Statistica Sinica Preprint No: SS-2024-0331	
Title	Fast Cost-Constrained High Dimensional Regression
Manuscript ID	SS-2024-0331
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0331
Complete List of Authors	Hyeong Jin Hyun and Xiao Wang
Corresponding Authors	Xiao Wang
E-mails	wangxiao@purdue.edu
Notice: Accepted author version.	

Statistica Sinica

Fast Cost-Constrained High Dimensional Regression

Hyeong Jin Hyun and Xiao Wang

Purdue University

Abstract: The conventional statistical models assume the availability of covariates without associated costs, yet real-world scenarios often involve acquisition costs and budget constraints imposed on these variables. Scientists must navigate a trade-off between model accuracy and expenditure within these constraints. In this paper, we introduce fast cost-constrained regression (FCR), designed to tackle such problems with computational and statistical efficiency. Specifically, we develop fast and efficient algorithms to solve cost-constrained problems with the loss function satisfying a quadratic majorization condition. We theoretically establish nonasymptotic error bounds for the algorithm's solution, considering both estimation and selection accuracy. We apply FCR to extensive numerical simulations and four datasets from the National Health and Nutrition Examination Survey. Our method outperforms the latest approaches in various performance measures, while requiring fewer iterations and a shorter runtime. The FCR package is available at `https://github.com/anonymous1116/FCR`.

Key words and phrases: budget, cost constraint, high dimensional regression, nonasymptotic error bound, nonconvex optimization

1. Introduction

Statistical models are conventionally developed under the assumption that all covariates are freely observable. However, in many real-world scenarios, there are certain costs for acquiring each feature, as well as budgets limiting the total expenditure. Cost refers not only to the monetary loss, but also to any *measured* obstacle that prevents the user of the data from acquiring the features. It includes financial price, privacy impacts, fairness, or subjective disfavor. Taking these into account, data analysts must find a compromise between model accuracy and budget constraints.

Such problems have recently been emphasized in relevant health fields due to resource constraints specific to clinical environments (Kachuee et al., 2019; Nguyen et al., 2021; Erion et al., 2022). To predict whether a patient requires hospitalization or to estimate a certain health signal, performing all possible clinical tests is ideal. However, relatively inexpensive tests alone may provide enough information to inform the decision maker, while allowing a better allocation of money, time, and human resources. At the peak of the 2019 coronavirus disease (COVID-19) pandemic, many hospitals in the United States reached near capacity, putting into question the efficiency of resource allocation in the US healthcare system. Recent studies have found that patients are often diagnosed through a larger than necessary set of

Table 1: Top: NHANES data summary; Bottom: data cost allocation (Yu et al., 2022)

	p	# conti	# ctgr	total cost
diabetes	45	29	4	\$155
hypertension	26	18	3	\$101
arthritis	46	17	8	\$100
heart	158	2	38	\$131

	Examples	Cost
Demographics	Age; Income; Education Level	\$2 to \$4
Questionnaire	Average sleep length (in hours)	\$4
Examination	Diastolic blood pressure; Systolic blood pressure	\$5
Laboratory	Cholesterol; Triglyceride; Fibrinogen	\$9

testing services, which delays the care of patients with a more immediate medical need (Nguyen et al., 2021).

Our research is driven by the four datasets from the National Health and Nutrition Examination Survey (NHANES), preprocessed by Kachuee et al. (2019). These datasets are used to perform classification tasks for predicting whether a patient has diabetes, hypertension, arthritis, or heart disease. Each dataset contains a response variable indicating the presence of the disease and the associated covariates in each patient. Table 1 is an overview of the datasets, providing the total number of variables p, the number of continuous and categorical variables among them, and the total cost. The p explanatory variables are categorized into demographics, questionnaire, examination, and laboratory, with each variable in these categories priced at

\$2 — \$4, \$4, \$5, and \$9, respectively. Acquiring all variables in the diabetes dataset, for example, would cost \$155. When a patient's budget is very small, say \$10, the patient must choose whether to take a single laboratory test or a combination of examinations, questionnaires, and demographics for the best predictive accuracy within their constraints. One can easily imagine how fast the complexity of the problem would grow when a larger budget is considered, such as \$50. This example nicely illustrates our central question: How can we establish a reasonably effective model for prediction, while reflecting the limitations imposed by restricted resources?

1.1 Problem Statements

Consider the response of interest $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, a set of features $\mathbf{X} \in \mathbb{R}^{n\times p}$, a cost vector $\boldsymbol{c} = (c_1, \ldots, c_p) \in \mathbb{R}^p$, and a budget $C \in \mathbb{R}$, where c_j is the cost needed to use the j-th feature for $j = 1, \ldots, p$. We denote the jth column of $\mathbf{X}$ by $X_j \in \mathbb{R}^n$ and the row by $\boldsymbol{x}_i \in \mathbb{R}^p$. Suppose that a set of predictors $\boldsymbol{x}_i$ is associated with y_i through a link function $\mu_i \equiv \boldsymbol{\beta}^\top \boldsymbol{x}_i$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and β_j's are the coefficients of the covariate X_j. We denote the collection of $\boldsymbol{\beta}$s that satisfy the cost constraint as $\mathcal{C} = \{\boldsymbol{\beta} : \sum_{j=1}^p c_j I(\beta_j) \leq C\}$ where $I(\beta_j) = 0$ if $\beta_j = 0$, $I(\beta_j) = 1$ otherwise. For the high-dimensional case with $n < p$, assume that $\boldsymbol{\beta}$ is sparse. The true value of $\boldsymbol{\beta}$ is denoted by $\boldsymbol{\beta}^{true}$ and $\beta_j^{true} \neq 0$ for $j = 1, \ldots, p_0$, while $\beta_j^{true} = 0$ otherwise, where the sparsity of $\boldsymbol{\beta}$ is p_0 with $p_0 \leq n$.

Let $\Phi(y, \mu)$ be the loss function used to fit the model. The cost-constrained problem can be written as a nonconvex minimization problem:

$$\min_{\boldsymbol{\beta}} \mathbb{E}\left[\Phi(y, \boldsymbol{x}^\top \boldsymbol{\beta})\right] \text{ subject to } \sum_{j=1}^{p} c_j I(\beta_j) \leq C. \tag{1.1}$$

Given a dataset $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, a natural way to approximate this problem is to solve its emprical version:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \Phi(y_i, \boldsymbol{x}_i^\top \boldsymbol{\beta}) \text{ subject to } \sum_{j=1}^{p} c_j I(\beta_j) \leq C. \tag{1.2}$$

1.2 Literature review

Natarajan (1995) studied the best subset selection problem when the ℓ_0 -norm of $\boldsymbol{\beta}$ is constrained. This can be considered a simplification of optimization (1.2) where all c_j's are considered equal. Even in this case, they show that the problem is NP-hard. Tibshirani (1996) and Zhao and Yu (2006) proved that by relaxing the constraint to the ℓ_1-norm, the consistency of model selection is guaranteed under reasonable conditions. This relaxation allows the incorporation of a convex optimization solver that makes the problem of Natarajan (1995) less demanding. However, in the case of optimization (1.2), a similar relaxation is not natural due to the complex relationship between the cost $\boldsymbol{c}$ and the budget C. In addition, the nonlinear best subset selection using neural networks has been studied in Chen et al. (2020). The shape constrained regression has been studied extensively in Wang and Woodroofe (2007); Wang (2008); Wang and Shen (2010, 2013).

Alternative methods for solving cost-constrained problems in recent literature tend to prioritize accuracy over interpretability or vice versa. Peter et al. (2017) improved prediction accuracy using deep gradient boosting, and Kachuee et al. (2019) used reinforcement learning. Nguyen et al. (2021) used approaches based on well-known models, such as linear regression and random forest, to make a greedy selection of covariates. These methods ensure the interpretability of the estimates, but they also use heuristics that could call into question their consistency and accuracy when the size of the feature is large.

Another line of approach to solve (1.2) is to use a penalized regression technique (Tibshirani, 1996; Fan and Li, 2001; Zou and Hastie, 2005) and greedily compare the evaluations of the loss function of $\boldsymbol{\beta}$ that satisfy the cost constraint. These methods have two shortcomings. First, their empirical solutions occasionally narrow their search to only the set of covariates with nonzero β_js. However, the true solutions may include "insignificant" variables which are highly correlated to the significant ones and are cheaper but sufficient to explain the response. Second, even among the true p_0 signals, a brute-force comparison of the 2^{p_0} combinations can be undesirable, especially when p_0 is large.

