the way of updating $\boldsymbol{\beta}^k_{\mathcal{I}^k}$ and $\boldsymbol{d}^k_{\mathcal{I}^k}$ based on current nonzero coefficients. The difference from Section 2.2 is to project into $\tilde{\mathcal{C}}$, instead of $\mathcal{C}$, after updating the signal candidate. We derive the following

$$\underset{\boldsymbol{\beta} \in \tilde{\mathcal{C}}}{\arg\min} \|\boldsymbol{\beta} - (\boldsymbol{\beta}^k + \boldsymbol{d}^k)\|_2^2 = \underset{\boldsymbol{\beta} \in \tilde{\mathcal{C}}}{\arg\min} \sum_{g=1}^{G} \|\boldsymbol{\beta}_g - (\boldsymbol{\beta}^k_g + \boldsymbol{d}^k_g)\|_2^2, \tag{S4.31}$$

where $\boldsymbol{\beta}_g$ and $\boldsymbol{d}_g$ are p_g-dimensional ($p_g \in \mathbb{N}$) subset of $\boldsymbol{\beta}$ and $\boldsymbol{d}$, respectively. Equation (S4.30) and the right-hand side of (S4.31) imply that if $\tilde{c}_g$ is spent and the gth group is chosen, the value lost is 0, but $\|\boldsymbol{\beta}^k_g + \boldsymbol{d}^k_g\|_2^2$ otherwise. By entering the loss and cost vector pair $(\{\|\boldsymbol{\beta}^k_g + \boldsymbol{d}^k_g\|_2^2\}, \{\tilde{c}_g\})$ for $g = 1, \ldots, G$, and the budget C, the knapsack algorithm provides the exact projection in $\tilde{C}$, that is, it solves the precise solution (S4.31). Incorporating this broadens

the practical effectiveness of Algorithm 2 when the cost at the group level should be considered.

S5 Numerical Details and Results for Simulations 2-6

For Simulation 2 (**S2**), we continue to show the performance of the algorithms, varying the correlation ρ among the covariates of $\mathbf{X}$ to better capture the applicability of the real world. The predictor is generated from a Gaussian distribution with mean $\mathbf{0}_p$ and covariance Σ where $\Sigma_{ij} = \rho^{|i-j|}$ for $i, j = 1, \ldots, p_0$. For $i, j = p_0+1, \ldots, p$, we set $\Sigma_{ij} = \mathbf{1}[i = j]$. The inclusion of the variable ρ complicates the computational process of identifying the optimal value for $\boldsymbol{\beta}^*$, so we reduce p_0 to 16. The elements of $\boldsymbol{\beta}^{true}$ are generated in the same way as in Simulation 1, and we set $C = 80$, and we vary ρ among $\{0.2, 0.4, 0.6, 0.8\}$.

Figure S1 shows the performance of the three algorithms over ρ. HCR and FCR increase its predictive performance as ρ increases. This implies the practical applicability of cost-constrained regression methods. Although the budget is limited to buy all variables, the algorithm can improve its explainability by choosing correlated cost-efficient variables. Therefore, for cost-constrained regression methods, increasing ρ has a similar effect of increasing the sample size. On the other hand, the second row of Figure S1

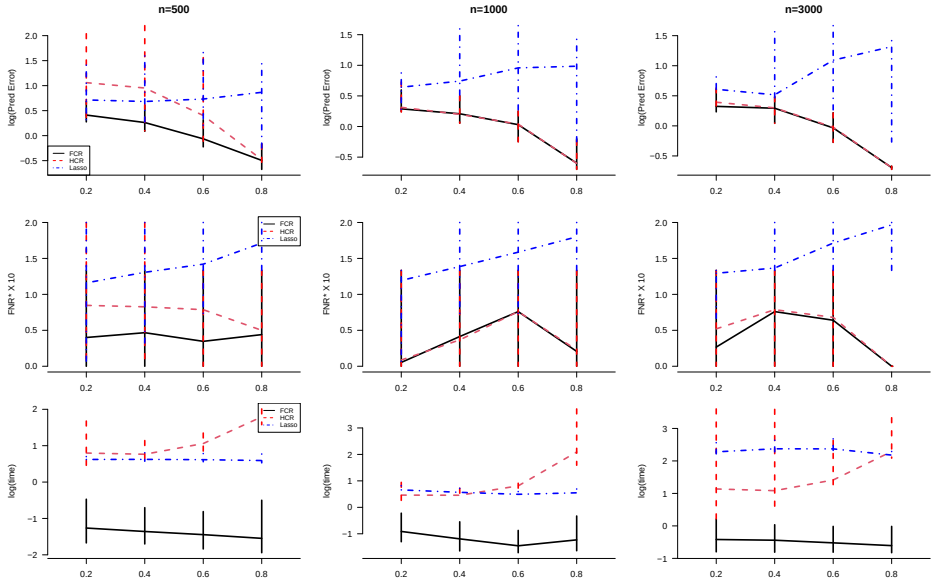

Figure S1: Simulation 2 results of the log of prediction error (the first row), and the FNR* (second), and the elapsed time (third) over $\rho = \{0.2, 0, 4.0.6, 0.8\}$.

