AM221 Final Project Report

Juntao Wang
School of Engineering and Applied Science
Harvard University
juntaowang@g.harvard.edu

Jiayu Yao
School of Engineering and Applied Science
Harvard University
jiy328@g.harvard.edu

Abstract

Previous research has shown that many social science datasets present moderation effect, which refers to a specific type of interaction among features. Given such datasets, it is often challenging to estimate the relationship between features and the target variable with linear regression methods. Since linear regression models assume linear independence among features, it may lead to low prediction accuracy or misinterpretation of the relationship between variables, which is critical in social science studies. In this project, we study the problem of explaining the relationship between variables with the existence of moderators. We firstly show that when the candidate moderators are mutually independent, the moderator selection problem is submodular. We then propose an efficient algorithm to identify these moderators, which enables us not only to predict the target variable accurately but also to interpret the interactions between variables.We also conduct extensive experiments, comparing the performance of our approach to that of regularized linear regression with introduction of extra interaction terms. The experiment results show that our approach works well even when the mutually independence condition is violated. Our model gives better interpretability and, parameter selection is more robust than regularized linear regression.

1 Introduction

Linear regression is widely used in social science to tackle prediction tasks. Although the model is simple, it usually gives decent prediction accuracy and it is easier to interpret comparing to other black box algorithms in machine learning, such as deep neural networks. One major disadvantage of the linear regression model is that it assumes linear independence among dependent variables so it may not perform well when there are moderators in the dataset. Moderators refer to a set of variables Z, which is usually categorical, that affects the strength of the relationship between the independent variables X and the dependent variable y. There are many examples involving the moderators in social science. For example, previous study show that for students who apply to work in for-profit organizations, in average, the annual salary in dollar is approximately 20,000 times the number of the job offers students receive. In contrast, for those who apply to work in non-profit organizations, the annual salary in dollar is only approximately 7,000 times the number of the job offers they receive [1]. In this example, we can see that the relationship between two continuous variables, the salary (the dependent variable) and the number of job offers (the independent variable) is affected by a categorical moderator variable, whether a student applies for work in a for-profit organization or a non-profit organization.

The existence of categorical moderators, as in the previous example, gives rise to difficulty in estimating the relationship between the features and the target variable. If we know these moderators in advance, a common method to tackle this problem is to fit a separate linear regression on each sub-dataset classified based on moderator values. This approach can be reformulated as a moderated multiple regression problem (MMR) [5], which simply includes interaction terms of moderators and

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

the rest of independent variables in the linear regression model. However, in a situation where we do not know the set of moderators (which is usually the case), treating these moderators as independent variables and then fitting the data points with linear regression may lead to high prediction error. We illustrate such phenomenon with a simple example 1. On the other hand, if we treat all categorical features as moderators and then train separate linear regression models on each subset of data divided according to the values of all categorical features, then overfitting problem may occur. The reason is that this approach may include variables that do not have moderation effect which fit the noise existing in the data.

In this project, we aim to attack the prediction task with moderators unknown in advance. More specifically, we consider the situation where there are both categorical and continuous features in the dataset and the task is to predict the target variable accurately. The categorical features may be correlated with each other while continuous features are linearly independent. We take the categorical features as the moderator candidates. The goal is to firstly identify the proper set of moderators, and then, given this set of moderators, to learn the relationship between independent continuous features and the target variable.

To solve the learning problem, we develop two approaches. In the first approach, we formulate the learning problem as an optimal moderator selection problem by adding a penalty term on the size of the moderator set to the residual sum of squares. We show that when these categorical features are mutually independent, under our problem formulation, the objective function is submodular. In Section 4.1, we prove the submodularity of our objective function of a special case when there is no continuous variables(Example 1). In other words, in each regression model, there is only a bias that is moderated by moderators. This special case closely relates to the Analysis of Variance problem (ANOVA), which is of great importance in statistics [4][1]. We further test submodularity on more general case with simulation data. Submodular minimization problem can be solved accurately in polynomial time. However, these algorithms are too complex. Therefore, based on the submodularity of the fitting error, we propose a much simpler greedy algorithm to solve this minimization problem.

In the second approach, we utilize regularized linear regression. We treat all categorical features as moderators and introduce product of categorical moderators and independent continuous features as interaction terms. We then fit a regularized linear regression model on the new feature space. This linear regression model adds a $l1$ norm penalty term on the coefficients, which allows us to do feature selection and do prediction at the same time. Although regularized linear regression perform well in terms of prediction error, it is hard to interpret in our case. Suppose we are given l binary features, there will be 2^{l-1} interaction terms that are related to the feature. However, regularized linear regression doesn't guarantee the coefficients before these terms will be zero or non-zero simultaneously. In addition, the model is sensitive to parameter selection depending on the data.

We conduct intensive simulations comparing the performance of these two approaches under cases when there are continuous features and not, and when the categorical features are correlated and not. We also provide a discussion about the pros and cons of each approach.