Recently, Yu et al. (2022) introduced an algorithm designed for high-dimensional cost-constrained regression (HCR). It uses the first-order gradient descent approach to solve the optimization problem (1.2), which occasionally leads to the successful selection of non-zero entries. While the method ensures stable convergence, it is often hampered by slow convergence and diminished performance in high-dimensional regression tasks.

1.3 Our contributions

As such, we propose the Fast Cost-Constrained Regression (FCR) algorithm. This scheme is systemically tailored to minimize (1.2) for a cost-constrained problem. FCR consistently tends to outperform existing methods in identifying nonzero entries and delivering accurate regression predictions, particularly in high-dimensional settings. It achieves these results with significantly fewer iterations and reduced computation time, as demonstrated through extensive experiments using both synthetic data and the four NHANES datasets. Additionally, we establish a non-asymptotic error bound for the solutions produced by the algorithm, quantifying both estimation and selection accuracy. Our theoretical analysis shows that, under certain conditions, the FCR algorithm guarantees an exact solution when all true signal satisfies the budget constraint, offering a level of assurance that other methods do not provide.

The paper is organized as follows. We introduce the algorithm of the FCR method in a detailed manner and elaborate on how the desired characteristics are achieved in Section 2. We investigate the theoretical property of the algorithm by establishing the ℓ_2 error bound of the algorithm in Section 3. In Section 4 we numerically show that the FCR method not only delivers accurate predictions, but also achieves this precision in notably few iterations. In Section 5 we show that our approach shows superior performance when applied to real-world data from the NHANES, indicating the practicability of our methods. The FCR package is available at `https://github.com//anonymous1116/FCR`.

2. Fast Cost-constrained Regression

We initially present the algorithm for linear regression with cost constraints. Building on the principles established in linear regression, we generalize the approach to encompass broader scenarios, including logistic regression. We explore the extension to incorporate group-wise costs.

2.1 Fast cost-constrained linear regression

Consider when the loss function $\Phi(y, \boldsymbol{x}^\top \boldsymbol{\beta})$ is an ℓ_2-loss that leads (1.1) to

$$\boldsymbol{\beta}^\dagger = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p} c_j I(\beta_j) \leq C, \tag{2.3}$$

where $\boldsymbol{\beta}^\dagger$ is a global minimizer. The NP hardness of the optimization problem (2.3) adds complexity to our discussion. It should be noted that $\boldsymbol{\beta}^\dagger$ also serves as a coordinate-wise minimizer. This observation enables us to address cost constraints in a more nuanced way, viewing the impact of each coordinate as a determining factor in deciding whether the use of cost c_j is justified. We delve into the properties of the coordinate-wise minimizer of the optimization problem in Theorem 1, where the proof is given in Supplement S1.

Theorem 1. *Let $\boldsymbol{\beta}^\dagger$ be a coordinate-wise minimizer of (2.3) and $\mathcal{I}^\dagger$ be a* *set of nonzero indices of $\boldsymbol{\beta}^\dagger$. Then each coordinate $\beta_j^\dagger$ of $\boldsymbol{\beta}^\dagger$ satisfies*

$$\boldsymbol{\beta}^\dagger = T(\boldsymbol{\beta}^\dagger + \boldsymbol{d}), \tag{2.4}$$

where $T : \mathbb{R}^p \to \mathbb{R}^p$ is a thresholding operator defined by its coordinate

$$T_j(\boldsymbol{\beta}) = \begin{cases} 0 & \textit{if } j \notin \mathcal{I}^\dagger \\ \beta_j & \textit{if } j \in \mathcal{I}^\dagger \end{cases}, \tag{2.5}$$

where $\boldsymbol{d}$ is defined by its coordinate $d_j = X_j^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^\dagger)/\|X_j\|_2^2$, $j = 1, \cdots, p$. Moreover, T satisfies, for all $\boldsymbol{\beta} \in \mathcal{C}$,

$$\left\| T(\boldsymbol{\beta}^\dagger + \boldsymbol{d}) - (\boldsymbol{\beta}^\dagger + \boldsymbol{d}) \right\|_2 \leq \left\| \boldsymbol{\beta} - (\boldsymbol{\beta}^\dagger + \boldsymbol{d}) \right\|_2. \tag{2.6}$$

From (2.4) in Theorem 1, we note that if j is in the index set $\mathcal{I}^\dagger$, the value of d_j should be zero by $\beta_j = \beta_j + d_j$. In other words, once the nonzero coefficients are determined, the subvector $\boldsymbol{d}_{\mathcal{I}^\dagger}$ is zero, where $\boldsymbol{d}_{\mathcal{I}^\dagger}$ is a subvector of $\boldsymbol{d} := (d_1, \ldots, d_p)$ indexed by $\mathcal{I}^\dagger$. This implies that the column vector X_j of the design matrix and $\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^\dagger$ is orthogonal, where j

is in the corresponding set. This gives us

$$\mathbf{0} = \mathbf{X}_{\mathcal{I}^\dagger}^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^\dagger) = \mathbf{X}_{\mathcal{I}^\dagger}^\top(\boldsymbol{y} - \mathbf{X}_{\mathcal{I}^\dagger}\boldsymbol{\beta}_{\mathcal{I}^\dagger}^\dagger),$$

where $\mathbf{X}_{\mathcal{I}^\dagger} \in \mathbb{R}^{n\times|\mathcal{I}|}$ and $\boldsymbol{\beta}_{\mathcal{I}^\dagger}^\dagger \in \mathbb{R}^{|\mathcal{I}|}$ are the sub design matrix of $\mathbf{X}$ and vectors of $\boldsymbol{\beta}^\dagger$ indexed by $\mathcal{I}^\dagger$. This leads to the fact that $\boldsymbol{\beta}_{\mathcal{I}^\dagger}^\dagger$ is actually the least squares estimate with its design matrix $\mathbf{X}_{\mathcal{I}^\dagger}$ given by

$$\boldsymbol{\beta}_{\mathcal{I}^\dagger}^\dagger = (\mathbf{X}_{\mathcal{I}^\dagger}^\top\mathbf{X}_{\mathcal{I}^\dagger})^{-1}\mathbf{X}_{\mathcal{I}^\dagger}^\top\boldsymbol{y}, \quad \boldsymbol{d}_{\mathcal{I}^\dagger} = \mathbf{0}. \tag{2.7}$$

We further notice the role of the thresholding operator T by inequality (2.6) in Theorem 1. This suggests that the operator T projects its input onto the cost constraint space $\mathcal{C}$, minimizing the ℓ_2-distance with the input $\boldsymbol{\beta}^\dagger + \boldsymbol{d}$ to produce the output $T(\boldsymbol{\beta}^\dagger + \boldsymbol{d})$. We introduce the 0-1 knapsack algorithm to implement the operator T. Formally, given a p-dimensional vector $a \in \mathbb{R}^p$, this knapsack algorithm provides an ℓ_2-projection $\widehat{a} \in \mathcal{C}$ that satisfies

$$\|a - \widehat{a}\|_2^2 \leq \|a - b\|_2^2 \quad \forall\, b \in \mathcal{C}. \tag{2.8}$$

The knapsack algorithm is originally designed to identify the optimal selection of items from a given set, each assigned a cost and loss. It minimizes the overall loss while the cumulative cost does not exceed a specified limit. The right hand side of (2.8) implies that if c_j is spent and jth index is chosen, the value lost is none, but a_j^2 otherwise. By inputting the loss and cost pair $(\{a_j^2\}, \{c_j\})$ for $j = 1, \ldots, p$, and the budget C, the algorithm provides projection (2.8) that satisfy $\arg\min_{b\in\mathcal{C}} \|a - b\|_2^2$.

A commonly used approach to solve the 0-1 knapsack problem is dynamic programming, with a pseudo-polynomial time complexity of $O(pC)$ (Martello and Toth, 1990). Recent research continues to explore ways to reduce computational costs, such as achieving complexities of $O(p + c_{max}^2 \log^4 c_{max})$ where $c_{max} = \max_j c_j$ (see Jin (2024) and the references therein). We primarily use the dynamic programming-based algorithm from the R package `adagio` (Borchers, 2022) in our simulation studies.