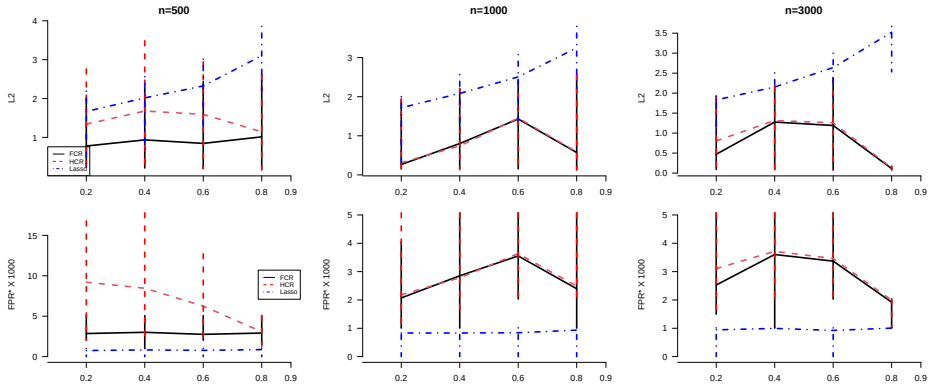

Figure S2: Simulation 2 results of L2 (the first row), and the FPR* (second) over $\rho = \{0.2, 0, 4.0.6, 0.8\}$.

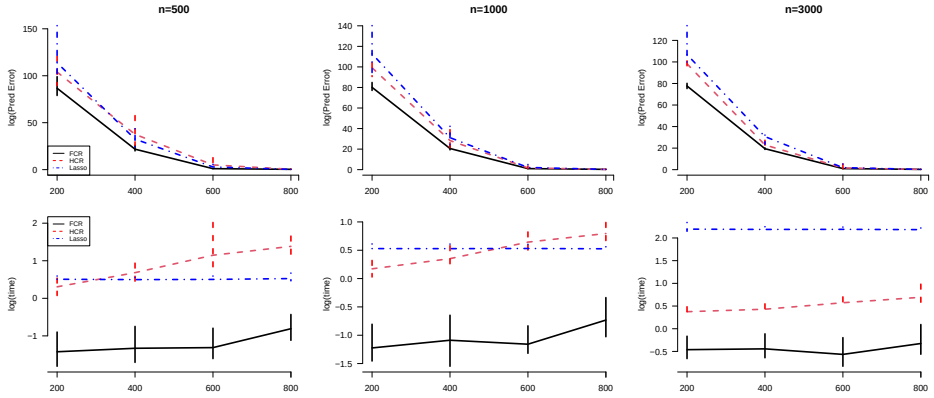

Figure S3: Simulation 3 results of the log of prediction error (the first row) and the elaspsed time (second) over $C = \{200, 400, 600, 800\}$

shows that LASSO tends not to select cost-effective variables as ρ increases. This implies the statistical inefficiency of using regularization methods in a cost-constrained situation. As similar to **S1** , FCR shows superior predictive performance and its small variance (cf. the first row of Figure 2), especially for high-dimensional regression settings. Furthermore, the superior performance of FCR is achieved in much shorter time than other methods (cf. the third row in Figure S1).

In Simulation 3 (**S3**), we extend the range of cost sets to investigate how these algorithms work. We sample costs uniformly from {1,2,. . . , 50} where the total cost is 686. We generate $\mathbf{X}$ the same way in **S2** where the correlation parameter ρ is equal to 0.3. As a result, FCR shows superior performance compared to the other methods, and the implementation takes

