Technometrics

ISSN: (Print) (Online) Journal homepage: https://amstat.tandfonline.com/loi/utch20

High-Dimensional Cost-constrained Regression Via Nonconvex Optimization

Guan Yu, Haoda Fu & Yufeng Liu

To cite this article: Guan Yu, Haoda Fu & Yufeng Liu (2022) High-Dimensional Cost-constrained Regression Via Nonconvex Optimization, Technometrics, 64:1, 52-64, DOI: 10.1080/00401706.2021.1905071

To link to this article: https://doi.org/10.1080/00401706.2021.1905071

TECHNOMETRICS
2022, VOL. 64, NO. 1, 52–64
https://doi.org/10.1080/00401706.2021.1905071

Taylor & Francis
Taylor & Francis Group

High-Dimensional Cost-constrained Regression Via Nonconvex Optimization

Guan Yu[a], Haoda Fu[b], and Yufeng Liu[c]

[a]Department of Biostatistics, State University of New York at Buffalo, NY; [b]Advanced Analytics and Data Sciences, Eli Lilly and Company, Indianapolis, IN; [c]Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC

ABSTRACT

Budget constraints become an important consideration in modern predictive modeling due to the high cost of collecting certain predictors. This motivates us to develop cost-constrained predictive modeling methods. In this article, we study a new high-dimensional cost-constrained linear regression problem, that is, we aim to find the cost-constrained regression model with the smallest expected prediction error among all models satisfying a budget constraint. The nonconvex budget constraint makes this problem NP-hard. In order to estimate the regression coefficient vector of the cost-constrained regression model, we propose a new discrete first-order continuous optimization method. In particular, our method delivers a series of estimates of the regression coefficient vector by solving a sequence of 0-1 knapsack problems. Theoretically, we prove that the series of the estimates generated by our iterative algorithm converge to a first-order stationary point, which can be a globally optimal solution under some conditions. Furthermore, we study some extensions of our method that can be used for general statistical learning problems and problems with groups of variables. Numerical studies using simulated datasets and a real dataset from a diabetes study indicate that our proposed method can solve problems of fairly high dimensions with promising performance.

ARTICLE HISTORY
Received October 2019
Accepted March 2021

KEYWORDS
Budget constraint; Dynamic programming; High-dimensional regression; Nonconvex optimization; 0-1 knapsack problem

1. Introduction

High-dimensional predictive modeling plays a fundamental role in modern statistical machine learning. In order to obtain a good model to fit high-dimensional data, many popular methods use penalized regression techniques. The corresponding optimization problem is to minimize an objective function with the form of a loss function plus a convex or nonconvex penalty function (Frank and Friedman 1993; Tibshirani 1996; Fan and Li 2001; Zou and Hastie 2005; Zou 2006; Zhang 2010). Most existing penalized regression methods sought to improve the accuracy of estimation and prediction but often failed to account for the costs on data collection. However, in some predictive modeling applications, especially in health care, it is essential to account for the costs associated with data collection. Note that the notion of cost here can be general. It includes both the actual monetary cost and some nonmonetary costs such as time, patient discomfort in medical procedures, and privacy impacts of data collection (Kachuee et al. 2019; Krishnapuram, Yu, and Rao 2011; Pattuk et al. 2015).

As an example, in the treatment of diabetes mellitus, it is important to predict the patients' treatment responses (e.g., the change in HbA1c) before assigning treatments so that doctors can select the treatment with the largest potential outcome for each patient. According to a randomized, double-blind, parallel-group comparison phase III study of diabetes (Charbonnel et al. 2005), twenty biomarkers for the prediction of diabetes patients' treatment responses and their costs are shown in Table 1. These twenty biomarkers are divided into 10 groups naturally according to the data collection process. The acquisition of some biomarkers incurs higher costs than others. For example, the values of HDL, LDL, Total cholesterol, and Triglycerides are generated together by a blood test. The price of the blood test is about $200, which is much more expensive than the measurement of blood pressure. As it can be expensive and inconvenient to let patients measure all these 20 biomarkers to predict their response, the traditional regression approach incorporating all biomarkers are infeasible in some situations. It is of practical importance to develop flexible predictive models that can help doctors predict the patients' treatment responses as accurately as possible while controlling the diagnostic cost. As a second example, in the diagnosis of kidney stones, medical imaging techniques such as computed tomography (CT), ultrasonography, kidney ureter bladder (KUB) plain film radiography, and magnetic resonance imaging (MRI) are widely used. The comparison of different imaging modalities for kidney stones is shown in Brisbane, Bailey, and Sorensen (2016). All imaging techniques have advantages and disadvantages. Some modalities such as MRI are much more expensive than the KUB. Besides the diagnostic accuracy, doctors generally prescribe the imaging modality for kidney stones based on a number of factors including monetary costs, patient body habitus, and tolerance of ionizing radiation. Cost-constrained predictive modeling methods are needed to help doctors make the decision. Besides, the above two applications in health care, as shown in Clark et al. (2019),

CONTACT Yufeng Liu yfliu@email.unc.edu Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TECH.

Table 1. Biomarkers and their costs in a diabetes study.

Predictor	Cost	Predictor	Cost
HDL	$200	Creatinine	$50
LDL		Fasting blood glucose	$20
Total cholesterol		HbA1c at baseline	$30
Triglycerides		ALT	$100
Fasting insulin	$50	AST	
Age	$20	GGT	$100
Weight		C-peptide	
BMI		Diastolic blood pressure	$10
Waist		Systolic blood pressure	
Duration of diabetes	$5	Pulse	

cost-constrained predictive modeling methods are useful for the problem of optimally placing sensors under a cost constraint that arises naturally in the design of industrial and commercial products, as well as in scientific experiments.

In the literature, there are a few predictive modeling methods accounting for the cost of collecting variables. For the linear regression problem, Yue (2010) suggested adding the cost of obtaining variables to the least-squares loss function. A branch and bound algorithm is used to search for a model which minimizes the total loss. For the binary classification problem, Fouskakis and Draper (2008) compared several stochastic optimization methods such as simulated annealing (Kirkpatrick et al. 1983), genetic algorithm (Holland 1992), and Tabu search (Glover 1977, 1986, 1989) to heuristically find subsets of variables that maximize a utility function which trades prediction accuracy against data collection cost. Fouskakis, Ntzoufras, and Draper (2009) proposed a Bayesian approach that accounts for the cost of variables via prior model weights, which leads to a generalized cost-adjusted version of the Bayesian Information Criterion. They used the reversible-jump Markov chain Monte Carlo (RJMCMC) method to search the predictive model. Although these existing methods deliver good performance for the problem with a small number of variables (e.g., $p < 100$ as shown in their simulation studies), they require expensive computation to search the optimal model when there are a lot of variables (e.g., $p = 1000$). In addition, they seek the variable subset by minimizing an objective function which trades prediction accuracy against the data collection cost. This is a soft approach of handling a hard budget constraint that demands the total cost of collecting the variables to not exceed a given budget. For many modern predictive modeling problems such as the prediction of disease risk using medical images, we need to choose variables from a large variable set. In addition, hard budget constraints become more and more common due to the high cost of collecting data using new techniques. It is important to develop new and efficient high-dimensional cost-constrained predictive modeling methods.

In this article, to address the above challenge, we first study a new high-dimensional cost-constrained regression problem, that is, we aim to find the cost-constrained regression model that satisfies the budget constraint and has the best prediction accuracy. This problem generalizes the best subset selection problem where the costs of all variables are assumed to be equal. The nonconvex budget constraint makes this problem NP-hard. For the high-dimensional cost-constrained regression problem considered in this article, the parameter vector of interest is the regression coefficient vector in the cost-constrained regression model. This parameter vector of interest can be different from the parameters in the underlying data-generating linear model when the underlying true model does not satisfy the budget constraint. As shown in Section 2.2, it can happen that none of the selected variables in the cost-constrained regression model is used in the underlying true linear model. Therefore, we cannot implement a simple two-step strategy that uses the existing penalized regression techniques to select the variables first and then searches for the cost-constrained regression model among those selected variables. In order to estimate the regression coefficients in the cost-constrained regression model directly, we propose a new discrete extension of the first-order continuous optimization methods (Nesterov 2013). Our proposed method delivers a convergent series of estimates of the parameter of interest by solving a sequence of 0-1 knapsack problems. Theoretically, we show that the series of the estimates of the parameter of interest generated by our iterative algorithm converge to a first-order stationary point, which can be a globally optimal solution when some conditions are satisfied. There are many extensions of our proposed cost-constrained regression method. It can be extended to statistical learning problems using loss functions with a Lipschitz continuous gradient, for example, the cost-constrained logistic regression problem. We can also adjust the proposed cost-constrained regression method for problems with groups of variables, or combine it with regularization techniques to reduce overfitting. Our numerical studies indicate that our algorithm can solve the high-dimensional problems efficiently with good estimation, prediction, and model selection performance. Both theoretical and numerical studies demonstrate the effectiveness of our proposed method.

The rest of this article is organized as follows. In Section 2, we introduce the high-dimensional cost-constrained regression problem. In Section 3, we introduce our new method. In Section 4, we show the extensions of our proposed high-dimensional cost-constrained regression method. In Section 5, we show some theoretical results about our iterative algorithm. In Sections 6 and 7, we demonstrate the effectiveness of our method using simulated datasets and a real dataset from a diabetes study. We conclude this article in Section 8. All proofs and the comparison of the computational time of different methods are shown in the supplementary materials for this article that are available online.

2. High-Dimensional Cost-Constrained Regression

In this section, we introduce the high-dimensional cost-constrained regression problem. More general cost-constrained predictive modeling problems will be discussed in Section 4.