We create the FCR algorithm, outlined in Algorithm 1, to embody the properties of a coordinate-wise minimizer throughout its iterations. First, starting from the initial value $\boldsymbol{\beta}^0$, typically $\boldsymbol{\beta}^0 = \mathbf{0}$, we update the coordinate $\boldsymbol{\beta}'s$ to be $\boldsymbol{z}^k := \boldsymbol{\beta}^k + \boldsymbol{d}^k$ in the kth iteration. Since $\boldsymbol{z}^k$ is not in $\mathcal{C}$, we project it into the space of $\mathcal{C}$ and denote it by $\boldsymbol{u}^k := \Pi_{\mathcal{C}}(\boldsymbol{z}^k)$ where the knapsack algorithm is performed. We now detect the nonzero components of $\boldsymbol{u}^k$ and denote its index set by $\mathcal{I}^k$. Detecting the nonzero set $\mathcal{I}^k$ allows

Algorithm 1 FCR

```
Input: Explanatory variable $\mathbf{X}$, response variable $y$, costs vector
$(c_1, \cdots, c_p)$, budget $C$, maxiter $K$, tol $\delta$,
$\boldsymbol{\beta}^0 = \mathbf{0}, \boldsymbol{d}^0 = 1/diag(\mathbf{X}^\top \mathbf{X}) \cdot \mathbf{X}^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^0)$
Output: $\hat{\boldsymbol{\beta}}$
  for $k = 0, \cdots, K$ do
      Calculate $\boldsymbol{z}^k = \boldsymbol{\beta}^k + \boldsymbol{d}^k$
      Project to $\mathcal{C}$: $\boldsymbol{u}^k = \Pi_{\mathcal{C}}(\boldsymbol{z}^k)$
      $\mathcal{I}^k = \{j : u_j^k \neq 0\}$
      $\boldsymbol{\beta}^{k+1}_{\mathcal{I}^k} = (\mathbf{X}^\top_{\mathcal{I}^k}\mathbf{X}_{\mathcal{I}^k})^{-1}\mathbf{X}^\top_{\mathcal{I}^k}\boldsymbol{y}$ ; $\boldsymbol{\beta}^{k+1}_{(\mathcal{I}^k)^c} = 0$
      $\boldsymbol{d}^{k+1}_{\mathcal{I}^k} = 0$; $\boldsymbol{d}^{k+1}_{(\mathcal{I}^k)^c} = 1/diag(\mathbf{X}^\top_{(\mathcal{I}^k)^c}\mathbf{X}_{(\mathcal{I}^k)^c}) \cdot \mathbf{X}^\top_{(\mathcal{I}^k)^c}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^k)$
      if Stopping criteria is met then
          $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^k$; break
      end if
  end for
```

us to update $\boldsymbol{\beta}$ using least squares estimates, as we observe in Theorem 1 and Equation (2.7). Therefore, we update

$$\boldsymbol{\beta}^{k+1}_{\mathcal{I}^k} = (\mathbf{X}^\top_{\mathcal{I}^k}\mathbf{X}_{\mathcal{I}^k})^{-1}\mathbf{X}^\top_{\mathcal{I}^k}\boldsymbol{y}, \qquad \boldsymbol{\beta}^{k+1}_{(\mathcal{I}^k)^c} = \mathbf{0},$$

$$\boldsymbol{d}^{k+1}_{\mathcal{I}^k} = \mathbf{0}, \qquad \boldsymbol{d}^{k+1}_{(\mathcal{I}^k)^c} = \frac{1}{\|X_j\|_2^2} X_j^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^{k+1}).$$

The FCR algorithm is in fact oriented to generalize non-convex optimization methods with the constraints $\sum_{j=1}^p I(\beta_j) \leq s$. Foucart (2011) and Huang et al. (2018) detected the support of the solution based on the sum of the primal ($\boldsymbol{\beta}^k$) and dual ($\boldsymbol{d}^k$) approximations using the hard thresholding operator, and the least squares solution is calculated on the detected support. These methods are well known to show good empirical performance and low complexity with regard to the best subset selection problem. For the best subset selection problem where all c_js are the same, projection $\Pi_{\mathcal{C}}$ is equivalent to choosing the largest β_js in order. The knapsack algorithm replaces this procedure for general cost constraints.

The primary difference between HCR (Yu et al., 2022) and FCR lies in the approach inspired by Theorem 1, which ensures that once the active set $\mathcal{I}^k$ is identified, the solution minimizes the loss function within that set. This process leads to dynamic updates of $\boldsymbol{\beta}^{k+1}$ and results in different active sets $\mathcal{I}^k$ as the iterations progress. It has a significant impact on the speed of each algorithm, as we demonstrate in Sections 4 and 5.

2.2 Extensions with QM* condition

Our algorithm is not limited to the case where the loss function (1.1) is an ℓ_2-loss. In fact, based on the case of ℓ_2-loss, we generalize our algorithm to the scope of quadratic majorization* (QM*). Let the empirical loss be

$$L(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} \Phi(y_i, \boldsymbol{\beta}^\top \boldsymbol{x}_i). \tag{2.9}$$

Define the QM* condition as follows.

Definition 1. The loss function L is said to satisfy the QM* condition, if and only if the following two assumptions hold:

(i) $L(\boldsymbol{\beta})$ is differentiable as a function of $\boldsymbol{\beta}$, i.e. $\nabla L(\boldsymbol{\beta})$ exists everywhere.

(ii) There exists a semipositive definite $\mathbf{H}$ such that for all $\boldsymbol{\beta}, \boldsymbol{\eta} \in \mathbb{R}^p$,

$$L(\boldsymbol{\beta}) \leq L(\boldsymbol{\eta}) + (\boldsymbol{\beta} - \boldsymbol{\eta})^\top \nabla L(\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\eta})^\top \mathbf{X}^\top \mathbf{H} \mathbf{X} (\boldsymbol{\beta} - \boldsymbol{\eta}). \tag{2.10}$$

The QM* condition indeed holds for many popular loss functions used in regression and classification. It includes regression loss functions for the generalized linear model (GLM) such as the linear and logistic models. In addition, it incorporates the loss functions for support vector machines, such as the squared hinge and Huberized hinge loss functions. We provide conditions in the Supplement S2 that are sufficient to characterize a class of loss functions that satisfies the QM* conditions.

The QM* condition is a modification of the QM condition (Yang and Zou, 2015), which offers specialization without imposing significant constraints on the eligible classes of loss functions. The quadratic term $\mathbf{X}^\top \mathbf{H} \mathbf{X}$ found in (2.10) is extended to a semipositive definite matrix $\tilde{\mathbf{H}}$ within the QM condition. However, the selection of $\tilde{\mathbf{H}}$ typically corresponds to an upper bound of the Hessian of Φ with respect to $\boldsymbol{\beta}$. In such cases, $\tilde{\mathbf{H}}$ is expected to take the form of $\mathbf{X}^\top\mathbf{H}\mathbf{X}$ for some $\mathbf{H}$, indicating a strong connection between the two. It should be noted that all four loss functions discussed in Yang and Zou (2015) satisfy the QM* condition.

Let $\Phi'_\mu(y, \mu)$ be the partial derivative of Φ with respect to μ. We have $\nabla L(\boldsymbol{\eta}) = \mathbf{X}^\top \boldsymbol{g}(\boldsymbol{\eta})$, where $g_i = \Phi_\mu(y_i, \boldsymbol{x}_i^\top\boldsymbol{\eta})/n$ for $i = 1, \ldots, n$. If the loss function satisfies the QM* condition, we can find $M \in \mathbb{R}$ that bounds $\mathbf{H}$ such that $\mathbf{H} \leq M \cdot I_n$. Define $Q(\boldsymbol{\beta}, \boldsymbol{\eta})$ as

$$Q(\boldsymbol{\beta}, \boldsymbol{\eta}) := L(\boldsymbol{\eta}) + (\boldsymbol{\beta} - \boldsymbol{\eta})^\top\mathbf{X}^\top\boldsymbol{g}(\boldsymbol{\eta}) + \frac{M}{2}(\boldsymbol{\beta} - \boldsymbol{\eta})^\top\mathbf{X}^\top\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\eta}).$$

$Q(\boldsymbol{\beta}, \boldsymbol{\eta})$ is also the upper bound for $L(\boldsymbol{\beta})$ for any $\boldsymbol{\beta}, \boldsymbol{\eta} \in \mathbb{R}^p$ and

$$Q(\boldsymbol{\beta}, \boldsymbol{\eta}) = \frac{M}{2}\|(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\eta}) + \frac{1}{M}\boldsymbol{g}(\boldsymbol{\eta})\|_2^2 + h(\boldsymbol{\eta}) \tag{2.11}$$

where $h(\boldsymbol{\eta})$ is a function that only depends on $\boldsymbol{\eta}$, not $\boldsymbol{\beta}$. We then have

$$\boldsymbol{\beta}^\dagger = \arg\min_{\boldsymbol{\beta}\in\mathcal{C}} L(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\mathcal{C}} Q(\boldsymbol{\beta}, \boldsymbol{\beta}^\dagger) = \arg\min_{\boldsymbol{\beta}\in\mathcal{C}} \|\mathbf{X}\boldsymbol{\beta} - \tilde{\boldsymbol{y}}\|_2^2, \tag{2.12}$$

where $\tilde{\boldsymbol{y}} = \mathbf{X}\boldsymbol{\beta}^\dagger - \boldsymbol{g}(\boldsymbol{\beta}^\dagger)/M$. The first and third equality are straightforward. The second equality follows from two key facts: $L(\boldsymbol{\beta}) \leq Q(\boldsymbol{\beta}, \boldsymbol{\beta}^\dagger)$ and

$L(\boldsymbol{\beta}^{\dagger}) = Q(\boldsymbol{\beta}^{\dagger}, \boldsymbol{\beta}^{\dagger})$. Specifically, $Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{\dagger})$ attains its lower bound $L(\boldsymbol{\beta}^{\dagger})$ at $\boldsymbol{\beta}^{\dagger}$. The formulation (2.12) links the QM* loss function to the case with the ℓ_2-loss. It leads to establish Theorem 2 which generalizes Theorem 1.