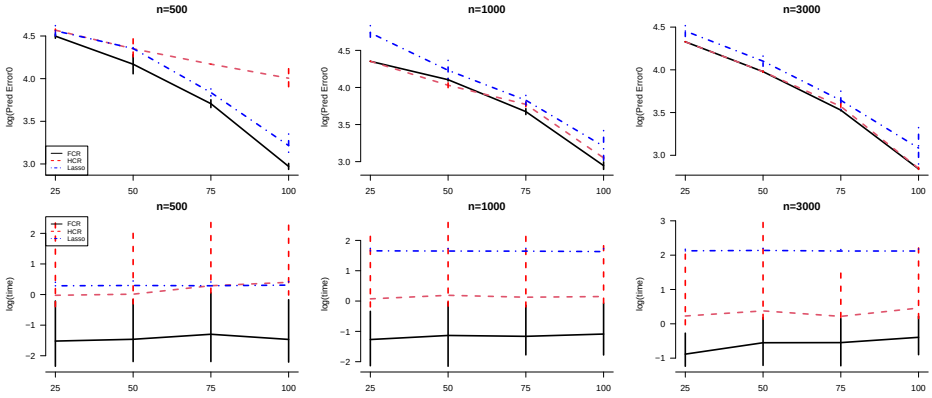

Figure S4: Simulation 4 results of the log of prediction error (the first row), and the elapsed time (third) over $C = \{25, 50, 75, 100\}$.

much shorter time than the others, just like the previous simulations.

In Simulation 4 (**S4**), the grouped costs are considered in this simulation. The predictor $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top$ is generated from a Gaussian distribution with mean $\mathbf{0}_p$ and covariance I_p. We consider where the total number of groups G is 250 and where each group has a size $p_g = 4$. Regarding the generation of the first $p_0 = 32$ elements of $\boldsymbol{\beta}^{true}$, where the 16 groups have nonzero elements. We generate the first two nonzero coefficients $N(2, 0.25)$ for those groups and zeros for the rest. We vary the budget by $C = \{25, 50, 75, 100\}$. For each $j \in \{1, 2, \cdots, p\}$, we choose c_j, the cost of collecting the jth variable, randomly from the set of integers $\{1, 2, \ldots, 10\}$. Figure $S4$ shows that FCR outperforms other methods in these cases. First, in grouped cost situations, FCR shows the best excep-

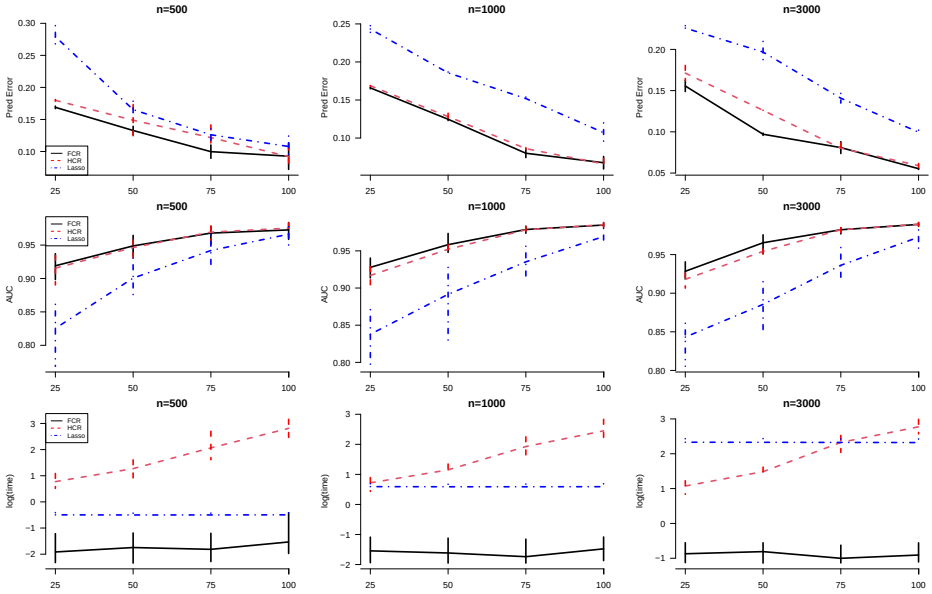

Figure S5: Simulation 5 results of the prediction error (the first row), and the elpased time (second) over $C = \{25, 50, 75, 100\}$.

tional performance in various budget scenarios, especially when n is 500. As the sample size grows, the performance of HCR tends to converge towards that of FCR. However, it should be noted that FCR achieves good performance at a significantly faster rate.