Example 1. *Suppose we are given a dependent variable y and two independent categorical features Z_1 and Z_2.*

$$y = \begin{cases} 1+\epsilon & \textit{for } Z_1 = 0, Z_2 = 0 \\ -1+\epsilon & \textit{for } Z_1 = 0, Z_2 = 1 \\ -1+\epsilon & \textit{for } Z_1 = 1, Z_2 = 0 \\ 1+\epsilon & \textit{for } Z_1 = 1, Z_2 = 1 \end{cases} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad Z_1, Z_2 \sim \textit{Bernoulli}(0.5)$$

In this case, y is moderated by two categorical features Z_1 and Z_2. It is easy to see that y cannot be expressed by a linear combination of Z_1, Z_2. If we only look at the relationship between y and Z_1, we get $E[y|Z_1 = 0] = 0, E[y|Z_1 = 1] = 0$. The effect of Z_1 on Y is concealed. We get similar results if we only look at the relationship between y and Z_2. This phenomena shows that any algorithm which only exams the correlation between the label and a proper subset of features may fail to identify the real correlation pattern between these two parties.

[1]The ANOVA investigates the problem that if data samples are divided into different groups according to the values of a set of features, whether the population mean of groups are statistically significantly different. The intuition behind ANOVA is that if the data is divided into groups properly, the variance within groups will be small while the variance among group population means should be large. Detailed discussion is provided in Section 4.1

2 Related Work

2.1 Moderate Multiple Regression (MMR)

Regression with moderation effect has been discovered and studied in social science and statistics for a long time [11, 5, 12, 1, 2, 4]. However, studies in these fields focus on testing whether a given variable is a moderator and whether the moderation effect is statistically significant in data. The detection of the moderators in these fields emphasizes much on prior knowledge of the data [1] and heuristic methods like hierarchical linear regression.

MMR allows the relationship between independent variables and the dependent variable to depend on the level of another set of variables, usually referred as moderator variables. The procedure of moderation effect test is described as follows.

For simplicity, suppose the relationship between two variables $X,\ Y$ varies as a function of a third variable Z. We analyze the relationship among X, Y, Z with two regression analysis. The output has a noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The first regression model M_1 assumes Z does not have moderating effect and it is formulated as follow:

$$Y = \beta_1 X + \beta_2 Z + \epsilon$$

The second regression model M_2 assumes the correlation of X, Y depends on Z by introducing a product term XZ. M_2 is formulated as follow:

$$Y = \beta_1 X + \beta_2 Z + \beta_3 XZ + \epsilon$$

We denote the correlation coefficients of M_1, M_2 as R_1^2, R_2^2 respectively. An *F*-test is performed on:

$$F = [(R_2^2 - R_1^2)/(k_2 - k_1)]/[(1 - R_2^2)/(N - k_2 - 1)]$$

where k_1, k_2 are the number of predictors in M_1, M_2 respectively. N is the sample size.

The time complexity of this approach is exponential. Suppose instead of having one moderator variable, we have a set of candidates Z. Then we need to perform hypothesis testing $2^{|Z|}$ times. MMR also does not guarantee well-fitness of the data since it only looks at R^2.

2.2 Linear Regression with Regularization

During regression analysis, people often add regularization terms to encourage sparsity of coefficients if there is a large number of variables.

Given dependent variable y and independent variable X. The ordinary linear regression model takes the form, $y = X\beta + \epsilon$. The ordinary least square estimate of coefficients is obtained with,

$$\hat{\beta} = \text{argmin}\|y - X\beta\|_2^2$$

$\hat{\beta}$ has a closed form when XX^T is invertible.

To encourage sparsity, people often add penalization on the norm of β. Specifically, lasso regression add l_1 norm penalization, which allows us to do feature selection and do prediction at the same time. The objective function is formulated as,

$$\min_{\beta}(\|y - X\beta\|_2^2 + \lambda\|\beta\|_1)$$

2.3 Submodular feature and rule selection

Our problem essentially can be seen as a feature selection problem. We aim to select a set of categorical features, which have moderating effect on the data. Traditional feature selection methods, such as Principal Component Analysis (PCA), can select out features that are of importance to the dataset but they don't tell us whether those variables are moderators or not.

Submodular feature selection or rule-based selection is widely studied to deal with feature and rule-based selection in general framework [3, 9, 8, 15, 14, 7]. In those methodology, a submodular function is developed to evaluate a set of features or rules, and the set of features that achieves optimal function value is selected.

[3] shows that when the covariance matrix of features satisfies certain condition, the linear regression prediction accuracy is a submodular function of feature set. Therefore, they propose to use Forward Regression algorithm to maximize the accuracy, which gives an (1-ϵ)-approximate solution. [9] selects a subset of features from high dimension feature space. It adopts two submodular functions evaluating the quality of a feature set, and the optimal feature set has been shown to achieve promising performance in many data sets. [8] studies the problem of selecting a proper number of rules from a given rule set for classification task. It shows that there exists a set of submodular objective functions. They then used submodular optimization algorithm to solve the objective function. [15] and [14] selects attributes based on some submodular criterion. We follow similar fashion in existing literature but develop a specific submodular objective function for our specific problem.

3 Problem Formulation

Consider a learning problem where each sample consists of a set of of categorical features $Z = \{Z_1, Z_2, ..., Z_m\}$ ($\forall i \in [m], Z_i$can take value from 1 to k_i), a set of continuous features $X = \{X_1, X_2, ..., X_l\}$ ($X_j \in \mathbb{R}, \forall j \in [l]$) and a label y ($y \in \mathbb{R}$). y is generated from a linear function of X. We are interested in estimating the coefficients (including the bias) which are moderated by X. We aim to identify the moderators in Z and also to minimize the prediction error.