Theorem 2. *Let $\boldsymbol{\beta}^{\dagger}$ be a coordinate-wise minimizer of (1.2), and let $L(\boldsymbol{\beta})$ satisfy the QM* condition. Then each coordinate $\beta_j^{\dagger}$ of $\boldsymbol{\beta}^{\dagger}$ satisfies*

$$\boldsymbol{\beta}^{\dagger} = T(\boldsymbol{\beta}^{\dagger} + \boldsymbol{d}), \tag{2.13}$$

where $T : \mathbb{R}^p \to \mathbb{R}^p$ is a thresholding operator defined in (2.5) and $\boldsymbol{d}$ is defined by its coordinate $d_j = -\tilde{X}_j^{\top} \boldsymbol{g}(\boldsymbol{\beta}^{\dagger}) / \|\tilde{X}_j\|_2^2$, $j = 1, \cdots, p$. Moreover, T satisfies, for all $\boldsymbol{\beta} \in \mathcal{C}$,

$$\|T(\boldsymbol{\beta}^{\dagger} + \boldsymbol{d}) - (\boldsymbol{\beta}^{\dagger} + \boldsymbol{d})\|_2 \leq \|\boldsymbol{\beta} - (\boldsymbol{\beta}^{\dagger} + \boldsymbol{d})\|_2. \tag{2.14}$$

Theorem 2 generalizes the FCR algorithm when the loss function satisfies the QM* condition. Note from (2.13) that $d_j = 0$ for $j \in \mathcal{I}^{\dagger}$, which implies

$$0 = \mathbf{X}_{\mathcal{I}^{\dagger}}^{\top} \boldsymbol{g}(\boldsymbol{\beta}^{\dagger}) = \mathbf{X}_{\mathcal{I}^{\dagger}}^{\top} \left[\frac{1}{n} \sum_{i=1}^{n} \Phi(y_i, (\boldsymbol{\beta}^{\dagger})^{\top} \boldsymbol{x}_i) \right] = \mathbf{X}_{\mathcal{I}^{\dagger}}^{\top} \left[\frac{1}{n} \sum_{i=1}^{n} \Phi(y_i, (\boldsymbol{\beta}^{\dagger})_{\mathcal{I}^{\dagger}}^{\top} (\boldsymbol{x}_i)_{\mathcal{I}^{\dagger}}) \right],$$

which is the gradient of L when the design matrix is $\mathbf{X}_{I^{\dagger}}$ in lieu of $\mathbf{X}$. To reflect this observation, we incorporate an update $\boldsymbol{\beta}_{\mathcal{I}^k}^{k+1} = \arg\min L(\boldsymbol{\beta}|_{\mathcal{I}^k})$ where $(\boldsymbol{\beta}|_{\mathcal{I}})_j = \beta_j \mathbf{1}(j \in \mathcal{I})$. Since this minimization only concerns the submatrix $\mathbf{X}_{\mathcal{I}^k}$, seen as a lower-dimensional regression, a computationally stable algorithm can be used, such as the iterative reweighted least squares algorithm for logistic regression. The rest of the algorithm is similar to Algorithm 1. We update the coordinate $\boldsymbol{\beta}'s$ to $\boldsymbol{z}^k = \boldsymbol{\beta}^k + \boldsymbol{d}^k$ in the kth iteration and then project it into the space of $\mathcal{C}$ based on the observation of (2.14) using the knapsack algorithm. Since we now detect the nonzero components of $\boldsymbol{u}^k$, $\boldsymbol{\beta}$ is updated as above, $\boldsymbol{d}_{\mathcal{I}^k}^{k+1} = 0$, and $\boldsymbol{d}_{(\mathcal{I}^k)^c}^{k+1} = -1/diag(\mathbf{X}_{(\mathcal{I}^k)^c}^\top \mathbf{X}_{(\mathcal{I}^k)^c}) \cdot \mathbf{X}_{(\mathcal{I}^k)^c}^\top g(\boldsymbol{\beta}^k)$. We summarize the generalized FCR method (GFCR) in Algorithm 2, and discuss its stopping criteria in Supplement S3.

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. We discuss such issues in the Supplement Section S4.

Algorithm 2 Generalized FCR

```
Input: Explanatory variable $\mathbf{X}$, response variable $y$, costs vector
$(c_1, \cdots, c_p)$, budget $C$, maxiter $K$, tol $\delta$,
$\boldsymbol{\beta}^0 = \mathbf{0}, \boldsymbol{d}^0 = -1/diag(\mathbf{X}^\top \mathbf{X}) \cdot \mathbf{X}^\top \boldsymbol{g}(\boldsymbol{\beta}^0)$
Output: $\hat{\boldsymbol{\beta}}$
  for $k = 0, \cdots, K$ do
      Calculate $\boldsymbol{z}^k = \boldsymbol{\beta}^k + \boldsymbol{d}^k$
      Project to $\mathcal{C}$: $\boldsymbol{u}^k = \Pi_{\mathcal{C}}(\boldsymbol{z}^k)$
      $\mathcal{I}^k = \{j : u_j^k \neq 0\}$
      $\boldsymbol{\beta}_{\mathcal{I}^k}^{k+1} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}|_{\mathcal{I}^k})$; $\boldsymbol{\beta}_{(\mathcal{I}^k)^c}^{k+1} = \mathbf{0}$
      $\boldsymbol{d}_{\mathcal{I}^k}^{k+1} = \mathbf{0}$; $\boldsymbol{d}_{(\mathcal{I}^k)^c}^{k+1} = -1/diag(\mathbf{X}_{(\mathcal{I}^k)^c}^\top \mathbf{X}_{(\mathcal{I}^k)^c}) \cdot \mathbf{X}_{(\mathcal{I}^k)^c}^\top g(\boldsymbol{\beta}^k)$
      if Stopping criteria is met then
          $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^k$; break
      end if
  end for
```

3. Theoretical Analysis

In this section, we establish nonasymptotic error bounds for the solution sequence generated by Algorithm 1. We investigate with the solution $\boldsymbol{\beta}^*$ of (1.1) where we write $\mathcal{I}^* = \{j : \beta_j^* \neq 0\}$. We focus on high-dimensional linear regression $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^{true} + \boldsymbol{\zeta}$, where $\boldsymbol{\beta}^{true}$ is the true unknown coefficient, $\mathbf{X}$ is assumed to be centered and $\sqrt{n}$-normalized, and $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)^T$ is the mean zero random error vector. We assume that $\zeta_1, \ldots, \zeta_n$ are i.i.d with a sub-Gaussian distribution that satisfies $\mathbb{E}[\exp(t\zeta_i)] \leq \exp(\sigma^2 t^2/2)$ for all $t \in \mathbb{R}$ and $i = 1, \ldots, n$. Initially, we adopt various definitions from Zhang and Huang (2008) and Huang et al. (2018) that guide the examination of FCR's properties.

Definition 2. $\mathbf{X}$ satisfies the sparse Riesze condition (SRC) with order s and spectrum bound $\{b_-(s), b_+(s)\}$ if

$$0 < b_-(s) \leq \frac{\|\mathbf{X}_{\mathcal{I}}\boldsymbol{\gamma}\|_2^2}{n\|\boldsymbol{\gamma}\|_2^2} \leq b_+(s) < \infty \quad \forall\, 0 \neq \boldsymbol{\gamma} \in \mathbb{R}^{|\mathcal{I}|},$$

for $\mathcal{I} \subset \{1, 2, \ldots, p\}$ and $|\mathcal{I}| \leq s$, and we denote this condition by $\mathbf{X} \sim \mathrm{SRC}\{s, b_-(s), b_+(s)\}$.