We use the following notations throughout this article. The complement of a set S is denoted by S^c. For a vector V and a set S, we use V_S to denote the subvector $\{V_j : j \in S\}$, and $||V||_2$ to denote the ℓ_2 norm of the vector. For two vectors X and Y, we use $\langle X, Y\rangle$ to denote the inner product. For a matrix $\mathbf{M}$ and sets S_1 and S_2, we use $\mathbf{M}_{S_1S_2}$ to denote the submatrix of $\mathbf{M}$ with the row indices in the set S_1 and the column indices in the set S_2. For a symmetric matrix $\mathbf{A}$, we use $\lambda_{\max}(\mathbf{A})$ to denote the largest eigenvalue of the matrix. We use $\mathcal{I}(\beta)$ to

denote the indicator function which equals to 1 if $\beta \neq 0$ and 0 otherwise. The gradient of a function $g(\beta)$ is denoted by $\nabla g(\beta)$. For a vector $X = (x_1, x_2, \ldots, x_p)$, we use $(X)_+$ to denote $(\max\{x_1, 0\}, \max\{x_2, 0\}, \ldots, \max\{x_p, 0\})$. Given an arbitrary set $S \subseteq \mathbb{R}^p$ and a function $g : \mathbb{R}^p \to \mathbb{R}$, the arg min over the subset S is defined by $\arg\min_{\beta \in S} g(\beta) := \{\beta | \beta \in S \text{ and } g(\beta) \leq g(\gamma) \text{ for any } \gamma \in S\}$.

2.1. High-Dimensional Cost-Constrained Regression Problem

Suppose that our data are generated from the following linear model:

$$y = x_1\beta_1^0 + x_2\beta_2^0 + \cdots + x_p\beta_p^0 + \epsilon, \tag{1}$$

where y is a continuous response variable, $\beta_1^0, \beta_2^0, \ldots, \beta_p^0$ are p true unknown parameters, $(x_1, x_2, \ldots, x_p)^T$ is a p-dimensional predictor vector following a multivariate distribution with mean $0_{p\times 1}$ and a positive-definite covariance matrix $\mathbf{\Sigma}$, and ϵ is the random error with mean 0 and variance σ^2. We assume that the random predictor vector $(x_1, x_2, \ldots, x_p)$ and the random error ϵ are independent. Suppose we have n iid samples generated from the model (1). Denote $Y = (y_1, y_2, \ldots, y_n)^T$, $\beta^0 = (\beta_1^0, \beta_2^0, \ldots, \beta_p^0)^T$, and the $n \times p$ design matrix by $\mathbf{X}$ where the ith row of $\mathbf{X}$ *is* $(\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{ip})$. Then, in matrix notation, we have

$$Y = \mathbf{X}\beta^0 + \xi, \tag{2}$$

where $\xi = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ is a vector of n iid realizations of the random variable ϵ.

We consider the high-dimensional regime ($p \gg n$) and assume that the p-dimensional regression coefficient vector β^0 is sparse. Denote $S = \{j : \beta_j^0 \neq 0\}$, and $S^c = \{j : \beta_j^0 = 0\}$. Without loss of generality, suppose that there are monetary costs on collecting variables. For each sample, we need to spend c_j dollars on collecting the value of the jth predictor x_j, where $j = 1, 2, \ldots, p$. If we do not consider the costs on data collection, then we can use any penalized regression technique to learn a linear model, which can be used in the future to predict the value of the response variable y given specific values of the predictors $x_1, x_2, \ldots, x_p$. However, for some applications such as medical diagnosis problems, we need to consider the costs on data collection and our budget on collecting values of $x_1, x_2, \ldots, x_p$ (e.g., different medical tests) is C dollars. If the budget limit C is small, then we may not have enough money to collect the values of all predictors used in the linear models learned by penalized regression methods. Therefore, due to the budget constraint, linear models learned by traditional penalized regression techniques can be infeasible for the prediction of y. We can only consider feasible linear models satisfying the budget constraint $\sum_{j=1}^p c_j\mathcal{I}(\beta_j) \leq C$, where β is the regression coefficient vector of the linear model and $\sum_{j=1}^p c_j\mathcal{I}(\beta_j)$ is the cost of the model. It is important to develop new methods to find the optimal feasible linear model that has the smallest expected prediction error among all feasible models. This optimal feasible model is called the cost-constrained regression model in this article.

Assume that the future observation $(y_*, x_{*1}, \ldots, x_{*p})$ is also generated from model (1). We can show that the regression coefficient vector of the cost-constrained regression model is

$$\beta^* \in \arg\min_\beta \mathrm{E}[(y_* - \sum_{j=1}^p x_{*j}\beta_j)^2] \quad \text{subject to} \quad \sum_{j=1}^p c_j\mathcal{I}(\beta_j) \leq C,$$

which is also a global minimizer of the following problem:

$$\min_\beta (\beta - \beta^0)^T\mathbf{\Sigma}(\beta - \beta^0) \quad \text{subject to} \quad \sum_{j=1}^p c_j\mathcal{I}(\beta_j) \leq C. \tag{3}$$

Note that there may be multiple global minimizers of the problem (3) due to the budget constraint. In that case, we use β^* to denote one of those global minimizers. Given the training data $\{Y, \mathbf{X}\}$, it is natural to estimate β^* by solving the following sample-average approximation (SAA) problem:

$$\min_\beta \frac{1}{2n}||Y - \mathbf{X}\beta||_2^2 \quad \text{subject to} \quad \sum_{j=1}^p c_j\mathcal{I}(\beta_j) \leq C. \tag{4}$$

The above problem is called the high-dimensional cost-constrained regression problem in this article. It can be viewed as a generalized best subset selection problem (the case with $c_1 = c_2 = \cdots = c_p$). Note that the constraint $\sum_{j=1}^p c_j\mathcal{I}(\beta_j) \leq C$ makes this problem (4) NP-hard. Indeed, even for the special case such as the best subset selection problem, as shown in Bertsimas et al. (2016), in order to find the global solution, most state-of-the-art algorithms as implemented in the popular statistical packages do not scale to problems with more than 30 variables.

2.2. The Difference Between β^* and β^0

In this article, since we aim to find the cost-constrained regression model, the parameter of interest is β^* rather than the parameter vector β^0 in the true linear model (1). If $\sum_{j=1}^p c_j\mathcal{I}(\beta_j^0) = \sum_{j\in S} c_j \leq C$, we know that β^0 is a feasible solution to (3) and therefore $\beta^* = \beta^0$. However, as showing in the following analysis, β^* *and* β^0 can be different if $\sum_{j\in S} c_j > C$.

Consider the case where the true important variables $\{x_j : j \in S\}$ and the unimportant variables $\{x_j : j \in S^c\}$ are uncorrelated, that is $\mathbf{\Sigma}_{SS^c} = \mathbf{0}$, we can show that the problem (3) is equivalent to the following problem

$$\min_\beta (\beta_S - \beta_S^0)^T\mathbf{\Sigma}_{SS}(\beta_S - \beta_S^0) + \beta_{S^c}^T\mathbf{\Sigma}_{S^cS^c}\beta_{S^c}$$
$$\text{subject to} \sum_{j\in S} c_j\mathcal{I}(\beta_j) + \sum_{j\in S^c} c_j\mathcal{I}(\beta_j) \leq C.$$

Since $\beta_{S^c}^T\mathbf{\Sigma}_{S^cS^c}\beta_{S^c} \geq 0$ for any β, it implies $\beta_{S^c}^* = 0$ and

$$\beta_S^* \in \arg\min_\beta (\beta_S - \beta_S^0)^T\mathbf{\Sigma}_{SS}(\beta_S - \beta_S^0)$$
$$\text{subject to} \sum_{j\in S} c_j\mathcal{I}(\beta_j) \leq C.$$

If $\sum_{j\in S} c_j > C$, we have $\beta_S^* \neq \beta_S^0$ and $\{j : \beta_j^* \neq 0\} \subset \{j : \beta_j^0 \neq 0\}$. In this case, using the training data $(Y, \mathbf{X})$, we can implement

a two-step strategy that uses the penalized regression techniques (Tibshirani 1996; Fan and Li 2001; Zou and Hastie 2005; Zou 2006; Zhang 2010) to select a small number of variables first, and then search the cost-constrained regression model considering only those selected variables. Since many penalized regression techniques such as LASSO are model selection consistent (Zhao and Yu 2006), this two-step strategy could save a lot of computational time and deliver good performance for the estimation of the parameter of interest. However, in many cases, as shown in the following toy example, the selected variables in the cost-constrained regression model satisfying the budget constraint may not be in the important variable set S.

Toy example. Let $p = 3$ and $\beta^0 = (t_0, t_0, 0)^T$, where $t_0 > 0$. Assume that the diagonal and off-diagonal elements of the covariance matrix $\mathbf{\Sigma}$ are 1 and ρ, respectively. Furthermore, assume that the costs of the variables and the budget C satisfy $0 < c_1, c_2, c_3 \le C$, $c_1 + c_2 > C$, $c_1 + c_3 > C$, and $c_2 + c_3 > C$. If $\rho \in (-1/2, -1/3)$, we can check that $\beta^* = (0, 0, 2\rho t_0)^T$. Therefore, we have $\{j : \beta_j^0 \neq 0\} = \{1, 2\}$, $\{j : \beta_j^* \neq 0\} = \{3\}$, and $\{j : \beta_j^0 \neq 0\} \cap \{j : \beta_j^* \neq 0\}$ is an empty set.

3. Motivation and Methodology

In this section, we first provide the motivation of our method using the orthogonal design case in Section 3.1. Then, we introduce our proposed method for the general high-dimensional cost-constrained regression problem (4) in Section 3.2. The extensions of our proposed method to some other statistical learning problems and problems with groups of variables will be discussed in Section 4.