Formally, we denote $\mathbf{z} \in ([k_1] \times [k_2] \times [k_m])$ the value vector of the categorical features, and $\mathbf{x} \in \mathbb{R}^l$ the value vector of the continuous features n. We denote Z' as the set of true moderators, $Z' \subseteq Z$and I as the set of index of moderators ($I \subseteq [m]$). That is, $Z' = \{Z_i | i \in I\}$. $\mathbf{z}_S$ denotes the value vector of under a subset of categorical features, i.e. $\mathbf{z}_S = (z_i)_{i|Z_i \in S}$. We use $\mathbf{w_s}$ to denote the coefficient vector when $\mathbf{z}$ takes value $\mathbf{s}$ $\mathbf{w_s} \in \mathbb{R}^l$. In the learning problem we study, the label y is generated by one of multiple linear models as follows:

$$y = \mathbf{w_s}^\intercal \mathbf{x} + \epsilon, \text{ for } \mathbf{z} = \mathbf{s}, \mathbf{s} \in \prod_{i \in I}[k_i], \epsilon \sim \mathcal{N}(0, \sigma^2)$$

where ϵ models noise of y. The model implies when the value vector of moderators $\mathbf{z}$ differs, the relationship between independent features X and the target variable y changes. The goal is to learn the moderator index set I and the corresponding coefficients $\mathbf{w_s}$ $\forall prod_{i \in I}[k_i]$ so that the overall prediction error is minimized. In the next two sections, we introduce two approaches to solve this learning problem by formulating it as optimization problems.

3.1 Optimal Moderator Selection Formulation

In this section, we reformulate our problem as an optimal moderator selection problem.

We first formally define the objective function for optimal moderator selection. Suppose we are given a dataset $\mathcal{D}$ containing N samples. We denote the ith sample by a tuple $(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}, y^{(i)})$. Suppose we choose S as the potential moderator index set. Then $Z_S = \{Z_j | j \in S\}$ becomes our moderator set. We can fit y under each value combination of Z_S with a linear model on X. We can then calculate overall fitting error on $\mathcal{D}$. We evaluate the prediction error with residual sum of squares (RSS) and define the overall prediction error as follows,

$$f(S) = \frac{1}{N} \sum_{\mathbf{v} \in \prod_{j \in S}[k_j]} \min_{\mathbf{w} \in \mathbb{R}^l} \left(\sum_{i \in \{i | \mathbf{z}_S^{(i)} = \mathbf{v}\}} (y^{(i)} - \mathbf{w}^\intercal \mathbf{x}^{(i)})^2 \right) \tag{1}$$

The minimization operations in Equation 1 computes *RSS* of each sample by fitting it with corresponding linear regression model suggested by $\mathbf{v}$. The *RSS* of each sample is uniformly weighted across all data samples.

It is worth noticing that once adding a categorical feature $Z_i \in Z'$ into S, f will decrease dramatically since Z_i allows us to capture the relationship between X and y correctly. On the other hand, if $Z_i \notin Z'$, f will still decrease since this extra feature enables model to fit the noise of y. However, the model performs badly when new data sample come in since the noise may differ from the original data set. Such phenomenon is referred as overfitting in machine learning. We will refer the original data, the new data as training data, testing data respectively in the following description. In general,

we want to avoid overfitting. Therefore, we add a penalty term on the size S, which allows us to recover the true moderator set:

$$\min_{S \subseteq X} f(S) + \lambda |S| \tag{2}$$

The λ in the formula models the trade-off between the fitting error of the model given the selected categorical feature set and the size of the selected feature set.

For any categorical feature set $S,\ S \supsetneq Z'$, although in expectation over the random noise, it gives similar optimal value on the term ,$f(S)$, to that of true moderator feature set Z', it gets a larger penalty term $\lambda|S|$. Therefore, when λ is larger than the degree of the random noise in the data, the feature set $S, S \supsetneq Z'$ will have a larger objective value and thus will not be an optimal solution. For any categorical feature set $S, S \subseteq Z'$, it has a larger $f(S)$ than Z', because it cannot capture all moderating effect. Therefore, when λ is smaller than the fitting error induced by the moderator effect of a moderator, then S will have a larger objective function and won't be an optimal solution.

To sum up, if λ is set within a proper range, the true moderator set will be the optimal solution of the optimization problem given be Equation 2. In Section 4, we will show under a mild assumption, this optimization problem is a submodular minimization problem and we also illustrate how to tailor the training set to satisfy the assumption.

3.2 Linear Regression with Regularization Formulation

We keep the idea of MMR of introducing extra product terms. Instead of hypothesis testing, we filtering out product terms by using Lasso linear regression. We starts by considering all variables in Z as moderators. And we introduce extra product terms of XZ into the linear regression model. With Lasso linear regression, we formulate the optimization problem as follows,

$$\min_{\mathbf{w},\gamma} \|y - (X\mathbf{w} + + \sum_{S \subseteq Z} \sum_{j=1}^{l} \gamma_{i,j} \prod_{Z_i \in S} X_j Z_l)\|_2^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{l} |\gamma_{i,j}|$$

The product term XZ implies interaction between X and Z. Firstly, we notice that the number of interaction terms added are $O(\prod_{i=1}^{m} k_i)$. Fortunately, by adding a penalty term of $l1$ norm, lasso regression is able to drive the coefficients of product terms to zero if it is not statistically significant. That is, for product terms that contains non-moderator variables, the coefficient before them will be driven down to zero, which allows us to do moderator selection.