Definition 3. The sparse orthogonality constant θ_{s_1,s_2} is defined by the smallest value such that

$$\frac{\|\mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{J}}\boldsymbol{\gamma}\|_2}{n\|\boldsymbol{\gamma}\|_2} \leq \theta_{s_1,s_2},\ \forall\, 0 \neq \boldsymbol{\gamma} \in \mathbb{R}^{|\mathcal{J}|},$$

where $\mathcal{I}, \mathcal{J} \subset \{1, \ldots, p\}, |\mathcal{I}| \leq s_1, |\mathcal{J}| \leq s_2,$ and $\mathcal{I} \cap \mathcal{J} = \emptyset$.

The SRC condition outlines the range of eigenvalues for the diagonal submatrices of the Gram matrix $\mathbf{X}^\top\mathbf{X}/n$, while the sparse orthogonality constant relates to the off-diagonal elements. By definitions, the ranges of the constants are $b_-(s) \leq 1 \leq b_+(s)$ and $0 \leq \theta_{s_1,s_2} \leq 1$. We also define a projection $\bar{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^{true}$ into the cost constraint space $\mathcal{C}$ under the ℓ_2 distance, that is, $\bar{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathcal{C}} \|\boldsymbol{\beta} - \boldsymbol{\beta}^{true}\|_2$, and the distance $R_{\mathcal{C}} := \|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^{true}\|_2$.

Now, we assume the structure of the realized design matrix $\mathbf{X}$. Let us define q as an upper bound for the size of $\mathcal{I}^k$ and $\mathcal{I}^*$, that is, $q := \max\{|\mathcal{I}^k| : k \geq 1\} \vee |\mathcal{I}^*|$. Define a constant that depends on the spectrum bound and the sparse orthogonality constant of the design matrix $\mathbf{X}$ such that

$$r := \frac{2\theta_{q,q} + (1+\sqrt{2})\theta_{q,q}^2}{b_-(q)^2} + \frac{(1+\sqrt{2})\theta_{q,q}}{b_-(q)}.$$

Theorem 3 establishes the property of the FCR algorithm within the assumption $r < 1$. The value r provides an approximate measure of the correlation between the variables, with a smaller r indicating a weaker correlation. Sufficient conditions are provided for a design matrix to satisfy the SRC (Zhang and Huang, 2008; Huang et al., 2018). This assumption is satisfied, for instance, if $b_+(2q) \leq 1.1599$ and $b_-(2q) \geq 0.8401$ (Huang et al., 2018), where the best subset selection is discussed within this assumption. We generalize the results in the article to the setting of cost-constrained constraints beyond the ℓ_0-constraint.

Theorem 3. *Suppose $r < 1$. Then, for any $\alpha \in (0, 1/2)$, we have*

$$\|\boldsymbol{\beta}^* |_{\mathcal{I}^* \setminus \mathcal{I}^{k+1}} \|_2 \leq r^{k+1} \|\boldsymbol{\beta}^*\|_2 + b_1 b_2 R_{\mathcal{C}} + b_1 \epsilon, \tag{3.15}$$

$$\|\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^*\|_2 \leq b_3 r^k \|\boldsymbol{\beta}^*\|_2 + b_2 b_4 R_{\mathcal{C}} + b_4 \epsilon, \tag{3.16}$$

with probability at least $1-2\alpha$ *where*

$$b_1 = \frac{r}{(1-r)\theta_{q,q}},\ b_2 = \sqrt{\frac{b_+(q)b_+(p_0)b_+(p_0+q)}{b_-(p_0+q)}},\ b_3 = 1+\frac{\theta_{q,q}}{b_-(q)},$$
$$b_4 = \left(b_1b_3 + \frac{1}{b_-(q)}\right),\quad \textit{and}\quad \epsilon = \sigma\sqrt{\frac{2q\log(p/\alpha)}{n}}.$$

Theorem 3 establishes the ℓ_2 error bound for the approximation errors of the sequence generated by the FCR in the $(k+1)$th iteration. In particular, (3.15) gives the ℓ_2-bound of the elements in $\mathcal{I}^*$ not included in the nonzero set in the $(k+1)$th iteration. Inequality (3.16) provides an upper bound for the ℓ_2 estimation error of $\boldsymbol{\beta}^{k+1}$. With the certain assumption that $r<1$, we show that these error bounds geometrically decay to the model error measured by $b_1b_2R_{\mathcal{C}}+b_1\epsilon$ and $b_2b_4R_{\mathcal{C}}+b_4\epsilon$, respectively, with high probability. We present its proof in the Supplement S1.

The outcome of Theorem 3 provides an interpretation of the FCR property, consistent with the observations in the simulations outlined in Section 4. First, the larger n increases the efficiency of the algorithm. The part of the error bound ϵ diminishes with larger data sizes n, which aligns with the numerical results as described in the first row of Figure 3. Furthermore, when operating within a constrained budget, the task of identifying optimal combinations of cost-effective variables becomes more challenging.

In practical terms, a tighter budget amplifies the difficulty of excluding insignificant variables. In extreme cases, where the budget is just above the cost of each variable with identical pricing, pinpointing the relevant variables becomes more challenging after eliminating the insignificant ones. These findings are consistent with the results summarized in the second row of Figure 2 and in the first row of Figure 3.

We remark that the underlying assumptions of our results limit the general applicability of the problem. The value of r critically hinges on the sparse orthogonality constants $\theta_{q,q}$ being small, which basically requires that any pairs of groups of covariates cannot be too correlated. As a result, this assumption may not fully capture all situations. However, our results provide valuable insights into the practical use of the FCR method. In many modern data settings, where steps are taken to reduce multicollinearity (e.g. through design or preprocessing), the variables are weakly correlated with each other. Also, the theory's applicability extends to cases where the significant ones are weakly correlated with many cheaper variables, rather than just a single correlated variable.

When the budget increases to meet condition $\sum_{j=1}^{p_0} c_j I(\beta_j) \leq C$, the algorithm transforms effectively into a solver for the problem of selecting the best subset. We demonstrate that the FCR algorithm can successfully identify nonzero coefficients in this scenario. To facilitate this, we introduce a notation: $\bar{m}$, the minimum true value of $\boldsymbol{\beta}^*$ i.e. $\bar{m} := \min\{\beta_j^* : j \in \mathcal{I}^*\}$.

Corollary 1. *Let* $\sum_{i=1}^{p_0} c_j I(\beta_j^{true}) \leq C$. *Then,*

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}^{true}.$$

Under the same conditions as in Theorem 3, for any $\alpha \in (0, 1/2)$, *with probability at least* $1 - 2\alpha$,

$$\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^*\|_2 \leq 2b_4\epsilon \quad \textit{if} \quad k \geq \log_{1/r} \frac{b_3\|\boldsymbol{\beta}^*\|_2}{b_4\epsilon}.$$

Furthermore, assume that $\bar{m} > b_1\epsilon/\xi$ *for some* $0 < \xi < 1$, *then we have*

$$\mathcal{I}^* \subset \mathcal{I}^k \quad \textit{if} \quad k \geq \log_{1/r} \frac{\|\boldsymbol{\beta}^*\|_2}{(1-\xi)\bar{m}}.$$

Corollary 1 demonstrates the ability of the FCR algorithm to effectively detect nonzero coefficients in the situation of $\sum_{i=1}^{p_0} c_j I(\beta_j^{true}) \leq C$. Initially, in that case, we show that the desired solution is the nonzero coefficients. Moreover, it establishes that after a sufficiently large number of iterations, the ℓ_2-distance between the k-th estimates and the true coefficients has the bound of ϵ up to a constant after a sufficiently large iteration with high probability. Furthermore, under the additional assumption that the minimum value of the true $\boldsymbol{\beta}^*$ is sufficiently larger than ϵ, the FCR algorithm is shown to accurately identify a subset of nonzero coefficients. It provides a practical applicability of FCR that ensures that the budget is not wasted, at least when the budget is enough.

Our findings align with those presented in Theorem 4 and Corollary 8 of Huang et al. (2018), which delves into the best subset selection problems. Despite addressing the more challenging nonconvex constraints, we successfully obtain comparable outcomes under identical circumstances. The difficulty in formulating theorems lies in identifying sets $\mathcal{I}^{k+1}$,$\mathcal{I}^k$, and $\mathcal{I}^*$. Unlike Huang et al. (2018), where sets have fixed sizes, ours vary based on iteration and cost-constraint structures. Lemma 4 in the Supplement S1 establishes the association among these sets.