For Simulations 5 and 6 (**S5 and S6**), we extend our scope to logistic regression. The coefficient $\boldsymbol{\beta}^{true}$, and the cost sets are generated in the same way as in **S1**–**S3**. The predictor $\mathbf{X}$ is generated in the same way as in **S1**, but with $\rho = 0.5$ for **S5** and $\rho = \{0.2, 0.4, 0.6, 0.8\}$ for **S6**. For **S5**, we vary the budget C among $\{25, 50, 75, 100\}$ and investigate the performance

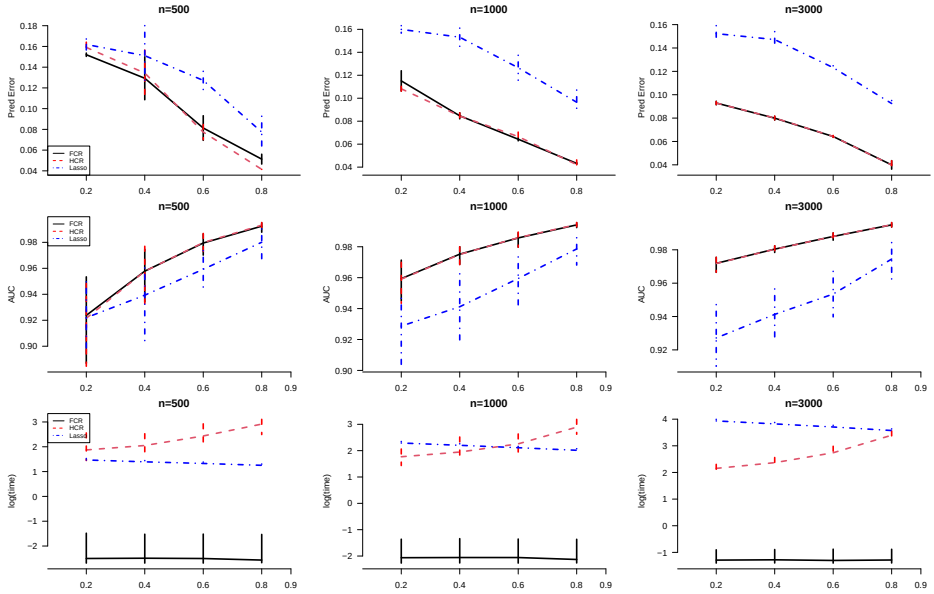

Figure S6: Simulation 6 results of the prediction error (the first row), and the elpased time (second) over $\rho = \{0.2, 0.4, 0.6, 0.8\}$.

of each method. Figures S5 and S6 summarize the results of **S5** and **S6**, respectively. The predictive performance of FCR exceeds that of other methods, and FCR also achieves its solution much more rapidly than HCR in these cases.

S6 More detailed results on NHANES data analysis

We report the loss function trajectories of the Hypertension, Arthritis, and Heart data in Figure S7.

The cost-constrained algorithm adeptly selects variables based on the

response \ costs	2	4	5	9
Diabetes	0.15	0.33	0.39	0.12
Hypertension	0.19	0.38	0.24	0.19
Arthritis	0.20	0.56	0.20	0.04
Heart	0.42	0.55	0	0.03