4 Solving the Optimal Moderator Selection Problem

In this section, we focus our analysis on a special case of the optimal moderator selection where there is no continuous feature. In other words, the label y is generated by a constant bias, and the bias is moderated by some of the categorical features. We then extend our analysis to more general setting.

4.1 A special case analysis

We consider the special case where there is no continuous feature and some of the categorical features moderate the bias term of the label. This special case is similar to the problem Analysis of Variance (ANOVA) in statistics.

Formally, in this case, the fitting error of a potential moderator set $S \subseteq X$ can be written as

$$f(S) = \frac{1}{N} \sum_{\mathbf{v} \in \prod_{j \in S}[k_j]} \min_{w \in \mathbb{R}} \left(\sum_{i \in \{i | \mathbf{z}_S^{(i)} = \mathbf{v}\}} (y^{(i)} - w \right)$$

Let Y denote the random variable. $Y \sim \mathcal{N}(\mathbf{wx}, \Sigma)$. $Z_i \sim \text{Multinomial}(N, p_i)$. $\mathbf{v}_S$ is a realization of Z_S Then, the fitting error can be expressed by $f(S) = \mathrm{E}_{X_S}\left[\mathrm{Var}\left[Y|X_S\right]\right]$. Next, we show that if Z_is are mutually independent, then the optimal moderator selection problem

$$\min_{S \subset X} \lambda|S| + \mathrm{E}_{X_S}\left[\mathrm{Var}\left[Y|X_S\right]\right] \tag{3}$$

is a submodular minimization problem.

Lemma 1. *Let S and A be two independent discrete random variables with finite support. Let Y be any random variable. Then, $E_S[Var[E_Y[Y|S,A]]] \geq Var[E_S[E_Y[Y|S,A]]]$*

The proof of Lemma 1 is shown in Appendix. The inequality in the lemma doesn't hold when the independence assumption does not hold. For example, one can verify the case where S, A, Y are binary random variables, $p(Y = S) = 1$ and $p(S = A) = 0.9$ such that $\mathrm{E}_S[\mathrm{Var}[\mathrm{E}_Y[Y|S,A]]] = 0$ and $\mathrm{Var}[\mathrm{E}_S[\mathrm{E}_Y[Y|S,A]]] > 0$.

Theorem 1. *Let Ω be a set of mutually independent discrete random variable with finite support, and Y be any random variable. The function $g : 2^\Omega \to \mathbb{R}$,*

$$g(S) = \lambda|S| + E_S\left[Var\left[Y|S\right]\right]$$

is submodular.

Proof. $|S|$ and $-|S|$ are both submodular functions. By additive property of submodular functions, we only need to prove the function $h(S) = E_{V_S}\left[Var\left[Y|V_S\right]\right]$ is a submodular function.

As for any random variable U, V, we have $\mathrm{Var}[U] = \mathrm{E}_V[Var[U|V]] + \mathrm{Var}[\mathrm{E}_U[U|V]]$. For any $S \subsetneq \Omega$ and A a singleton of Ω, $A \cap S = \emptyset$, we have

$$\begin{aligned}
h(S) =& \mathrm{E}_S\left[\mathrm{Var}\left[Y|S\right]\right] \\
=& \mathrm{E}_S\left[\mathrm{E}_A[\mathrm{Var}[Y|S,A]]\right] + \mathrm{E}_S\left[\mathrm{Var}[\mathrm{E}_Y[Y|S,A]]\right] \\
=& \mathrm{E}_{S\cup A}[\mathrm{Var}[Y|S\cup A]] + \mathrm{E}_S\left[\mathrm{Var}[\mathrm{E}_Y[Y|S,A]]\right] \\
=& h(S\cup A) + \mathrm{E}_S\left[\mathrm{Var}[\mathrm{E}_Y[Y|S,A]]\right]
\end{aligned}$$

Let S_1, S_2 be any subset of Ω such that $S_1 \subseteq S_2 \subsetneq \Omega$. Let A be a single of Ω, and $A \cup S_2 = \emptyset$. Denote $T = S_2 \backslash S_1$. We have

$$\begin{aligned}
& h(S_1\cup A) - h(S_1) - (h(S_2\cup A) - h(S_2)) \\
&= \mathrm{E}_{S_1}\left[\mathrm{Var}[\mathrm{E}_Y[Y|S_1,A]]\right] - \mathrm{E}_{S_2}\left[\mathrm{Var}[\mathrm{E}_Y[Y|S_2,A]]\right] \\
&= \mathrm{E}_{S_1}\left[\mathrm{Var}[\mathrm{E}_Y[Y|S_1,A]]\right] - \mathrm{E}_{S_1}\left[\mathrm{E}_T\left[\mathrm{Var}[\mathrm{E}_Y[Y|T,A]]\right]\middle|S_1\right] \\
&= \mathrm{E}_{S_1}\left[\mathrm{Var}[\mathrm{E}_Y[Y|A] - \mathrm{E}_T\left[\mathrm{Var}[\mathrm{E}_Y[Y|T,A]]\right]\middle|S_1\right] \\
&= \mathrm{E}_{S_1}\left[\mathrm{Var}[\mathrm{E}_T[\mathrm{E}_Y[Y|T,A]] - \mathrm{E}_T\left[\mathrm{Var}[\mathrm{E}_Y[Y|T,A]]\right]\middle|S_1\right] \\
&\geq 0
\end{aligned}$$