4. Numerical Results

4.1 A motivating example

Imagine a scenario in which the response $\boldsymbol{y}$ is generated from $\mathbf{X}\boldsymbol{\beta}^{true} + \zeta$. We choose $n = 2000$, $p = 1000$, $p_0 = 8$, and $\mathbf{X}$ is generated from a multivariate Gaussian distribution with mean 0, with correlations among the first 24 displayed in the left panel of Figure 1. In particular, variables

$(x_1, \ldots, x_8)$ and $(x_{17}, \ldots, x_{24})$ exhibit a significant correlation. In instances where certain variables are prohibitively expensive, alternative variables should be explored as substitutes.

In order to make such scenarios, the first 24 entries of $\boldsymbol{\beta}^{true}$ and $\boldsymbol{c}$ are set to have their components as following,

$$\boldsymbol{\beta}^{true}_{1:24} = (\underbrace{\beta_1, \ldots, \beta_4, \beta_5, \ldots, \beta_8}_{\mathbf{2}}, \underbrace{\beta_9 \ldots, \beta_{16}, \beta_{17} \ldots, \beta_{24}}_{\mathbf{0}})$$

$$\boldsymbol{c}_{1:24} = (\underbrace{c_1, \ldots, c_4}_{\$20}, \underbrace{c_5, \cdots c_8}_{\$10}, \underbrace{c_9 \ldots, c_{16}, c_{17} \ldots, c_{24}}_{\$1 \text{ or } \$2}),$$

and the rest of $\boldsymbol{\beta}^{true}$ is set to be 0 and that of $\boldsymbol{c}$ \$1 or \$2. In such case, consider a scenario where the maximum budget available is limited to \$50. Despite the significance of the first eight variables, using all of them incurs a cost of $\$20 \times 4 + \$10 \times 4 = \$120$. Given the relatively high cost of the first four variables, β_1–β_4, a sensible approach is to opt for variables that exhibit strong correlation with them but have a lower cost, such as β_{17}–β_{20}, requiring only \$1 or \$2 each. Through a brutal force search, we find that the optimal variables for this situation are β_5–β_8 and β_{17}–β_{20} (middle, and upper right solid line in Figure 1), and obtain the solution value of (1.1) as well. We experiment whether the feasible algorithms select the correct variables. To evaluate the performance under these conditions, we generate

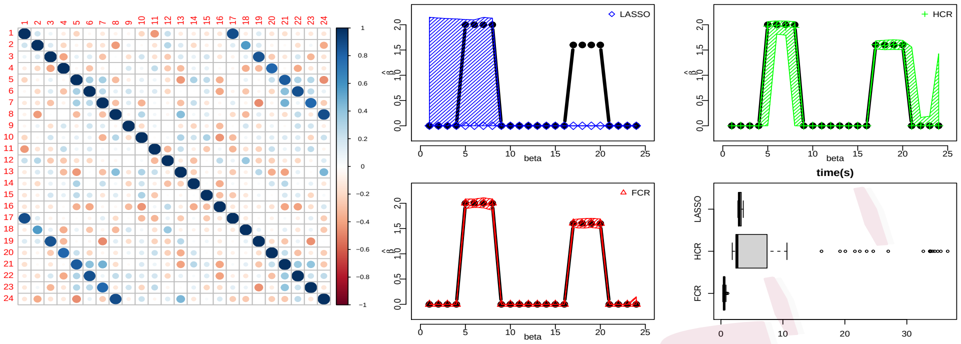

Figure 1: Left: correlation among the columns of $\mathbf{X}_{1:24}$. middle: $\beta's$ true solution of (1.2) (black line) and 90% bands of estimated values of LASSO methods (blue shade) and FCR (red shade) of 100 simulations. right: the HCR (green shade) performance and the implementation time

100 different pairs of $\mathbf{X}$ and $\boldsymbol{y}$ and compare their outcomes.

One of the naïve but feasible approaches is to try to find sufficient values of the hyperparameter of the regularization term and to find feasible sets of β_js that satisfy the cost restrictions. Among them, the prediction error can be estimated and the combination of β_js that minimize the error can be chosen. We apply LASSO (Tibshirani, 1996) to this approach. We compare LASSO, HCR, and FCR in the above scenario and plot the results on the middle and right sides of Figure 1. We iterate three algorithms 100 times with various pairs of $\mathbf{X}$ and $\boldsymbol{y}$, showcasing the 90% confidence bands for the estimated values of each method. These bands are then overlaid on the plot depicting the true solution. As depicted in the upper middle section of Figure 1, LASSO tends to exhibit conservative variable selection, occasionally favoring less favorable options among the initial four coefficients while neglecting other cost-effective variables. Although the methods successfully identify nonzero coefficients, this behavior results in lower prediction errors, primarily due to the lack of systematic consideration of cost constraints in variable selections. On the other hand, FCR and HCR tend to select relatively desirable variables (lower middle and upper right of Figure 1). The FCR accurately selects the desired variables β_5–β_8 and β_{17}–β_{20}, occasionally incorporating β_{24}, although with a negligible estimated value. Meanwhile, the HCR tends to make appropriate variable choices; however, its estimated values exhibit a larger variance compared to the FCR estimates. The HCR method occasionally selects cost-ineffective variables β_{22}-β_{24} with larger estimated values. We measure the time taken by the three methods and display the results in the lower right panel of Figure 1. The FCR completes its estimate in less than a second, whereas the HCR exhibits a longer time elapsed with considerable variability.

4.2 Simulation results

To study the properties of each algorithm, we investigated each method in various scenarios. The simulation settings in this section reflect typical scenarios encountered when costs are involved in the implementation of

Table 2: Simulation setting summaries

	Model	c_j	C_0	C	p	p_0	$\rho_{ij}(p_0)$
S1	Linear Regression	1:10	146	25, 50, 75, 100	1000	24	0
S2	Linear Regression	1:10	146	80	1000	16	0.2, 0.4, 0.6, 0.8
S3	Linear Regression	1:50	686	200, 400, 600, 800	1000	24	0.3
S4	Linear Regression	1:10 grouped	68	15, 30, 45, 60	1000	32	0
S5	Logistic Regression	1:10	146	25, 50, 75, 100	1000	24	0.5
S6	Logistic Regression	1:10	146	80	1000	24	0.2, 0.4, 0.6, 0.8

regression tasks. Both linear and logistic regression are considered, and so are the cases of high-dimensional ($n < p$) to low-dimensional ($n \geq p$). We fix $p = 1000$ for all simulations, while the cost scale, the amount of budget, and the correlation among the covariates $X_j's$ are varied to highlight the effectiveness of our method depending on various situations. We investigate six different simulations and denote them as **S1** to **S6**. While we describe the detailed simulation settings in the following subsection and Supplement S5, they are summarized in Table 2.

To evaluate the performance of these methods, we use five measures: (1). The prediction error $\|\boldsymbol{y}^{\text{test}} - \mathbf{X}^{\text{test}}\hat{\boldsymbol{\beta}}\|_2^2/n_{\text{test}}$ for the regression task and $\frac{1}{n}\left|\boldsymbol{y}^{\text{test}} \neq \text{sign}(\mathbf{X}^{\text{test}}\hat{\boldsymbol{\beta}})\right|$ for the logistic model; (2). The false negative rate (FNR*):$= |\{j : \beta_j^* \neq 0, \hat{\beta}_j = 0\}|/|\{j : \beta_j^* \neq 0\}|$; (3). The false positive rate (FPR*):$= |\{j : \beta_j^* = 0, \hat{\beta}_j \neq 0\}|/|\{j : \beta_j^* = 0\}|$; (4). The estimation error $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2^2$; (5). The elapsed time of the algorithms in the software R, where $\boldsymbol{\beta}^*$ is the solution to the optimization problem (1.1); (6) The area under the receiver operating characteristic curve (AUC) for the logistic model. Through an exhaustive search, we can identify $\boldsymbol{\beta}^*$ in straightforward scenarios such as **S1** and **S2** . However, for **S3** to **S6**, since obtaining the true set of $\boldsymbol{\beta}^*$ is challenging both analytically and computationally, we only report the prediction error, AUC and the elapsed time. FNR* measures how less the algorithm selects cost-efficient variables. If β_j^* is not selected, the algorithm will lose cost-efficiency even though the algorithm buys other nonzero significant variables. On the other hand, FPR* measures how much the algorithm selects cost-inefficient variables. However, the larger value of FPR* error does not necessarily mean that the algorithm works in a cost-inefficient way. Even when none of the variables is selected, the FPR* error should be zero, but it provides undesirable estimates. Therefore, careful consideration is needed to interpret the FPR* measures.