3.1. Orthogonal Design Case

For motivation, we first assume that $n > p$ *and* $\mathbf{X}^T\mathbf{X} = n\mathbf{I}_n$. We can show that

$$\begin{aligned}\frac{1}{2n}||Y - \mathbf{X}\beta||_2^2 &= \frac{1}{2n}||Y - \mathbf{X}\hat{\beta}||_2^2 + \frac{1}{2n}||\mathbf{X}\beta - \mathbf{X}\hat{\beta}||_2^2 \\ &= \frac{1}{2n}||Y - \mathbf{X}\hat{\beta}||_2^2 + \frac{1}{2}||\beta - \hat{\beta}||_2^2,\end{aligned}$$

where $\hat{\beta} = \frac{\mathbf{X}^T Y}{n}$ is the least-square estimate of the true regression coefficient β^0. Therefore, in this case, problem (4) is equivalent to the following optimization problem

$$\min_{\beta} ||\beta - \hat{\beta}||_2^2 \quad \text{subject to} \sum_{j=1}^{p} c_j \mathcal{I}(\beta_j) \le C. \tag{5}$$

As shown in the following Theorem 1, problem (5) is equivalent to a 0-1 knapsack problem.

Theorem 1. If $\hat{\beta}$ is an optimal solution to the following problem:

$$\min_{\beta} ||\beta - a||_2^2 \quad \text{subject to} \sum_{j=1}^{p} c_j \mathcal{I}(\beta_j) \le C, \tag{6}$$

then $\hat{\beta} = a \circ \hat{Z}$ *where* $\circ$ denotes the entrywise product of two vectors, and $\hat{Z} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_p)$ is the solution to the following 0-1 knapsack problem:

$$\max_{z_1, z_2, \ldots, z_p} \sum_{j=1}^{p} a_j^2 z_j \quad \text{subject to} \sum_{j=1}^{p} c_j z_j \le C, \text{ and } z_1, z_2, \ldots, z_p \in \{0, 1\}. \tag{7}$$

The 0-1 knapsack problem is a famous problem in combinatorial optimization, and has been extensively studied (see, e.g., chapter 8 in the book of Paschos (2013) and the references therein) in the operations research community. It is the problem of choosing a subset of p items such that the corresponding profit sum is maximized without having the weight sum to exceed a prespecified capacity C. Although this problem is NP-hard, algorithmic advances and hardware improvements enable us to solve it efficiently. For example, dynamic programming (Bellman 1966) is a popular algorithm to exactly solve the 0-1 knapsack problem. Theoretically, it is a pseudo-polynomial time algorithm and the complexity is $O(pC)$. Numerically, as shown in Martello, Pisinger, and Toth (1999), the hybrid algorithm combining dynamic programming and the branch-and-bound approach (Nauss 1976) is able to solve many test datasets, with up to 10,000 variables, in less than 0.2 s.

In this article, we choose the dynamic programming algorithm to solve the 0-1 knapsack problem. Consider problem (7) as an example. Denote $D[j, w]$ be the maximum value that can be attained with weight less than or equal to w using items up to j. We can define $D[j, w]$ recursively as follows:

- $D[0, w] = 0$;
- $D[j, w] = D[j-1, w]$ *if* $c_j > w$;
- $D[j, w] = \max(D[j-1, w], D[j-1, w-c_j] + a_j^2)$ *if* $c_j \le w$.

The solution can then be found by calculating $D[p, C]$. Throughout this article, we assume that the costs $c_1, c_2, \ldots, c_p$ and the budget C are nonnegative integers. If they are not integers, we can use a scaling method that multiplies the noninteger costs and the budget by the same factor so that all costs and the budget are integers.

For the orthogonal design case, as shown in Theorem 1, we can solve the SAA problem (4) by solving its corresponding 0-1 knapsack problem. This theoretical result is important for us to design the algorithm for the general case in Section 3.2 where the predictors are correlated.

3.2. General Case

To solve the general high-dimensional cost-constrained regression problem (4), similar to the method for the best subset selection problem (Bertsimas et al. 2016), we use projected gradient descent methods for the first-order convex optimization problems (Nesterov 2013). Denote

$$g(\beta) = \frac{1}{2n}||Y - \mathbf{X}\beta||_2^2 \quad \text{and}$$
$$\nabla g(\beta) = \frac{\partial g(\beta)}{\partial \beta} = -\frac{1}{n}\mathbf{X}^T(Y - \mathbf{X}\beta).$$

For any $\alpha, \beta \in \mathbb{R}^p$, and $L \geq \ell = \lambda_{\max}(\frac{\mathbf{X}^T\mathbf{X}}{n})$, we can check that

$$||\nabla g(\alpha) - \nabla g(\beta)||_2 = ||\frac{1}{n}\mathbf{X}^T\mathbf{X}(\alpha - \beta)||_2 \leq \lambda_{\max}(\frac{\mathbf{X}^T\mathbf{X}}{n})||\alpha - \beta||_2 \leq L||\alpha - \beta||_2.$$

In addition, for any $\eta, \beta \in \mathbb{R}^p$, and $L \geq \ell = \lambda_{\max}(\frac{\mathbf{X}^T\mathbf{X}}{n})$, denote $Q_L(\eta, \beta) = g(\beta) + \frac{L}{2}||\eta - \beta||_2^2 + \langle \nabla g(\beta), \eta - \beta \rangle$. We can check that $Q_L(\eta, \beta) - g(\eta) \geq \frac{L-\ell}{2}||\beta - \eta||_2^2 \geq 0$ and thus

$$g(\eta) \leq Q_L(\eta, \beta) = g(\beta) + \frac{L}{2}||\eta - \beta||_2^2 + \langle \nabla g(\beta), \eta - \beta \rangle \tag{8}$$

for all β, η with equality holding at $\beta = \eta$. Therefore, given a current approximate solution $\beta^{(m)}$ to the problem (4), we can upper bound the function $g(\eta)$ by the function $Q_L(\eta, \beta^{(m)})$, and update the solution by

$$\beta^{(m+1)} \in \arg\min_{\eta} Q_L(\eta, \beta^{(m)}) \quad \text{subject to} \quad \sum_{j=1}^{p} c_j \mathcal{I}(\eta_j) \leq C,$$

which is also a global minimizer of the following problem:

$$\min_{\eta} ||\eta - (\beta^{(m)} - \frac{1}{L}\nabla g(\beta^{(m)}))||_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p} c_j \mathcal{I}(\eta_j) \leq C. \tag{9}$$

According to Theorem 1, we can solve problem (9) efficiently by finding the solution to its corresponding 0-1 knapsack problem. Therefore, we propose the following high-dimensional cost-constrained regression (HCR) method to estimate the parameter of interest β^*.

High-Dimensional Cost-Constrained Regression (HCR)

Step 1: Choose $\delta > 0$, $L > \ell = \lambda_{\max}(\frac{1}{n}\mathbf{X}^T\mathbf{X})$, and initialize $\beta^{(1)}$ such that $\sum_{j=1}^{p} c_j \mathcal{I}(\beta_j^{(1)}) \leq C$.

Step 2: For $m \geq 1$, denote $\mu^{(m)} = \beta^{(m)} + \frac{1}{nL}\mathbf{X}^T(Y - \mathbf{X}\beta^{(m)})$, apply dynamic programming to find

$$\beta^{(m+1)} = \mu^{(m)} \circ Z^{(m)} = (\mu_1^{(m)} z_1^{(m)}, \mu_2^{(m)} z_2^{(m)}, \ldots, \mu_p^{(m)} z_p^{(m)}),$$

where

$$Z^{(m)} \in \arg\max_{z_1, z_2, \ldots, z_p} \sum_{j=1}^{p} (\mu_j^{(m)})^2 z_j \quad \text{subject to} \quad \sum_{j=1}^{p} c_j z_j \leq C.$$

Step 3: Repeat **Step 2**, until $g(\beta^{(m)}) - g(\beta^{(m+1)}) \leq \delta$.

In the HCR method, we need to choose the initial value $\beta^{(1)}$ such that $\sum_{j=1}^{p} c_j \mathcal{I}(\beta_j^{(1)}) \leq C$. We can choose $\beta^{(1)} = 0$. A better choice can be the LASSO estimate which satisfies the budget constraint and delivers the lowest mean cross-validated error. Our simulation studies in Section 6 show that this choice obtains good performance. Since the nonconvex optimization problem (4) may have some local minimiziers, it is worthwhile starting the algorithm with multiple different choices of $\beta^{(1)}$, and choosing the solution with the smallest value of the objective function. To solve the 0-1 knapsack problem, we can use the dynamic programming algorithm which is available in the `adagio` R package. We can also use some other methods such as the efficient hybrid algorithm (Martello, Pisinger, and Toth 1999). For each iteration in our algorithm, we need to calculate $\mu^{(m)} = \beta^{(m)} + \frac{1}{nL}\mathbf{X}^T(Y - \mathbf{X}\beta^{(m)})$ and use the dynamic programming to find the exact solution to the 0-1 knapsack problem. Since the dynamic programming algorithm for the 0-1 knapsack problem costs $O(pC)$ operations, each iteration in our algorithm costs $O(p(n + C))$ operations in total.

4. Extensions

In this section, we consider several extensions of our proposed HCR method to some general statistical learning problems (e.g., the cost-constrained logistic regression problem) and problems with groups of variables.

4.1. Convex Differential Loss With Lipschitz Continuous Gradient

Consider a general statistical learning problem where the statistical model links the predictors $x_1, x_2, \ldots, x_p$ to the response variable y via a linear function $f = \sum_{j=1}^{p} x_j\beta_j$. Let $\psi(y, f)$ be the loss function used to fit the model. In this general setting, considering the budget constraint, we are interested in finding the cost-constrained model with $f = \sum_{j=1}^{p} x_j\beta_j^*$, where

$$\beta^* \in \arg\min_{\beta} \mathrm{E}[\psi(y, \sum_{j=1}^{p} x_j\beta_j)] \quad \text{subject to} \quad \sum_{j=1}^{p} c_j \mathcal{I}(\beta_j) \leq C.$$

Denote $g(\beta) = \frac{1}{n}\sum_{i=1}^{n} \psi(y_i, \sum_{j=1}^{p} \mathbf{X}_{ij}\beta_j)$. To estimate β^*, we need to solve

$$\min_{\beta} g(\beta) \quad \text{subject to} \quad \sum_{j=1}^{p} c_j \mathcal{I}(\beta_j) \leq C. \tag{10}$$

Suppose that the gradient of the convex differential loss function $\psi(y, f)$ satisfies the following Lipschitz condition:

$$|\frac{\partial \psi}{\partial f}(y, f_1) - \frac{\partial \psi}{\partial f}(y, f_2)| \leq M_1|f_1 - f_2|, \tag{11}$$

for any y, f_1, f_2, and a positive constant M_1. According to Lemma 1 in Yang and Zou (2015), if $L \geq 2M_1 \cdot \lambda_{\max}(\frac{\mathbf{X}^T\mathbf{X}}{n})$, we can show that

$$g(\eta) \leq Q_L(\eta, \beta) = g(\beta) + \frac{L}{2}||\eta - \beta||_2^2 + \langle \nabla g(\beta), \eta - \beta \rangle, \tag{12}$$

for all β, η with equality holding at $\beta = \eta$.