Because we assume the the mutually independence, we can apply Lemma 1 to show the last inequality holds. □

Based on Theorem 1, when the set of discrete random variable $\{Z_1, ..., Z_m\}$ satisfies the mutually independence condition, then the optimal moderator selection problem is a submodular minimization problem. Here, when we consider the mutually independence condition on the training set, we refer the conditional probability $Pr(Z_i = v|Z_S = \mathbf{v}_S)$ for any $Z_i \in X, v \in \prod_{s\in S}[k_s]$ as the frequency of samples with $Z_i = v$ in all samples where $Z_S = \mathbf{v}_S$. For training set that does not not satisfy this assumption, we can always tailor the training set, i.e., remove samples or duplicate samples, to make it satisfies this assumption. For example, we can divide a training set into sub data set by the values of all categorical features. Then, we select a sub data set with a median size, and uniformly randomly duplicate (remove) samples in sub data sets of smaller (larger) size so as to keep the same size for all sub data set. Then, the new data set satisfies this mutually independence assumption. This trick is commonly used and have shown decent performance in various situations.

4.2 Solution to the special case

In the above section, we show that the optimal moderator selection problem is a submodular minimization problem. Although there exist algorithms [6] to solve the submodular minimization

problem in polynomial time, these algorithms are still too complex to achieve good empirical performance [13]. Here, we propose another algorithm to solve this specific optimization problem under mild assumptions.

Our algorithm works by eliminating features that based on diminish return, and end up with a smaller set of categorical features which contains all the moderators. We notice that at the optimal solution S^* to our optimization problem in Equation 2, we have the optimal condition $f(S^*) - f(S^* \backslash A) + \lambda < 0, \forall A \in S^*$. Meanwhile, f is also a submodular function. So for any S_1, S_2 such that $S_1 \subsetneq S_2 \subseteq X$ and any $A \in S_1$, we have

$$f(S_2) - f(S_2 \backslash \{A\}) + \lambda > 0 \Rightarrow f(S_1) - f(S_1 \backslash \{A\}) + \lambda > 0.$$

Therefore, if $f(Z) - f(Z \backslash \{A\}) + \lambda > 0$, then A must not be an element in the optimal solution. Therefore, we can reduce our search space into $2^{Z \backslash \{A\}}$ and apply the same operation again. We formally present this procedure in Algorithm 1. This algorithm operates in time $\mathcal{O}(|Z|^2|N|)$. When λ is set properly, our algorithm outputs the real moderator set. We show the robustness to the selection of λ in experiments in Section 5.

Algorithm 1 Optimal moderator selection algorithm

Input: The training set D, the set of all categorical features Z, the trade-off parameter λ
Output: A set of categorical features selected as moderators S, and a set W of linear regression models fitted on each sub data set classified by value combinations of categorical features in S.

procedure SELECTMODERATORS(D, X, λ)
  $S \leftarrow X$
  while $S \neq \emptyset$ **do**
    $A \leftarrow \arg\max_{A \in S}\{f(S) - f(S \backslash \{A\})\}$
    $t \leftarrow \max_{A \in S}\{f(S) - f(S \backslash \{A\})\}$
    if $t + \lambda > 0$ **then** $S \leftarrow S \backslash \{A\}$
    else break
  Fit linear regression models on each sub data set classified by categorical features in S, and record models in W
  return S, W

4.3 The general case

In the general case, there exist continuous features which linearly correlated with the label conditioned on some of the categorical features. To extend Algorithm 1 to general case, we just need to include the set Z of continuous features into the regressions when we compute the inner minimization problems of $f(S)$ for any S. This extended algorithm works in time $\mathcal{O}(|S|^2|X|^2|N|)$, where $\mathcal{O}(|X|^2|N|)$ is the time cost to solve the regressions in the computation of $f(\cdot)$, which is called by $\mathcal{O}(|S|^2)$ times.

However, when there exists the continuous features, the submodularity of the optimization problem in Equation 2 does not hold any more. In real life, when we divide data, we have less data points in each subset. Each subset of data may not be large enough to represent the true distribution. However, as long as the sample size is fair, the violation of submodularity is moderate, and our algorithm can still achieve great performance. We show this through intensive experiments in Section 5.

5 Experimental Result

We demonstrated submodularity of our objective function on simulated datasets and compared the performances of two approaches we proposed in previous section. We also ran our algorithms on a real dataset and interpreted the results we get.

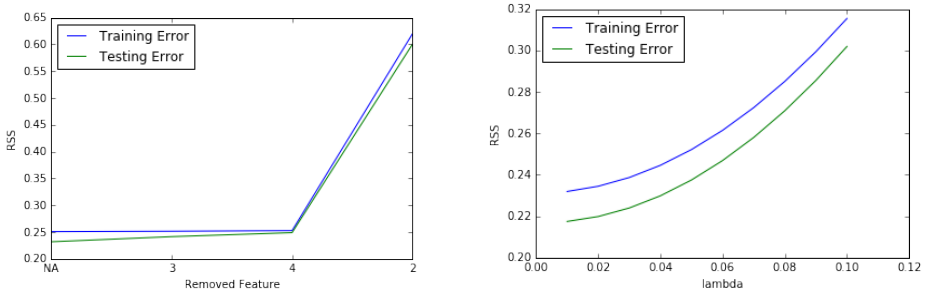
5.1 Experiments 1

We firstly demonstrated submodularity on the special case we proved in 4.1. We generated four independent binary variables $Z_1, \cdots, Z_4 \sim \text{Binomial}(n = 1000, p = [0.5,\ 0.4,\ 0.5, 0.6])$. The target variable y is only moderated by Z_1, Z_2. $y = w_{\mathbf{s}} + \epsilon,\ \epsilon \sim \mathcal{N}(0, 0.5)$