As illustrated above, we compare our method with two other methods. First, we take into account the comparison with HCR (Yu et al., 2022), which deals first with the cost-constrained regression problem. We choose the hyperparameter as the largest eigenvalue of $\mathbf{X}^\top\mathbf{X}/n$ plus 0.1 in the same way as Yu et al. (2022) to set in their simulations. Next, we compare

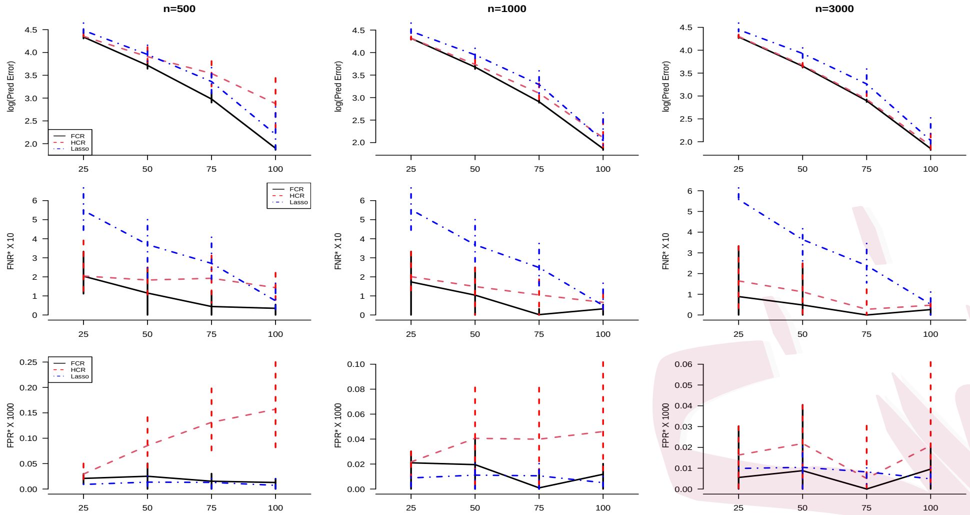

Figure 2: S1 results of the prediction error (the first row), and the FNR* (second), and FPR* (third). FCR, HCR, and LASSO are indicated in the solid, dashed, and dot dashed line, respectively

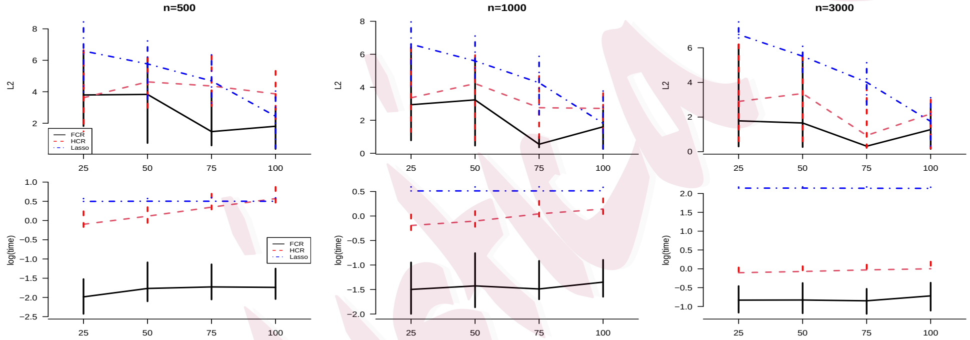

Figure 3: S1 results of the estimator error (the first row), and the elapsed time (second).

with LASSO, which finds a nonzero $\boldsymbol{\beta}'s$ using cross-validation using the package glmnet (Tay et al., 2023) and naïvely search for a combination of β_js that minimizes objective functions. The elapsed time for LASSO include solving LASSO for a path of regularization parameters and find feasible sets of parameters.

In Simulation 1 (**S1**), 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. Regarding the generation of the first $p_0 = 24$ elements of $\boldsymbol{\beta}^{true}$, the first quarter of them is generated from $N(4, 0.25)$, the next quarter from $N(3, 0.25)$, the next $N(2, 0.25)$, and finally $N(1, 0.25)$. 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 $\{1, 2, \ldots, 10\}$. Throughout this section, we generate different $\mathbf{X}$ 100 times and iterate each algorithm 100 times, and indicate the mean of the measures and its 95% confidence interval in the figures.

We divide the simulation results into Figures 2 and 3. Figure 2 summarizes the performance of the prediction error, FNR* and FPR*, and Figure 3 the estimation error and the time elapsed over the budget $C = \{25, 50, 75, 100\}$. We can see that the predictive error of the FCR method decreases as the budget increases. The total budget to fully explain the response variable is 146, which the model performs better as the budget approaches. Compared to other methods, the FCR method shows superior predictive performance and small variance (cf. the first row of Figure 2). This is distinctive when n is small (cf. top left in Figure 2), but it is still the case when n is medium and large (see the top middle and top right of Figure 2). The second row of Figure 2 indicates the FNR* for each method. The FCR method finds a cost-effective variable uniformly, regardless of whether the budget is tight or abundant. The third row of Figure 2 indicates the FPR* for each method. Compared to HCR, FCR chooses less cost-inefficient variables over budget and sample size. We can see that HCR tends to spend its budget inefficiently as the budget increases. Interestingly, the naïve LASSO method shows comparably good FPR* performance. This is because the naïve LASSO spends the budget conservatively so that it does not search for variables regardless of redundancy. The FPR* of FCR is comparable to that of this frugal method. The lower FNR* and FPR* of FCR lead to the top-quality performance in the estimator power as shown in the first row of Figure 3. Moreover, the superior achievement of FCR is achieved in a much shorter time than other methods. The naïve LASSO approach needs a similar time no matter what the sample size is due to its brutal search nature. For HCR, the elapsed time is similar to the naïve LASSO approach for small n, but it improves as the sample size increases. On the other hand, FCR takes only a much shorter time to implement compared with the other two methods.

For Simulation 2-6 (**S2-S6**), we obtain similar results to those above. As summarized in Table 2, we vary the correlation ρ among the covariates of **X** (**S2**), the cost distribution (**S3**), and grouped costs that are discussed in Section S4 (**S4**). We also consider the logistic regression discussed mainly in Section 2.2 in **S5** and **S6**.

We detail the simulation settings and report the corresponding results in Supplement S5. It is important to emphasize that for non-convex optimization problems like this, one method may not be universally superior to the other. Despite this, our proposed FCR method demonstrates distinct advantages over HCR as seen in the simulation results. The FCR method not only tends to perform better in high-dimensional cases but also achieves comparable or superior results with significantly fewer iterations, resulting in greater computational efficiency.

5. Real Data Analysis

We perform a real-world analysis with the four data sets described in Table 1. Kachuee et al. (2019) performed a survey to assign an associated cost (ranging from \$2 to \$9) to each feature, taking into account subject factors such as convenience and subjective usefulness of each feature. We also refine the data by removing less common questionnaire responses or other categories. As described in Table 1, most variables in diabetes data are continuous, while most heart data are categorical. Table S1 in Supplement

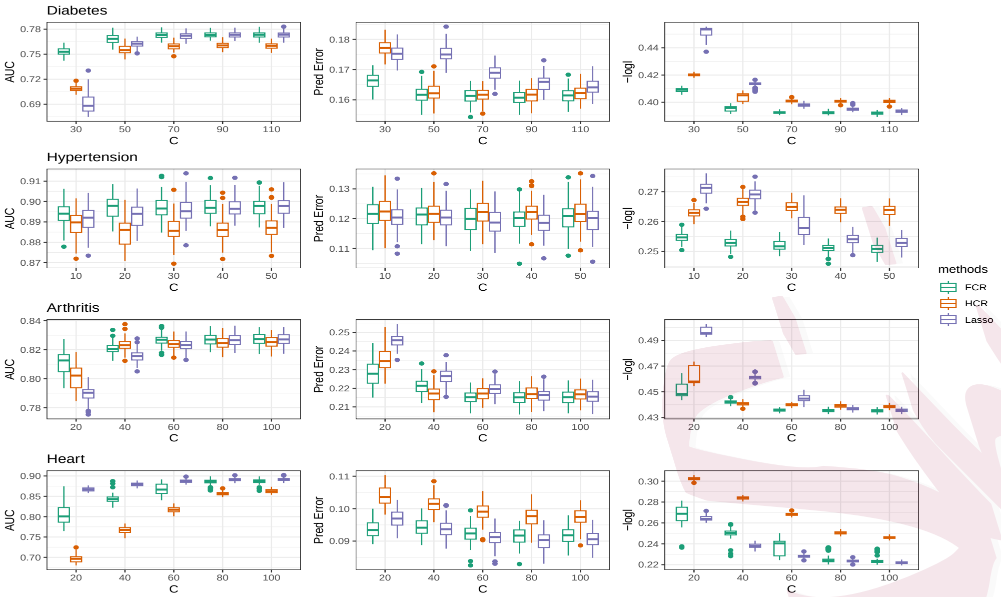

Figure 4: NHANES data analysis summary: For every budget C, the boxplot on the left represents FCR, while the one on the right represents HCR

Section S6 illustrates the cost distribution for each dataset. By examining the four sets of data, we can assess the effectiveness of the algorithms in various scenarios.