In addition, if

$$\frac{\partial \psi^2(y, f)}{\partial f^2} \text{ exists and } \frac{\partial \psi^2(y, f)}{\partial f^2} \leq M_2 \text{ for any } y \text{ and } f, \tag{13}$$

and $L \geq M_2 \cdot \lambda_{\max}(\frac{\mathbf{X}^T\mathbf{X}}{n})$, we can also show that

$$g(\eta) \leq Q_L(\eta,\beta) = g(\beta) + \frac{L}{2}||\eta-\beta||_2^2 + \langle \nabla g(\beta), \eta-\beta\rangle, \tag{14}$$

for all β, η with equality holding at $\beta = \eta$.

In Section 3, for the cost-constrained regression problem, we choose the quadratic least-square loss $\psi(y,f) = \frac{1}{2}(y-f)^2$. We can check that it satisfies the condition (13) with the constant $M_2 = 1$. Next, we show some other loss functions satisfying the condition (11) or (13). We use some loss functions for the binary classification problem as examples. For the binary classification problem, the response variable y is a class label. We code y *by* $\{+1,-1\}$.

The first example is the logistic regression loss defined by $\psi(y,f) = \log(1+\exp(-yf))$. We can show that

$$\frac{\partial\psi^2(y,f)}{\partial f^2} = \frac{e^{yf}}{(1+e^{yf})^2}, \text{ and therefore } \frac{\partial\psi^2(y,f)}{\partial f^2} \leq 1/4.$$

The second example is the squared hinge loss $\psi(y,f) = [(1-yf)_+]^2$, where $(1-t)_+ = 0$ if $t > 1$ and 0 otherwise. As shown in Yang and Zou (2015), we can verify that

$$\frac{\partial\psi}{\partial f}(y,f) = 0 \quad \text{if } yf > 1 \text{ and } -2y(1-yf) \text{ otherwise.}$$
$$|\frac{\partial\psi}{\partial f}(y,f_1) - \frac{\partial\psi}{\partial f}(y,f_2)| \leq 2|f_1-f_2|,$$

and therefore condition (11) holds. For all the convex differential loss functions satisfying the condition (11) or (13), using the same idea shown in Section 3, given a current approximate solution $\beta^{(m)}$ to the problem (10), we can upper bound the function $g(\eta)$ by the function $Q_L(\eta,\beta^{(m)})$, and update the solution by

$$\beta^{(m+1)} = \arg\min_{\eta} ||\eta - (\beta^{(m)} - \frac{1}{L}\nabla g(\beta^{(m)}))||_2^2$$
$$\text{subject to } \sum_{j=1}^{p} c_j\mathcal{I}(\eta_j) \leq C. \tag{15}$$

According to Theorem 1, we can solve problem (15) efficiently by finding the solution to its corresponding 0-1 knapsack problem.

In practice, the budget limit C may be relatively large while the costs $c_1, c_2, \ldots, c_p$ are small. Then, many feasible models may have k predictors where k can be close to or larger than the sample size n. In this case, we can combine our HCR method with regularization techniques to reduce overfitting. For example, if we use the elastic net penalty (Zou and Hastie 2005), we can estimate β^* by solving

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}\psi(y_i, \sum_{j=1}^{p}\mathbf{X}_{ij}\beta_j) + \sum_{j=1}^{p}\lambda(\alpha|\beta_j| + \frac{1-\alpha}{2}\beta_j^2)$$
$$\text{subject to } \sum_{j=1}^{p} c_j\mathcal{I}(\beta_j) \leq C.$$

To solve the above problem, similar to the method in Section 3, we only need to develop efficient algorithms to solve the following problem

$$\min_{\beta} \frac{1}{2}||\beta-a||_2^2 + \sum_{j=1}^{p}\lambda(\alpha|\beta_j| + \frac{1-\alpha}{2}\beta_j^2)$$
$$\text{subject to } \sum_{j=1}^{p} c_j\mathcal{I}(\beta_j) \leq C. \tag{16}$$

As shown in the following Proposition 1, in order to solve problem (16), we also only need to solve a 0-1 knapsack problem.

Proposition 1. If $\hat{\beta}$ is an optimal solution to the problem (16), then $\hat{\beta} = \frac{1}{1+\lambda(1-\alpha)} \cdot \text{sign}(a-\alpha\lambda) \circ (|a|-\alpha\lambda)_+ \circ \hat{Z}$, where $\hat{Z} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_p)$ is the solution to the following 0-1 knapsack problem

$$\max_{z_1,z_2,\ldots,z_p} \sum_{j=1}^{p} \frac{a_j^2 - 2\alpha\lambda|a_j| + \alpha^2\lambda^2}{2(1+\lambda(1-\alpha))} \cdot \frac{1+\text{sign}(|a_j|-\alpha\lambda)}{2} \cdot z_j$$
$$\text{subject to } \sum_{j=1}^{p} c_j z_j \leq C, \text{ and } z_1, z_2, \ldots, z_p \in \{0,1\}.$$

As shown in Proposition 1, both the shrinkage and the budget constraint can result in a sparse estimate of β^*. For our proposed HCR method with regularization, we can use cross-validation to choose the tuning parameters such as α *and* λ in the elastic net penalty. The budget limit C is assumed to be prespecified.

4.2. Groups of Variables

In some data collection processes, variables are collected group-by-group. As shown in the diabetes study in Section 1, biomarkers are collected from different lab tests. From a blood test, we get the values of four biomarkers (HDL, LDL, Total cholesterol, and Triglycerides) simultaneously. We need to spend \$200 if we need to collect the value of any of these four biomarkers to predict the treatment response. In this case, the variables are divided into different groups naturally according to the data collection process. It is more reasonable to consider the group costs of different lab tests rather than the separate cost of each biomarker.

Suppose we need to spend $\tilde{c}_g$ dollars to collect the values of all p_g variables in the gth group $\mathcal{A}_g$. If the loss function $\psi(y,f)$ is used, we are interested in finding the cost-constrained model with $f = \sum_{j=1}^{p} x_j\beta_j^*$ and

$$\beta^* \in \arg\min_{\beta} \mathrm{E}[\psi(y, \sum_{j=1}^{p} x_j\beta_j)]$$
$$\text{subject to } \sum_{g=1}^{G} \tilde{c}_g[1 - \prod_{j\in\mathcal{A}_g}(1-\mathcal{I}(\beta_j))] \leq C,$$

where we assume that we always need to spend $\tilde{c}_g$ dollars if there is at least one variable in the gth group $\mathcal{A}_g$ with a nonzero regression coefficient. Given the training data $\{Y, \mathbf{X}\}$, we estimate β^*

by solving the following SAA problem:

$$\min_{\beta} g(\beta) \quad \text{subject to} \sum_{g=1}^{G} \tilde{c}_g[1 - \prod_{j\in\mathcal{A}_g}(1-\mathcal{I}(\beta_j))] \le C. \quad (17)$$

Similar to our discussion in Section 4.1, if the loss function $\psi(y,f)$ is convex, differentiable, and satisfies the condition (11) or (13), given a current approximate solution $\beta^{(m)}$ to the problem (17), we can upper bound the function $g(\eta)$ by the function $Q_L(\eta,\beta^{(m)})$, and update the solution by

$$\beta^{(m+1)} \in \arg\min_{\eta} ||\eta - (\beta^{(m)} - \frac{1}{L}\nabla g(\beta^{(m)}))||_2^2$$
$$\text{subject to} \sum_{g=1}^{G} \tilde{c}_g[1 - \prod_{j\in\mathcal{A}_g}(1-\mathcal{I}(\beta_j))] \le C. \quad (18)$$

The following Proposition 2 shows how to solve problem (18).

Proposition 2. If $\hat{\beta}$ is an optimal solution to the following problem

$$\min_{\beta} ||\beta - a||_2^2 \quad \text{subject to} \sum_{g=1}^{G} \tilde{c}_g[1 - \prod_{j\in\mathcal{A}_g}(1-\mathcal{I}(\beta_j))] \le C,$$

then $\hat{\beta} = a \circ \hat{Z}$, where $\hat{Z} = (\hat{z}_1\mathbf{1}_{p_1}, \hat{z}_2\mathbf{1}_{p_2}, \ldots, \hat{z}_g\mathbf{1}_{p_g})^T$, $\mathbf{1}_{p_g}$ is a row vector of p_g 1's, and $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_g$ is the solution to the following 0-1 knapsack problem:

$$\max_{z_1,z_2,\ldots,z_g} \sum_{g=1}^{G}(\sum_{j\in\mathcal{A}_g} a_j^2)z_g \quad \text{subject to} \sum_{g=1}^{G} \tilde{c}_g z_g \le C, \text{ and}$$
$$z_1, z_2, \ldots, z_g \in \{0,1\}.$$

The proof of Proposition 2 is very similar to the proof of Theorem 1. In this case, the number of groups G is often much smaller than the dimension p, and we can use dynamic programming to solve the 0-1 knapsack problem efficiently. For the problem with groups of variables, we can also combine our HCR method with regularization techniques. We estimate β^* by solving

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}\psi(y_i, \sum_{j=1}^{p}\mathbf{X}_{ij}\beta_j) + \sum_{j=1}^{p}\lambda(\alpha|\beta_j| + \frac{1-\alpha}{2}\beta_j^2)$$
$$\text{subject to} \sum_{g=1}^{G} \tilde{c}_g[1 - \prod_{j\in\mathcal{A}_g}(1-\mathcal{I}(\beta_j))] \le C. \quad (19)$$

The following Proposition 3 can be used to solve the above optimization problem.