Table S1: Proportion of each cost for each dataset

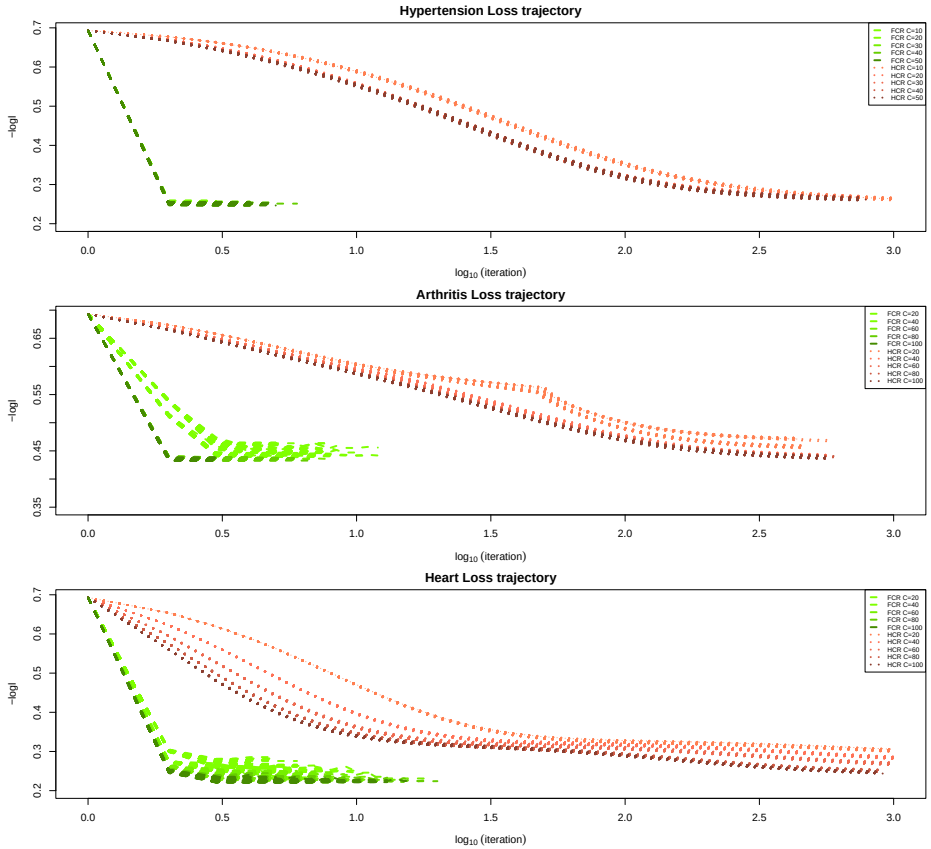

Figure S7: Loss function trajectories

cost limitations provided. We represent variable selections of FCR in the diabetes study in Figure S8, which illustrates its adaptability to different situations. We partition the variables of the data into demographic (Demo), questionnaire (Ques), examination (Exam), and laboratory (Lab) categories, numbering them according to the variable's nature. Additionally, we enclose its cost within parentheses, and indicate the variable's categorical nature with 'd'. If the variable is selected, we color it red and otherwise yellow. At the budget level of $C = 10$, the algorithm predominantly uses relatively inexpensive demographic and questionnaire variables, many of which are grouped variables. With a budget increase of 15–20, the algorithm acquires relatively costly examination variables (Exams 5 and 11). When reaching a budget of 25, it invests in the most expensive laboratory data (Lab 2) while excluding the examination variable (Exam 5). To address this, at a budget of 30, the model opts for a questionnaire instead of Exam 5 but reverts to acquiring Exam 5 when the budget reaches 35. Consequently, the algorithm dynamically adjusts its expenditure to acquire variables based on the available budget.

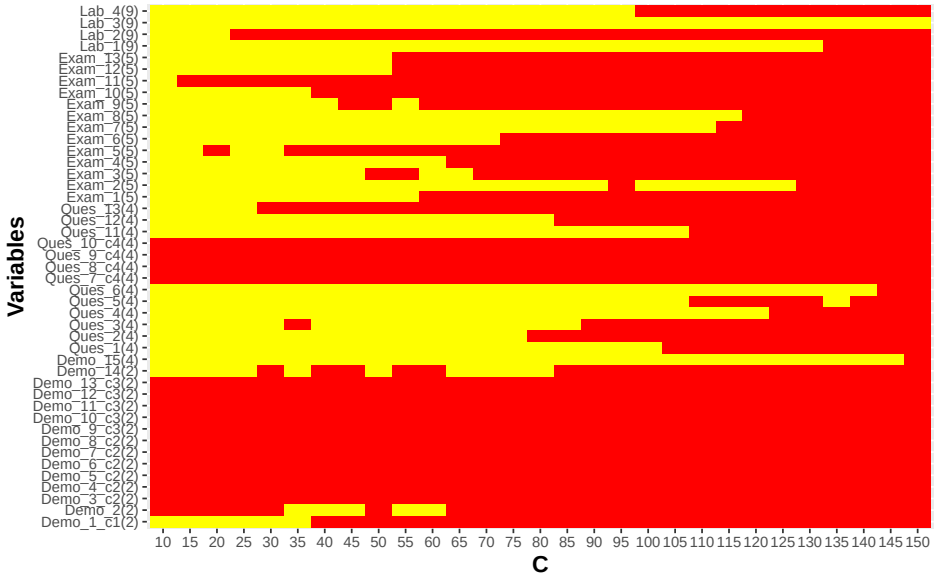

Figure S8: The variable selections of FCR in the diabetes study: the x-axis is budget C and the y-axis indicate the type of variables its cost in parenthesis. The categorical variables are indicated by adding c1–c4.

Bibliography

Foucart, S. (2011). Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM Journal on numerical analysis*, 49(6):2543–2563.

Huang, J., Jiao, Y., Liu, Y., and Lu, X. (2018). A constructive approach to l0 penalized regression. *The Journal of Machine Learning Research*, 19(1):403–439.

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25:1129–1141.

Yu, G., Fu, H., and Liu, Y. (2022). High-dimensional cost-constrained regression via nonconvex optimization. *Technometrics*, 64(1):52–64.