We implement logistic grouped cost-constrained regression for each dataset and compare it with HCR and LASSO. For each dataset, we randomly split it into training and test datasets, with a ratio of 0.8 and 0.2. While we iterate the split of the data 100 times, we record AUC, prediction error, negative log-likelihood, and time elapsed on various budgets C.

The superiority of FCR over HCR is evident in real-world applications. The box plot of each performance is presented in Figure 4 on the budgets for both methods (AUC, the prediction error, and the negative log-likelihood

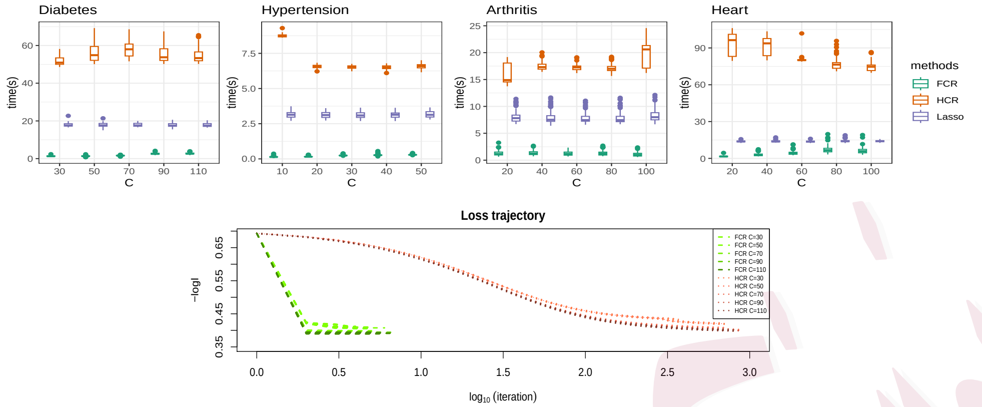

Figure 5: Top: the elapsed time in R for NHANES data analysis, bottom: the diabetes study's loss function trajectory of FCR (dashed) and HCR (dotted) for each budget

from left to right column). The first row corresponds to the diabetes study, followed by hypertension in the second, arthritis in the third, and heart in the final row. Despite variations in the characteristics of the four data sets, FCR consistently demonstrates a higher AUC, a lower prediction error, and a lower negative log-likelihood than HCR. This consistent superiority underscores the practical efficacy of FCR in its application to clinical data. In the context of budget considerations, improved performance suggests substantial potential for cost savings.

In addition, FCR achieves its outstanding performance in a considerably shorter time frame. In the diabetes study, FCR is, on average, 28 times faster than HCR, and in the hypertension study it is 30 times faster, as seen in the top of Figure 5. To closely see the difference, we plot the loss function trajectories for both methods in the diabetes study in the bottom of Figure 5. FCR accomplishes its task in fewer than 10 iterations, whereas HCR requires a minimum of 300 iterations to complete its task. We present the trajectories of the other studies in Figure S7 in Supplement Section S6.

When compared to LASSO, FCR performs better in terms of predictive accuracy for diabetes, arthritis, and hypertension datasets, but underperforms for the heart dataset. This discrepancy appears to stem from the configuration of the cost vectors in the heart dataset. The LASSO method can be preferable when (i) the significant variable is not expensive enough, so that (ii) the cheaper variables cannot sufficiently explain the response instead of the expensive one. This conservative approach works well for the heart dataset, where LASSO effectively selects the key variables. Table S1 shows that the number of variables in the heart dataset whose cost 5 is zero and cost 9 is very small, which turns out to be insignificant. The LASSO method, as expected, tends to identify nonzero coefficients conservatively, selecting significant variables, some of which are not that expensive, at most cost 4, and others are even cheap with cost 2. On the other hand, FCR selects nonzero coefficients in a more distributed manner, emphasizing variables with lower costs (e.g., cost 2), which results in suboptimal performance for the heart dataset.

In contrast, for the diabetes, arthritis, and hypertension datasets, the cost constraints are more intricate, making LASSO's conservative approach less effective. The complex cost structure in these datasets demands a method like FCR, which is specifically designed to handle cost-constrained regression problems. This aligns with our initial motivation for this article and highlights the need for a method tailored to cost-sensitive scenarios.

6. Conclusion

This paper introduces FCR, a novel method adapted for scenarios in which variables incur costs. The FCR gains its solution in an accelerated way and reaches it with robust predictive modeling capabilities. We establish a non-asymptotic error bound for FCR, ensuring desirable characteristics.

FCR has proven to be promising, but there is still room for growth in both theoretical and algoritmic ways. Elucidating the precise rationale behind the FCR poses a challenge due to the non-convex nature and the complicated structure of the cost constraint $\mathcal{C}$. Another direction for further development is to adapt the framework to more general regression settings, such as nonlinear models and random forests. Expanding the application of FCR to such models not only diversifies the options available to users, but improves its predictive capabilities.

References

Borchers, H. W. (2022). *adagio: Discrete and Global Optimization Routines*. R package version 0.8.5.

Chen, Y., Q. Gao, F. Liang, and X. Wang (2020, 09). Nonlinear variable selection via deep neural networks. *Journal of Computational and Graphical Statistics 30*, 1–23.

Erion, G., J. D. Janizek, C. Hudelson, R. B. Utarnachitt, A. M. McCoy, M. R. Sayre, N. J. White, and S.-I. Lee (2022). Coai: Cost-aware artificial intelligence for health care. *Nature biomedical engineering 6*(12), 1384.

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.

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

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

Jin, C. (2024). 0-1 knapsack in nearly quadratic time. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 271–282.

Kachuee, M., O. Goldstein, K. Karkkainen, S. Darabi, and M. Sarrafzadeh (2019). Opportunistic learning: Budgeted cost-sensitive learning from data streams. *arXiv preprint arXiv:1901.00243*.

Martello, S. and P. Toth (1990). *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, Inc.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM journal on computing 24*(2), 227–234.

Nguyen, S., R. Chan, J. Cadena, B. Soper, P. Kiszka, L. Womack, M. Work, J. M. Duggan, S. T. Haller, J. A. Hanrahan, et al. (2021). Budget constrained machine learning for early prediction of adverse outcomes for covid-19 patients. *Scientific Reports 11*(1), 19543.

Peter, S., F. Diego, F. A. Hamprecht, and B. Nadler (2017). Cost efficient gradient boosting. *Advances in neural information processing systems 30*.

Tay, J. K., B. Narasimhan, and T. Hastie (2023). Elastic net regularization paths for all generalized linear models. *Journal of statistical software 106*.

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

Wang, X. (2008). Bayesian free-knot monotone cubic spline regression. *Journal of Computational and Graphical Statistics 17*(2), 373–387.

Wang, X. and J. Shen (2010, 06). A class of grouped brunk estimators and penalized spline estimators for monotone regression. *Biometrika 97*(3), 585–601.

Wang, X. and J. Shen (2013). Uniform convergence and rate adaptive estimation of convex functions via constrained optimization. *SIAM Journal on Control and Optimization 51*(4), 2753–2787.

Wang, X. and M. Woodroofe (2007). A Kiefer–Wolfowitz comparison theorem for Wicksell's problem. *The Annals of Statistics 35*(4), 1559 – 1575.

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

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

Yu, G., D. Witten, and J. Bien (2022). Controlling costs: Feature selection on a budget. *Stat 11*(1), e427.

Zhang, C.-H. and J. Huang (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics 36*, 1567–1594.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*, 2541–2563.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology 67*(2), 301–320.