Proposition 3. If $\hat{\beta}$ is an optimal solution to the following problem:

$$\min_{\beta} \frac{1}{2}||\beta - a||_2^2 + \sum_{j=1}^{p}\lambda(\alpha|\beta_j| + \frac{1-\alpha}{2}\beta_j^2) \quad \text{subject to}$$
$$\sum_{g=1}^{G} \tilde{c}_g[1 - \prod_{j\in\mathcal{A}_g}(1-\mathcal{I}(\beta_j))] \le C,$$

then $\hat{\beta} = \frac{1}{1+\lambda(1-\alpha)} \cdot \text{sign}(a - \alpha\lambda) \circ (|a| - \alpha\lambda)_+ \circ \hat{Z}$, where $\hat{Z} = (\hat{z}_1\mathbf{1}_{p_1}, \hat{z}_2\mathbf{1}_{p_2}, \ldots, \hat{z}_g\mathbf{1}_{p_g})^T$ and $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_g$ is the solution to the following 0-1 knapsack problem

$$\max_{z_1,z_2,\ldots,z_g} \sum_{g=1}^{G}\left(\sum_{j\in\mathcal{A}_g} \frac{a_j^2 - 2\alpha\lambda|a_j| + \alpha^2\lambda^2}{2(1+\lambda(1-\alpha))} \cdot \frac{1+\text{sign}(|a_j|-\alpha\lambda)}{2}\right) z_g$$
$$\text{subject to} \sum_{g=1}^{G} \tilde{c}_g z_g \le C, \text{ and } z_1, z_2, \ldots, z_g \in \{0,1\}.$$

The proof of Proposition 3 is very similar to the proof of Proposition 1. We omit the details of the proof.

5. Theoretical Properties

In this section, we consider the problem (10) with a general convex differential loss function $\psi(y,f)$ satisfying the condition (11) or (13), and the budget constraint $\sum_{j=1}^{p} c_j\mathcal{I}(\beta_j) \le C$. Denote $\ell = 2M_1\lambda_{\max}(\frac{1}{n}\mathbf{X}^T\mathbf{X})$ if $\psi(y,f)$ satisfies the condition (11), and $M_2\lambda_{\max}(\frac{1}{n}\mathbf{X}^T\mathbf{X})$ if $\psi(y,f)$ satisfies the condition (13). Since we choose to use the dynamic programming algorithm to solve the 0-1 knapsack problem in the HCR algorithm, we assume that all costs c_j's and the budget C are integers. If we have some noninteger costs or budget, we can use the scaling method to scale c_j's and C first and then use the dynamic programming algorithm. We generalize some theoretical results shown in Bertsimas et al. (2016) about the best subset selection problem. The theoretical results of the methods for the other extensions shown in Section 4 can be derived similarly.

We first show the asymptotic convergence of our proposed method.

Theorem 2. Assume that all costs c_j's and the budget C are integers. For any $L \ge \ell$, the sequence $g(\beta^{(m)}) = \frac{1}{n}\sum_{i=1}^{n}\psi(y_i, \sum_{j=1}^{p}\mathbf{X}_{ij}\beta_j^{(m)})$ is decreasing, converges and satisfies $g(\beta^{(m)}) - g(\beta^{(m+1)}) \ge \frac{L-\ell}{2}||\beta^{(m+1)} - \beta^{(m)}||_2^2$. Furthermore, if $L > \ell$, then $\beta^{(m+1)} - \beta^{(m)} \to 0$ as $m \to \infty$.

According to Theorem 2, given a small positive number δ in our HCR method, our algorithm is guaranteed to converge in finite steps. Next, we introduce the first-order stationary point of the high-dimensional cost-constrained regression problem, which can be considered as a near optimal solution. Then we show that the sequence generated by our algorithm, $\{\beta^{(m)}\}$, converges to a first-order stationary point.

Definition 1. Given a positive constant $L \ge \ell$, a vector $\eta \in \mathbb{R}^p$ is said to be a first-order stationary point of the high-dimensional cost-constrained regression problem (10) if it satisfies the following condition:

$$\eta \in \arg\min_{\beta} ||\beta - (\eta - \frac{1}{L}\nabla g(\eta))||_2^2 \quad \text{subject to}$$
$$\sum_{j=1}^{p} c_j\mathcal{I}(\beta_j) \le C.$$

As shown in our following Theorem 3, a global minimizer of the problem (10) is a first-order stationary point.

Theorem 3. Let $L > \ell$. If $\hat{\beta}$ is a global minimizer of the problem (10), then it is a first-order stationary point.

As the problem (10) is a nonconvex optimization problem, it is generally difficult to find a global minimizer. According to the above Theorem 3, a global minimizer is also a first-order stationary point. We can develop some efficient algorithms to find a first-order stationary point which can be a global minimizer of the problem (10).

Theorem 4. If η is a first-order stationary point and

$$\max_{j\in A} c_j + \sum_{j\in A^c} c_j \le C, \tag{20}$$

where $A = \{j : \eta_j = 0\}$, then η is a global minimizer of the problem (10).

The condition in Theorem 4 can be used to check whether a first-order stationary point is a global minimizer. By Theorems 3 and 4, we can view the first-order stationary point as a near optimal solution of the problem (10). In the following Theorem 5, we will show that the sequence $\{\beta^{(m)}\}$ generated by our method converges to a first-order stationary point which can be a global minimizer of the problem (10).

Theorem 5. Assume that all costs c_j's and the budget C are integers. If $L > \ell$ and

$$\lim\inf_{m\to\infty} \min\{|\beta_j^{(m)}| : |\beta_j^{(m)}| > 0\} > 0, \tag{21}$$

then the sparsity pattern sequence $\{Z^{(m)}\}$ converges after finitely many steps, that is, there exists an iteration index M^* such that $Z^{(m)} = Z^{(m+1)}$ for all $m \ge M^*$. Furthermore, the sequence $\{\beta^{(m)}\}$ is bounded and converges to a first-order stationary point.

Note that Bertsimas et al. (2016) used a condition similar to the condition (21) for the best subset selection problem. This condition implies that the support of $\beta^{(m)}$ stabilizes after several iterations and our proposed HCR method behaves like a gradient descent algorithm thereafter. Theorem 5 shows that the sequence $\{\beta^{(m)}\}$ converges to a first-order stationary point, which is only a near global minimizer. We can use the sufficient condition (20) to check whether this first-order stationary point is a global minimizer. Besides the case where the condition (20) is satisfied, we can show that the sequence $\{\beta^{(m)}\}$ converges to a global minimizer when some other conditions on the function $g(\beta)$, the nonconvex set $\mathcal{F} = \{\beta : \sum_{j=1}^p c_j\mathcal{I}(\beta_j) \le C\}$, and the initial estimate $\beta^{(1)}$ are satisfied.

Let $\hat{\beta}$ be a global minimizer of problem (10). Suppose that the convex differentiable loss function $\psi(y,f)$ satisfies condition (11) or (13). We have shown that the function $g(\beta) = \frac{1}{n}\sum_{i=1}^n \psi(y_i, \sum_{j=1}^p x_{ij}\beta_j)$ satisfies

$$g(\eta) \le g(\beta) + \frac{L}{2}||\eta-\beta||_2^2 + \langle \nabla g(\beta), \eta-\beta\rangle,$$

for all β, η and $L \ge \ell$. Denote $\mathcal{B} = \{\beta : ||\beta - \hat{\beta}||_2 \le ||\beta^{(1)} - \hat{\beta}||_2\}$, where $\beta^{(1)}$ is the initial value used in our algorithm. In the following Theorem 6, we further assume that the function $g(\beta)$ satisfies the restricted strong convexity condition (Barber and Ha (2018))

$$g(\eta) \ge g(\beta) + \frac{\alpha}{2}||\eta-\beta||_2^2 + \langle \nabla g(\beta), \eta-\beta\rangle, \tag{22}$$

for all $\beta, \eta \in \mathcal{B}$, where $\alpha \in (0, L)$ is a constant. In addition, we assume the nonconvex set $\mathcal{F} = \{\beta : \sum_{j=1}^p c_j\mathcal{I}(\beta_j) \le C\}$ satisfies

$$\max_{\beta,\eta\in\mathcal{F}\cap\mathcal{B}} \gamma_\beta(\mathcal{F}) \cdot ||\nabla g(\eta)||_2 \le \frac{(1-t_0)\cdot\alpha}{2}, \tag{23}$$

where $t_0 \in (0,1)$ is a constant and $\gamma_\beta(\mathcal{F})$ is the local concavity coefficient measuring the concavity in the feasible set $\mathcal{F}$ relative to the point β (Barber and Ha (2018)). Using Theorem 3 in Barber and Ha (2018), we can prove the following Theorem 6.

Theorem 6. Assume that all costs c_j's and the budget C are integers. Suppose that the convex differentiable loss function $\psi(y,f)$ satisfies condition (11) or (13). Furthermore, assume that conditions (22) and (23) hold. Then, we have

$$||\beta^{(m+1)} - \hat{\beta}||_2^2 \le (1 - \frac{2\alpha t_0}{L+\alpha})^m \cdot ||\beta^{(1)} - \hat{\beta}||_2^2.$$

The above result indicates that the sequence $\{\beta^{(m)}\}$ generated by our proposed algorithm can converge linearly to the global minimizer $\hat{\beta}$. If we further assume that $\hat{\beta}$ is the unique global minimizer of the problem (10), according to the existing theoretical result on the consistency of SAA estimators (e.g., (Shapiro, Dentcheva, and Ruszczyński 2014, Theorem 5.3)), we can show that $\hat{\beta}$ is a consistent estimator of β^* under some regularity conditions. By combining the consistency of SAA estimators and the result shown in Theorem 6, we can conclude that $||\beta^{(m+1)} - \beta^*||_2 \le ||\beta^{(m+1)} - \hat{\beta}||_2 + ||\hat{\beta} - \beta^*||_2$, where the error $||\beta^{(m+1)} - \hat{\beta}||_2$ converges to 0 as $m \to \infty$ and the error $||\hat{\beta} - \beta^*||_2$ converges to 0 in probability as $n \to \infty$.