According to submodular function property: If Ω is a finite set, if function $f : 2^{\Omega} \to \mathbb{R}$ is submodular, then for $X, Y \subseteq \Omega$ with $X \subseteq Y$ and every $x \in \Omega \backslash Y$, we have

$$f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$$

We don't list the diminish return here due to space limit. The table is attached in file *exp_discrete.ipynb*.



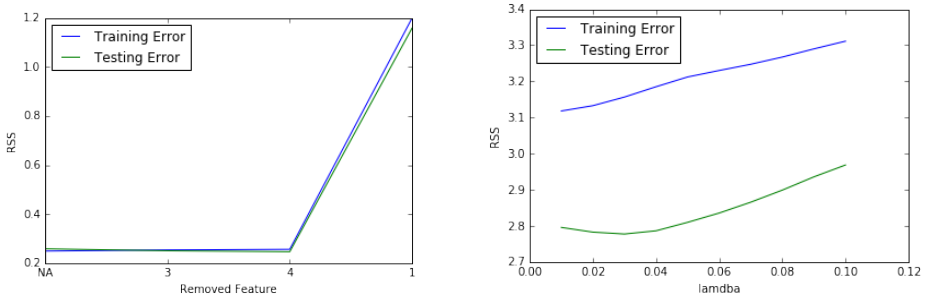
(a) 1 (b) Lasso

Figure 1: (a) A demonstration of performance of 1. The x-axis represents the feature got removed from each round. NA indicates we use all Z (b) A demonstration of performance of Lasso. x-axis indicates penalization λ. For both figures, y-axis represents prediction error. the blue line indicates the training error and the green line indicates the testing error

We fitted the simulated dataset with 1, which identifies the moderator set accurately. We started by including all Z as moderators. In this case, we see the error is the lowest since the model also fits the noise well. Once we removed Z_2, which is one of the true moderators of the dataset, the error increases significantly. We cut the threshold here and we recovered the true moderator set. The significant increase in the error also allows a large range of λ in our objective function. Comparing to Lasso linear regression (feature space is $\mathbb{R}^15$ now), when λ is small, we see larger difference between training error and testing error, which indicates overfitting. If we further increases λ, the performance is worse than our algorithm. We set $\lambda = 0.05$, and we are left with non-zero interaction terms Z_1, Z_1Z_2. This is hard to interpret. If we take Z_1, Z_2 both as moderators, then we are expecting Z_2 to be non-zero as well. If only one of $Z_1,\ Z_2$ are moderators, we still have Z_1Z_2 interaction terms left, whose coefficient is significant. Additionally, we see the prediction error is quite sensitive to the selection of λ as it increases quickly as λ increases.

5.2 Experiments 2

We extended the experiment to a more general case with four binary variable $Z_1, \cdots, Z_4$, two continuous independent variables X and a target variable y. $Z_1, \cdots, Z_4 \sim$ Binomial$(n = 1000, p = [0.5,\ 0.4,\ 0.5, 0.6])$. $X \sim \mathcal{N}(0, 1)$. The relationship between X and y is moderated by Z_1, Z_2. $y = \mathbf{w_s} + \epsilon,\ \epsilon \in \mathcal{N}(0, 0.5)$
We again tested submodularity on the simulated data set. Because of sample size, each subset of data samples is only the approximation of the true distribution. For some $X, Y \in \Omega,\ X \subseteq Y$, the submodularity is slightly violated, but $f(X \cup \{x\}) - f(X) \approxeq f(Y \cup \{x\}) - f(Y)$. The difference magnitude is within 10^{-5}. The table of diminish return table is attached in file *exp_continuous.ipynb*.



(a) 1 (b) Lasso

Figure 2: (a)Performance demonstration of 1. The x-axis represents the feature got removed from each round. NA indicates all variables included (b)Performance demonstration of Lasso. x-axis indicates penalization λ. For both figures, y-axis represents prediction error. the blue line indicates the training error and the green line indicates the testing error

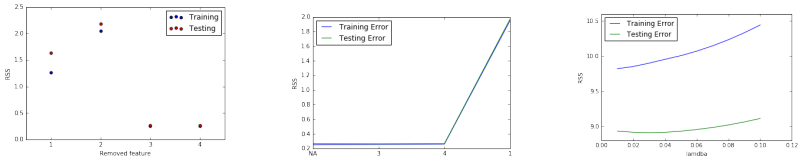
We again fitted the simulated dataset with 1. We see it performs well in terms of identifying the true moderator set and prediction error. The penalization parameter λ can be flexible in a fairly large range due to the huge increase in $f(S)$ when removing Z_2. Lasso linear regression (now with 20 features) performs poorly in terms of prediction error. We examinied the coefficients, and most non-zero coefficients are not related to Z_1, Z_2, which indicates the lasso linear regression model fail to capture the moderating effect in this situation.