6. Simulation Study

In this section, we demonstrate the effectiveness of our proposed HCR method using some simulated examples. Since the existing methods (Fouskakis and Draper 2008; Fouskakis, Ntzoufras, and Draper 2009; Yue 2010) that trades prediction accuracy against data collection cost requires expensive computation for the high-dimensional examples used in this study, we only compare our proposed method with the LASSO method and two weighted LASSO methods. The first weighted LASSO method (WLASSO1) is the adaptive LASSO regression (Zou 2006) using the costs of predictors c_j's as the weights. The second weighted LASSO method (WLASSO2) is the adaptive LASSO regression using $c_j/|\hat{\beta}_j^{\text{initial}}|$'s as the weights, where $\hat{\beta}_j^{\text{initial}}$ is the ridge regression estimate which obtains the lowest mean cross-validated error.

The LASSO method and the two weighted LASSO methods do not handle the budget constraint directly. For all these three methods, in the tuning process, we use the cross-validation method to choose the estimate in the regularization path

which satisfies the budget constraint and obtains the lowest mean cross-validated error. The glmnet R package (Friedman, Hastie, and Tibshirani 2010) is used for these three methods. For our HCR method, we choose $\delta = 10^{-4}$, and $\beta^{(1)}$ to be the LASSO estimate which satisfies the budget constraint and delivers the lowest mean cross-validated error. We choose L to be $\lambda_{\max}(\frac{1}{n}\mathbf{X}^T\mathbf{X}) + 0.1$ for the four regression examples (Examples 1–4) with the quadratic least-squares loss and $0.25\lambda_{\max}(\frac{1}{n}\mathbf{X}^T\mathbf{X}) + 0.1$ for the two classification examples (Examples 5 and 6) with the logistic regression loss. For each experiment, we run the iterations in the HCR algorithm at most 1000 times. For each example, we repeat the experiment 100 times. For all six examples, the dimension p is 1000, the sample size of the training dataset is 200, and the sample size of the testing dataset is 10,000.

6.1. Examples

We study four regression examples and two classification examples. For the regression examples, the response variables are generated from the linear model shown in Equation (1). For the classification examples, the class labels are generated from logistic regression models.

Example 1. The predictors $(x_1, x_2, \ldots, x_p)^T \sim N(0, \mathbf{I}_p)$. The first 10 elements of β^0 are generated from $N(4, 1)$ independently, and the other $p - 10$ elements of β^0 are 0. The budget $C = 12$ and the variance $\sigma^2 = 0.25$. For each $j \in \{1, 2, \ldots, p\}$, we choose c_j, the cost of collecting the jth variable, randomly from the set {1, 2, 3}, and then use those c_j's for all the 100 experiments. In this example, β^* is different from β^0 since the simulated costs satisfy $\sum_{j=1}^{10} c_j = 20 > C$. As $\mathbf{\Sigma} = \mathbf{I}_p$, we can find the true parameter of interest β^* by solving the corresponding 0-1 knapsack problem of the problem (3).

Example 2. In this example, we consider the case where the true important variables (the first 10 variables) and the unimportant variables (the other $p - 10$ variables) are uncorrelated. The predictors $(x_1, x_2, \ldots, x_{10})^T \sim N(0, \mathbf{A})$, where $a_{jt} = 0.2$ if $j \neq t$ and 1 otherwise. The other $p - 10$ predictors are generated from $N(0, \mathbf{B})$, where $b_{jt} = 0.2$ if $j \neq t$ and 1 otherwise. The first 10 elements of β^0 are generated from $N(4, 1)$ independently, and the other $p - 10$ elements of β^0 are 0. The budget $C = 12$ and the variance $\sigma^2 = 0.25$. For each $j \in \{1, 2, \ldots, p\}$, we choose c_j randomly from the set {1, 2, 3}, and then use those c_j's for all the 100 experiments. In this case, since $\beta_j^* = 0$ for $j > 10$, we use the enumeration method to find the true β_j^* for $j \leq 10$. As the simulated costs in this example satisfy $\sum_{j=1}^{10} c_j = 20 > C$, β^* is different from β^0.

Example 3. The predictors $(x_1, x_2, \ldots, x_p)^T \sim N(0, \mathbf{\Sigma})$, where $\sigma_{jt} = 0.5^{|j-t|}$. The first 10 elements of β^0 are generated from $N(4, 1)$ independently, and the other $p - 10$ elements of β^0 are 0. The budget $C = 100$ and the variance $\sigma^2 = 0.25$. For each $j \in \{1, 2, \ldots, p\}$, we choose c_j to be 10 for all the 100 experiments. Since all costs c_j's are equal, the cost-constrained regression problem in this example is equivalent to the best subset selection problem with the constraint $\sum_{j=1}^{p} \mathcal{I}(\beta_j) \leq C/c_j = 10$. In addition, as C is equal to the cost of collecting all important variables, β^0 is a feasible solution and therefore the true parameter of interest β^* is the same as β^0.

Example 4. This example is the same as Example 3 except that for each $j \in \{1, 2, \ldots, p\}$, we choose c_j randomly from the set $\{1, 2, 3, \ldots, 50\}$, and then use those c_j's for all the 100 experiments. It requires expensive computation to find the exact true parameter of interest β^* by solving the NP-hard problem (3). For this example, we do not know the true parameter of interest β^*. However, since the simulated costs satisfy $\sum_{j=1}^{10} c_j = 283 > C$, we know that β^* is different from β^0.

Example 5. The predictors $(x_1, x_2, \ldots, x_p)$ and β^0 are generated by the same method showing in Example 2. For the i-th observation, the class label (+ 1 or –1) is generated by a binomial distribution. The probability of being 1 is $\exp(\sum_{j=1}^{p} \mathbf{X}_{ij}\beta_j^0)/(1 + \exp(\sum_{j=1}^{p} \mathbf{X}_{ij}\beta_j^0))$. The budget C is chosen to be 30. For each $j \in \{1, 2, \ldots, p\}$, we choose c_j randomly from the set $\{1, 2, 3, \ldots, 10\}$, and then use those c_j's for all the 100 experiments. Similar to Example 4, it requires expensive computation to find the exact true parameter of interest β^*. As the simulated costs in this example satisfy $\sum_{j=1}^{10} c_j = 45 > C$, β^* is different from β^0.

Example 6. The predictors $(x_1, x_2, \ldots, x_p)^T \sim N(0, \mathbf{\Sigma})$, where $\sigma_{jt} = 0.5^{|j-t|}$. The other settings are the same as Example 5.

To evaluate different methods, we use five measures: (i) estimation error: $||\hat{\beta} - \beta^*||_2$, (ii) prediction error: $||Y_{\text{test}} - \mathbf{X}_{\text{test}}\hat{\beta}||_2^2/n_{\text{test}}$ for regression examples or the misclassification error for classification examples, (iii) false-positive rate (FPR) $|\{j : \beta_j^* = 0 \text{ and } \hat{\beta}_j \neq 0\}|/|\{j : \beta_j^* = 0\}|$, (iv) false-negative rate (FNR) $|\{j : \beta_j^* \neq 0 \text{ and } \hat{\beta}_j = 0\}|/|\{j : \beta_j^* \neq 0\}|$, and (v) elapsed time of the R software to calculate $\hat{\beta}$ (the results about the elapsed time are shown in the supplementary materials). For Examples 4, 5, 6, since we cannot calculate the true parameter of interest β^*, we only compare the elapsed time and the prediction error of different methods.

6.2. Estimation and Prediction Performance

Figure 1 shows the estimation errors of HCR, LASSO, and the weighted LASSO methods. As shown in those three bar plots, compared with the other three methods, our proposed HCR method has the best estimation performance for all three examples. For the third example, as $\beta^* = \beta^0$, penalized regression methods generally have good performance. Both the LASSO method and the weighted LASSO methods deliver relatively low estimation errors. Our proposed HCR method still obtains the lowest estimation error. One possible reason is that HCR uses a nonconvex constraint rather than a convex penalty and therefore has a smaller estimation bias.

The comparison of the prediction performance between HCR and the other three methods is shown in Figure 2. Our proposed HCR method delivers better prediction performance

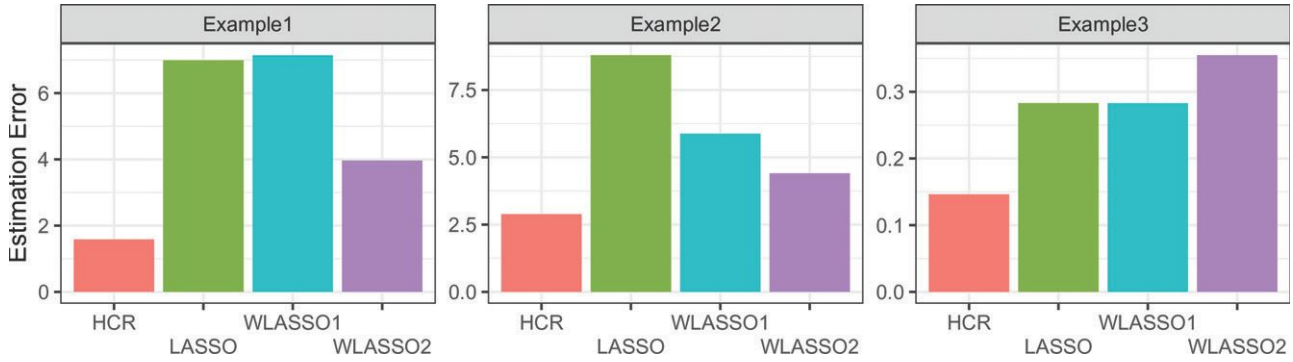

Figure 1. Estimation errors of HCR, LASSO, and the weighted LASSO methods.

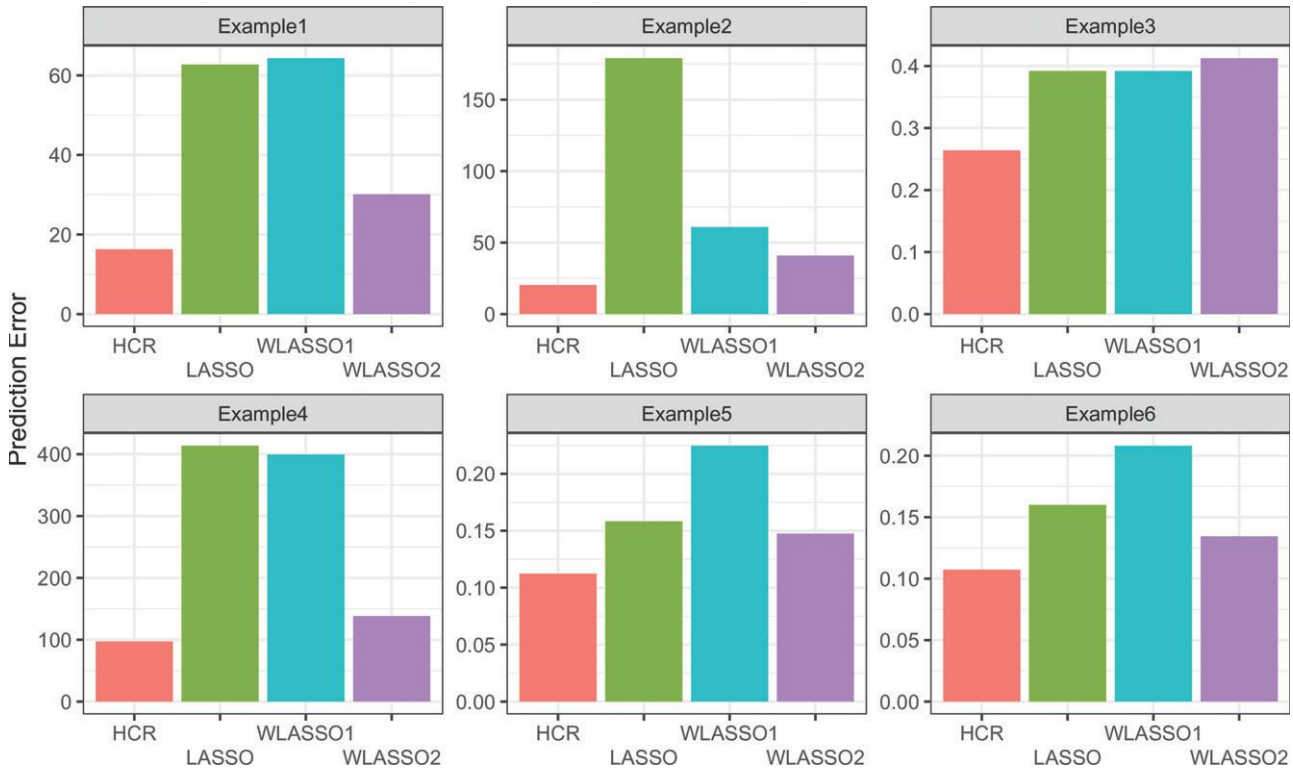

Figure 2. Prediction errors of HCR, LASSO, and the weighted LASSO methods.

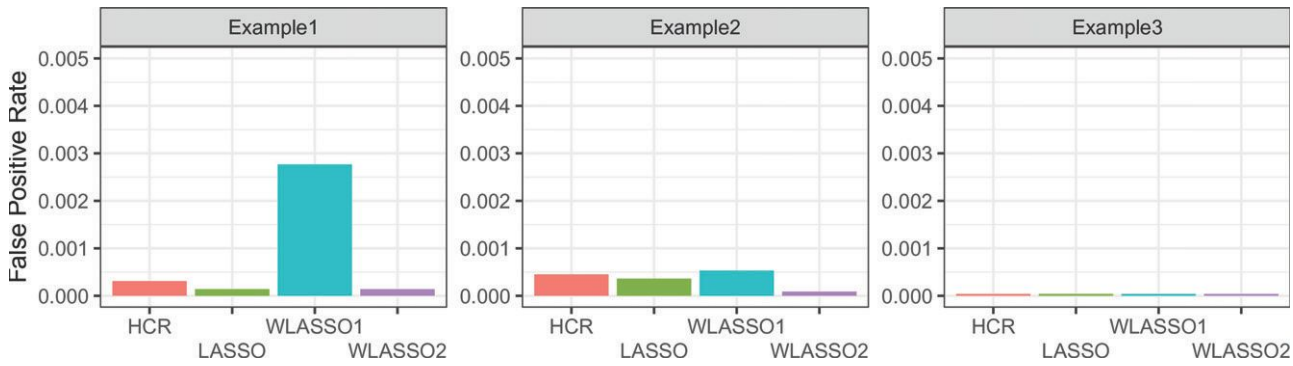

Figure 3. False positive rates of HCR, LASSO, and the weighted LASSO methods.

than the other three methods in all six examples. Compared with the LASSO method, the two weighted LASSO methods perform better in Examples 1, 2, and 4. However, they can perform worse than the LASSO methods as shown in the bar plots of Examples 5 and 6. For the cost-constrained linear regression problem, how to choose effective weights in the

weighted LASSO method considering both the costs and the initial estimate of the true regression coefficient vector is still an open question.

6.3. Model Selection Performance

We compare the model selection performance using the false-positive rate and false-negative rate in Figures 3 and 4, respectively. As shown in Figures 3, the false-positive rates of all four methods are close to 0. WLASSO1 delivers a slightly higher false-positive rate than the other three methods in Example 1. In terms of false negative rate, as shown in Figure 4, WLASSO2 delivers the best performance in Examples 1 and 2. For Example 1, our proposed method delivers a lower FNR than the LASSO and WLASSO1 methods. For Example 2, the false-negative rate of HCR is much lower than that of the LASSO while it is slightly higher than that of WLASSO1. For Example 3, the true parameter of interest β^* is the same as β^0. Both the LASSO and weighted LASSO methods are expected to have good model selection performance. Our proposed HCR method also obtains very good model selection performance on this example.

Overall, our simulation studies demonstrate the effectiveness of our proposed HCR method. Compared with the penalized regression techniques, our proposed HCR method could deliver a feasible model with better estimation, prediction, and model selection performance.

7. Application to a Diabetes Study

Diabetes mellitus, or simply diabetes, is a disease characterized by elevated blood glucose. It is a major cause of kidney failure, nontraumatic lower-limb amputations, blindness, heart disease and stroke. As a result, diabetes is one of the leading causes of death. The goal of treating diabetes patients is to lower their blood glucose. It is important to predict patient's treatment responses before assigning treatments so that we can select the treatment with the largest potential outcome for each patient. For this prediction problem, the best linear predictor may incorporate some biomarkers that are expensive to measure. To accurately predict treatment response while controlling the diagnostic cost is of practical importance. Our proposed HCR method is developed to address this important problem.

In this study, we use a dataset from a randomized, double-blind, parallel-group comparison phase III study that compares drug efficacy of gliclazide (control) versus pioglitazone (treatment). A total of 1270 patients with Type 2 diabetes were randomized in the phase III study with poorly controlled HbA1c (7.5%-11%). Patients were either received pioglitazone up to 45 mg once daily or gliclazide up to 160 mg two times a day. The primary efficacy endpoint was change in HbA1c from baseline to the end of the study (52 weeks). More details on this study design are shown in Charbonnel et al. (2005). In our analysis, we include 20 biomarkers measured at baseline. We delete patients if the value of change in HbA1c from baseline to the end of the

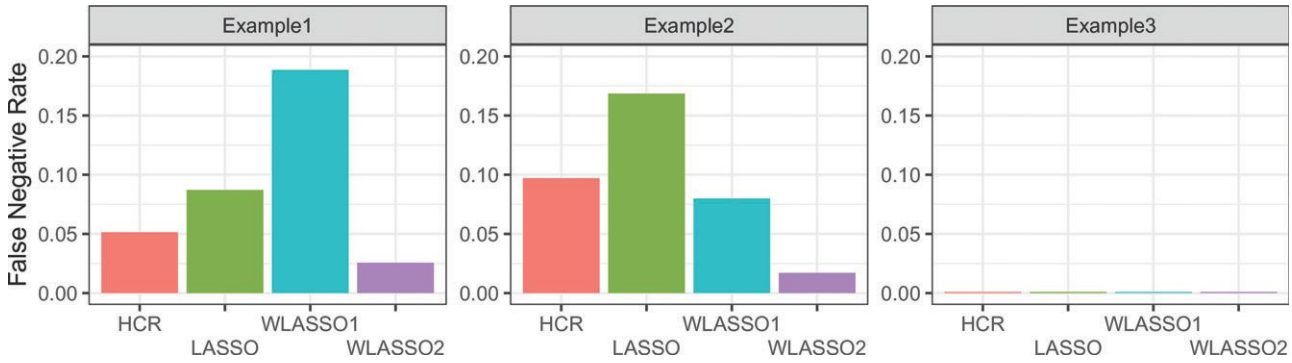

Figure 4. False negative rates of HCR, LASSO, and the weighted LASSO methods.