5.3 Experiments 3

We also ran experiments of the case when there is correlation among Z. Although submodularity does not hold anymore, the experiment shows that our greedy algorithm still works well. The simulated dataset is generated as follows. We replicated the setting from 5.2. In addition to that, Z_1, Z_3 is correlated in the following way

$$\begin{cases} p(Z_3 = 0|Z_1 = 0) = 0.2 \\ p(Z_3 = 1|Z_1 = 0) = 0.8 \\ p(Z_3 = 0|Z_1 = 1) = 0.8 \\ p(Z_3 = 1|Z_1 = 1) = 0.2 \end{cases}$$

$$\rho(Z_1, Z_3) = \frac{E[Z_1 Z_3] - E[Z_1]E[Z_3]}{\sqrt{Var(Z_1)Var(Z_3)}} = \frac{0.1 - 0.25}{0.25} = -0.6$$



(a) First round (b) 1 (c) Lasso

Figure 3: (a)Prediction error of removing a single feature. The x-axis represents the feature got removed. (b) Performance demonstrantion of 1 on correlated data set. The x-axis represents the feature got removed from each round. (c) Performance of Lasso. x-axis indicates penalization λ. For both figures, y-axis represents prediction error. the blue line indicates the training error and the green line indicates the testing error

We included Figure 3(a) to illustrate the effect of correlation on our greedy algorithm. Because Z_1 and Z_3 is correlated, removing Z_1 gives smaller diminish return than removing Z_2. But the diminish return is still significant for the algorithm to keep it in the selected feature set. In contrast, when there

are correlation among Z, the Lasso linear regression tends to give more non-zero terms. Thus, its performance is worse than when Z are mutually independent.

5.4 Experiments 4

We further tested our greedy algorithm on a real data set from *Good Judgement Project (GJP) dataverse*[2]. The GJP is a social science project focusing on the study of how to motivate crowds to report accurate predictions and how to effectively draw consensus predictions. The dataset contains three part: the forecasters' background, the forecasting events information, and the forecasters' predictions for each event in terms of probability between 0 and 1. The data were collected throughout four consecutive years, where forecasters were recruited each year and they were allowed to quit any time. Our task is to predict the forecasters' prediction score given their background. We used features suggested by [10], which includes two categorical features that indicate forecasters' level of training and levels of elicitation, and a continuous feature that indicates forecasters' raven IQ test score. Our task is to predict forecasters' prediction performance. Since we only have two moderator candidates, the task is fairly simple.

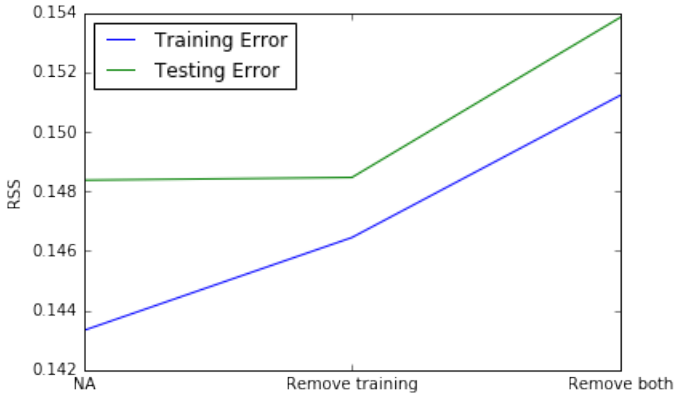


Figure 4: x-axis indicates the selected moderator set. NA means we include both as moderators. y-axis represents the prediction error

Note that on this real data set, the *RSS* doesn't increase significantly even if we remove both variables. The figure may imply that in this dataset, we don't have moderating effect.

6 Conclusion

In social science, study of moderation effect is of great importance since it is present in many problems. Traditional linear regression method performs poorly on solving those problems since the linear independence assumption is violated. In this project, we look at this particular problem of learning relationship between independent variables and the dependent variables, which is influenced by another set of moderators. We reformulate this problem as an optimization problem by proposing a submodular objective function. We also developed a greedy algorithm to solve this submodular minimization problem with fairly well approximation solution. Our algorithm works even if there is correlation among variables. We compared our approach to a widely used machine learning technique by introducing interaction terms and running regularized linear regression analysis. The latter can hardly achieve comparable performance with ours unless the parameters are extremely carefully tuned, while our algorithm is more robust to the parameter selection. Moreover, our algorithm provides the better interpretability as from our algorithm, we know the set of features acting as moderators. In contrast, we can hardly readout the moderators from latter approach.

[2]https://dataverse.harvard.edu/dataverse.xhtml?alias=gjp

References

[1] Herman Aguinis. *Regression analysis for categorical moderators*. Guilford Press, 2004.

[2] Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.

[3] Abhimanyu Das and David Kempe. Algorithms for subset selection in linear regression. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 45–54. ACM, 2008.

[4] Andrew Gelman et al. Analysis of variance—why it is more important than ever. *The annals of statistics*, 33(1):1–53, 2005.

[5] Julie R Irwin and Gary H McClelland. Misleading heuristics and moderated multiple regression models. *Journal of Marketing Research*, 38(1):100–109, 2001.

[6] Satoru Iwata and James B Orlin. A simple combinatorial algorithm for submodular function minimization. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1230–1237. Society for Industrial and Applied Mathematics, 2009.