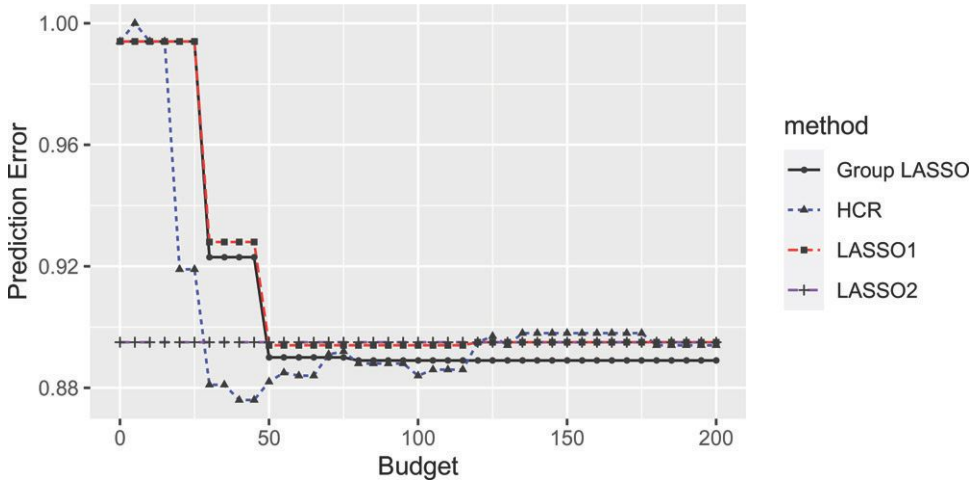

Figure 5. Prediction errors of different methods for the diabetes study. Note that the total cost of all the predictors used in the LASSO2 model selected by the 5-fold CV is $50. This LASSO2 model is infeasible when the budget is less than $50.

study or the value of some biomarker is missing. The biomarkers used in our study and their costs are shown in Table 1 in Section 1.

After the data preprocessing, our final dataset has 181 diabetes patients with one continuous response variable (the change in HbA1c from baseline to the end of the study) and 20 predictors. Considering the group structure of the predictors in this study, we used the HCR method (19) for the data with groups of variables to develop cost-constrained models for a sequence of budgets (from \$0 to \$200). We also used the LASSO penalty to reduce overfitting. We compared this HCR method with the LASSO and the Group LASSO (Yuan and Lin 2006) methods. For LASSO and Group LASSO, in order to choose the best tuning parameter λ, we used the cross-validation to choose the best tuning parameter that delivered the feasible solution (a solution that satisfies the budget constraint) with the lowest mean cross-validated error. In order to check the prediction performance of the linear model considering all predictors, we also used the LASSO method ignoring the budget constraint. We use LASSO1 and LASSO2 to denote the LASSO method considering the budget constraint and ignoring the constraint, respectively.

Before building cost-constrained models, we centered and standardized both the response variable and predictors. To compare the prediction errors of different methods, we used the 5-fold cross-validation (CV). For all methods, we used another inner 5-fold CV on the training data to choose the best tuning parameters. Figure 5 shows the prediction errors of different methods. The prediction errors of HCR, Group LASSO, and LASSO1 decrease as the budget increases. These three methods consider the budget constraint and therefore perform similarly when the budget is very small (e.g., less than \$20 in this real data analysis). The prediction error of LASSO2 is constant as it ignores the budget constraint. The total cost of all the predictors used in the LASSO2 model is \$50. We can expect that the prediction performance of all methods should be similar if the budget is more than \$50. Indeed, as shown in Figure 5, when the budget is larger than \$50, the prediction errors of all methods are close.

When the budget is less than \$50, the LASSO2 model is infeasible, and therefore it is not reasonable to compare the prediction performance of HCR, Group LASSO, and LASSO1 with that of LASSO2. As shown in Figure 5, when the budget is between \$20 and \$45, our proposed HCR method obtains significantly lower prediction errors than the Group LASSO and LASSO1 methods that also consider the budget constraint. Our numerical results also indicate that the prediction error of HCR can be even lower than that of LASSO2 in this case. One possible reason is that when the budget C is smaller than the cost of the best model fitted by LASSO2 and it is large enough to allow us to use some important predictors, both the budget constraint and the LASSO penalty play an important role in the modeling process. In that case, our HCR method can deliver a lower prediction error than LASSO2 by using both the budget constraint and the LASSO penalty to guide the modeling process. However, if C is larger than the cost of the best model fitted by LASSO2, the budget constraint will not affect the modeling process significantly and therefore our HCR method will be similar to the LASSO2 method.

8. Conclusion

In this article, in order to take into account the cost of data collection in the modeling process, we study a new high-dimensional cost-constrained regression problem. Although the nonconvex budget constraint makes this problem NP-hard, we propose a new discrete extension of the first-order continuous optimization methods to deliver a near optimal solution. Our HCR algorithm generates a series of estimates of the regression coefficient vector by solving a sequence of 0-1 knapsack problems that can be efficiently addressed by many existing algorithms such as dynamic programming. Our proposed HCR method can be extended to general statistical learning problems and problems with groups of variables. We can also combine our HCR method with regularization techniques to reduce overfitting. Theoretically, we show that the series of the estimates of the regression coefficient vector converge to a first-order stationary point, which is a near optimal solution. Our numerical study indicates that the proposed HCR method is computationally tractable to solve the nonconvex high-dimensional cost-constrained regression problem. It delivers promising estimation, prediction, and model selection performance.

Supplementary Materials

Supplementary file: All proofs and the comparison of the computational time of different methods are shown in this file. (pdf)

R codes: R codes for all the simulation examples and the diabetes study. (ZIP archive)

Funding

This research was supported in part by NSF grant DMS-1821231 and NIH grant R01GM126550.

References

Barber, R. F., and Ha, W. (2018), "Gradient Descent With Non-Convex Constraints: Local Concavity Determines Convergence," *Information and Inference: A Journal of the IMA*, 7, 755–806. [59]

Bellman, R. (1966), *Dynamic Programming Science*, 153(3731), 34–37. [55]

Bertsimas, D., King, A., Mazumder, R. (2016), "Best Subset Selection Via a Modern Optimization Lens," *The Annals of Statistics*, 44, 813–852. [54,55,58,59]

Brisbane, W., Bailey, M. R., and Sorensen, M. D. (2016), "An Overview of Kidney Stone Imaging Techniques," *Nature Reviews Urology*, 13, 654–662. [52]

Charbonnel, B., Matthews, D., Schernthaner, G., Hanefeld, M., Brunetti, P., and Group, Q. S. (2005), "A Long-Term Comparison of Pioglitazone and Gliclazide in Patients with Type 2 Diabetes Mellitus: A Randomized, Double-Blind, Parallel-Group Comparison Trial," *Diabetic Medicine*, 22, 399–405. [52,62]

Clark, E., Askham, T., Brunton, S. L., and Nathan Kutz, J. (2019), "Greedy Sensor Placement With Cost Constraints," *IEEE Sensors Journal*, 19, 2642–2656. [52]

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1361. [52,55]

Fouskakis, D. and Draper, D. (2008), "Comparing Stochastic Optimization Methods for Variable Selection in Binary Outcome Prediction, With Application to Health Policy," *Journal of the American Statistical Association*, 103, 1367–1381. [53,59]

Fouskakis, D., Ntzoufras, I., and Draper, D. (2009), "Bayesian Variable Selection Using Cost-adjusted BIC, With Application to Cost-Effective Measurement of Quality of Health Care," *The Annals of Applied Statistics*, 663–690. [53,59]

Frank, L. E. and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135. [52]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models Via Coordinate Descent," *Journal of Statistical Software*, 33, 1. [60]

Glover, F. (1977), "Heuristics for Integer Programming Using Surrogate Constraints," *Decision Sciences*, 8, 156–166. [53]

——— (1986), "Future Paths for Integer Programming and Links to Artificial Intelligence," *Computers & Operations Research*, 13, 533–549. [53]

——— (1989), "Tabu Search-Part I," *ORSA Journal on Computing*, 1, 190–206. [53]

Holland, J. H. (1992), *Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence*, Cambridge, MA: MIT Press. [53]

Kachuee, M., Karkkainen, K., Goldstein, O., Zamanzadeh, D., and Sarrafzadeh, M. (2019), "Cost-Sensitive Diagnosis and Learning Leveraging Public Health Data," arXiv preprint arXiv:1902.07102. [52]

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P. (1983), "Optimization by Simulated Annealing," *Science*, 220, 671–680. [53]

Krishnapuram, B., Yu, S., and Rao, R. B. (2011), *Cost-Sensitive Machine Learning*, Boca Raton, FL: CRC Press. [52]

Martello, S., Pisinger, D., and Toth, P. (1999), "Dynamic Programming and Strong Bounds for the 0-1 Knapsack Problem," *Management Science*, 45, 414–424. [55,56]

Nauss, R. M. (1976), "An Efficient Algorithm for the 0-1 Knapsack Problem," *Management Science*, 23, 27–31. [55]

Nesterov, Y. (2013), "Gradient Methods for Minimizing Composite Functions," *Mathematical Programming*, 140, 125–161. [53,55]

Paschos, V. T. (2013), *Paradigms of Combinatorial Optimization: Problems and New Approaches* (vol. 2), Hoboken, NJ: Wiley. [55]

Pattuk, E., Kantarcioglu, M., Ulusoy, H., and Malin, B. (2015), "Privacy-aware Dynamic Feature Selection," in *2015 IEEE 31st International Conference on Data Engineering*, Seoul, South Korea: IEEE, pp. 78–88. [52]

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014), *Lectures on Stochastic Programming: Modeling and Theory*, Philadelphia, PA: SIAM. [59]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [52,55]

Yang, Y., and Zou, H. (2015), "A Fast Unified Algorithm for Solving Group-lasso Penalize Learning Problems," *Statistics and Computing*, 25, 1129–1141. [56,57]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [63]

Yue, L. H. (2010), "Cost-efficient Variable Selection Using Branching LARS," *Electronic Thesis and Dissertation Repository*. [53,59]

Zhang, C. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [52,55]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563. [55]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [52,55,59]

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [52,55,57]