[7] Rajiv Khanna, Ethan Elenberg, Alexandros G Dimakis, Sahand Negahban, and Joydeep Ghosh. Scalable greedy feature selection via weak submodularity. *arXiv preprint arXiv:1703.02723*, 2017.

[8] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1675–1684. ACM, 2016.

[9] Yuzong Liu, Kai Wei, Katrin Kirchhoff, Yisong Song, and Jeff Bilmes. Submodular feature selection for high-dimensional acoustic score spaces. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7184–7188. IEEE, 2013.

[10] Ville Satopää Philip Tetlock Jon Baron Lyle Ungar, Barbara Mellers. The good judgment project: A large scale test of different methods of combining expert predictions. In *2012 AAAI Fall Symposium Series (RSS)*, 2012.

[11] Subhash Sharma, Richard M Durand, and Oded Gur-Arie. Identification and analysis of moderator variables. *Journal of marketing research*, pages 291–300, 1981.

[12] Piers D Steel and John D Kammeyer-Mueller. Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology*, 87(1):96, 2002.

[13] Peter Stobbe and Andreas Krause. Efficient minimization of decomposable submodular functions. In *Advances in Neural Information Processing Systems*, pages 2208–2216, 2010.

[14] Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, pages 1954–1963, 2015.

[15] Jingjing Zheng, Zhuolin Jiang, and Rama Chellappa. Submodular attribute selection for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2242–2255, 2017.

Appendix

Proof of Lemma 1

Proof. We prove $\mathrm{E}_S[\mathrm{Var}[\mathrm{E}[Y|S,A]]] \geq \mathrm{Var}[\mathrm{E}_S[\mathrm{E}[Y|S,A]]]$ under the case where S and A each follow a uniform distribution over the support and S and A are independent.

Suppose r.v. S takes value uniformly from a set $\{1, ..., m\}$, indexed by i, and r.v. A takes value uniformly from a set $\{1, ..., n\}$, indexed by j. We denote $\mathbb{E}[Y|S = i, A = j]$ by $a_{i,j}$, and denote $a_{i,j} - a_{i,k}$ as $\delta_{i,j,k}, \forall i \in [m], j, k \in [n]$. Then, we have

$$\begin{aligned}
\text{Var}[E[Y|S = i, A]] =& \frac{1}{n} \sum_{j\in[n]} a_{i,j}^2 - \left(\frac{\sum_{j\in[n]} a_{i,j}}{n} \right)^2 \\
=& \frac{1}{n^2} \left(\sum_{j\in[n]} n a_{i,j}^2 - \left(\sum_{j\in[n]} a_{i,j}^2 + \sum_{j,k\in[n], j\neq k} 2a_{i,j}a_{i,k} \right) \right) \\
=& \frac{1}{n^2} \left(\sum_{j\in[n]} (n-1) a_{i,j}^2 - \sum_{j,k\in[n], j\neq k} 2a_{i,j}a_{i,k} \right) \\
=& \frac{1}{n^2} \sum_{j,k\in[n]} (a_{i,j} - a_{i,k})^2 \\
=& \sum_{j,k\in[n]} \frac{1}{n^2} \delta_{i,j,k}^2
\end{aligned}$$

Therefore, we have

$$\text{E}_S[\text{Var}[\text{E}[Y|S, A]]] = \frac{1}{n^2 m} \sum_{i\in[m]} \sum_{j,k\in[n]} \delta_{i,j,k}^2.$$

On the other hand, we denote $\text{E}_S[E[Y|S, A = j]]$ as μ_j. We have $\mu_j = \frac{1}{m} \sum_{i\in[m]} a_{i,j}$. Then, we have

$$\begin{aligned}
\text{Var}[\text{E}_S[\text{E}[Y|S, A]]] =& \sum_{j,k\in[n]} \frac{1}{n^2} (\mu_j - \mu_k)^2 \\
=& \sum_{j,k\in[n]} \frac{1}{n^2} \left(\frac{1}{m} \sum_{i\in[m]} a_{i,j} - \frac{1}{m} \sum_{i\in[m]} a_{i,k} \right) \\
=& \sum_{j,k\in[n]} \frac{1}{n^2 m} \left(\sum_{i\in[m]} a_{i,j} - a_{i,k} \right)^2 \\
=& \frac{1}{n^2 m^2} \sum_{j,k\in[n]} \left(\sum_{i\in[m]} \delta_{i,j,k} \right)^2
\end{aligned}$$

Therefore, we have

$$\text{E}_S[\text{Var}[\text{E}[Y|S, A]]] - \text{Var}[\text{E}_S[\text{E}[Y|S, A]]] = \frac{1}{n^2} \sum_{j,k} \left(\sum_{i\in[n]} \frac{\delta_{i,j,k}^2}{m} - \left(\sum_{i\in[n]} \frac{\delta_{i,j,k}}{m} \right)^2 \right) \geq 0.$$

When S and A are not uniformly distributed but are independent, we can alway bijectively map a unit probability mass of S and A to a probability mass of two corresponding independent uniformly distributed r.v.s and keep the conditional distribution of Y on these two mapped unit probability mass the same. Then, $\text{E}_S[\text{Var}[\text{E}[Y|S, A]]]$ and $\text{Var}[\text{E}_S[\text{E}[Y|S, A]]]$ each will have the same value with the counterpart where we change S and A to the corresponding uniformly distributed r.v.s. Thus, the lemma holds. □