DISCRETE METHODS IN STATISTICS: FEATURE SELECTION AND FAIRNESS-AWARE DATA MINING

Kory D. Johnson

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

Robert A. Stine, Professor of Statistics

Graduate Group Chairperson

Eric Bradlow, K.P. Chao Professor; Professor of Marketing, Statistics, and Education

Dissertation Committee

First Member, Title

Second Member, Title

Third Member, Title

DISCRETE METHODS IN STATISTICS: FEATURE SELECTION AND FAIRNESS-AWARE DATA MINING

ACKNOWLEDGEMENTS

I am supremely grateful to my advisors, Bob Stine and Dean Foster, for their many years of guidance. I often find myself quoting their advice about life and graduate school, and they have shaped far more than just my research skills. To Dean: for inspiring me to get a PhD in statistics and to study any topic, regardless of its dubious connection to my research. To Bob: for being patient with my meandering research interests and providing steady encouragement and feedback. I owe much of the clarity of this work to his persistence and mentoring.

I am grateful to the other members of my committee, Lawrence Brown and Andreas Buja, for our insightful discussions and their insistence that I not move to the next slide unless they understand every detail on the current one.

I would like to thank my family—Mom, Dad, and Krysta— for your constant love and encouragement. Your steadfast confidence in my success has always been motivational. To both my parents and grandparents—Momma Joan and Pop—thank you for believing in and supporting my education. The opportunities I have had and the flexibility I currently enjoy are the result of your generosity.

ABSTRACT

DISCRETE METHODS IN STATISTICS: FEATURE SELECTION AND FAIRNESS-AWARE DATA MINING

Kory D. Johnson

Robert A. Stine

This dissertation is a detailed investigation of issues that arise in models that change discretely. Models are often constructed by either including or excluding features based on some criteria. These discrete changes are challenging to analyze due to correlation between features. Feature selection is the problem of identifying an appropriate set of features to include in a model, while fairness-aware data mining is the problem of needing to remove the *influence* of protected features from a model. This dissertation provides frameworks for understanding each problem and algorithms for accomplishing the desired goal.

The feature selection problem is addressed through the framework of sequential hypothesis testing. We elucidate the statistical challenges in repeatedly using inference in this domain and demonstrate how current methods fail to address them. Our algorithms build on classically motivated, multiple testing procedures to control measures of false rejections when using hypothesis testing during forward stepwise regression. Furthermore, these methods have much higher power than recent proposals from the conditional inference literature.

The fairness-aware data mining community is grappling with fundamental questions concerning fairness in statistical modeling. Tension exists between identifying explainable differences between groups and discriminatory ones. We provide a framework for understanding the connections between fairness and the use of protected information in modeling. With this discussion in hand, generating fair estimates is straight-forward.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : INTRODUCTION

This dissertation discusses a variety of methods in seemingly disparate domains. Broadly speaking, we address issues in model selection, sequential testing, inference after model selection, and fairness-aware data mining. The last of these may seem disconnected from the first three; however, it is fundamentally the same problem. All of these topics analyze the effect of either adding or removing features from a model. This discrete change between models is complex due to correlation between features.

Chapters 2-4 analyze the problem of selecting predictive features from a large feature space. Our data consists of n observations of (response, feature) sets, $(y_i, x_{i1}, \ldots, x_{im})$, where each observation has m associated features. Observations are collected into matrices and the following model is assumed for our data

$$Y = \mathbf{X}\beta + \epsilon \qquad\qquad \epsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n) \tag{1.1}$$

where $\mathbf{X}$ is an $n \times m$ matrix and Y is an $n \times 1$ response vector. Typically, most of the elements of β are 0. Hence, generating good predictions requires identifying the small subset of predictive features. The model (1.1) proliferates the statistics and machine learning literature. In modern applications, m is often large, potentially with $m \gg n$, which makes the selection of an appropriate subset of these features essential for prediction.

The model selection problem is to minimize the error sum of squares

$$\text{ESS}(\hat{Y}) = \|Y - \hat{Y}\|_2^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

while restricting the number of nonzero coefficients:

$$\min_{\beta} \text{ESS}(\mathbf{X}\beta) \quad \text{s.t.} \quad \|\beta\|_{l_0} = \sum_{i=1}^{m} I_{\{\beta_i \neq 0\}} \leq k, \tag{1.2}$$

where the number of nonzero coefficients, k, is the desired sparsity. Note that we are

not assuming a sparse representation exists, merely asking for a sparse approximation. In the statistics literature, the model selection problem (1.2) is more commonly posed as a penalized regression:

$$\hat{\beta}_{0,\lambda} = \text{argmin}_{\beta} \{\text{ESS}(\mathbf{X}\beta) + \lambda \|\beta\|_{l_0}\} \tag{1.3}$$

where $\lambda \geq 0$ is a constant. The classical hard thresholding algorithms C_p (Mallows, 1973), AIC (Akaike, 1974), BIC (Schwarz, 1978), and RIC (Foster and George, 1994) vary λ. The solution to (1.3) is the least-squares estimator on an optimal subset of features. Let $M \subset \{1, \ldots, m\}$ indicate the coordinates of a given model so that $\mathbf{X}_M$ is the corresponding submatrix of the data. If M^*_λ is the optimal set of features for a given λ then $\hat{\beta}^*_{0,\lambda} = (\mathbf{X}^T_{M^*}\mathbf{X}_{M^*})^{-1}\mathbf{X}_{M^*}Y$.

Given the combinatorial nature of the constraint, solving (1.2) quickly becomes infeasible as m increases and is NP-hard in general (Natarajan, 1995). Forward stepwise is the greedy approximation to the solution of (1.2). Let M_i be the features in the forward stepwise model after step i and note that the size of the model is $|M_i| = i$. The algorithm is initialized with $M_0 = \emptyset$ and iteratively adds the variable which yields the largest reduction in ESS. Hence, $M_{i+1} = \{M_i \cup j\}$ where

$$j = \underset{l \in \{1,\ldots,m\} \setminus M_i}{\arg\max} \text{ESS}(\mathbf{X}_{M_i \cup l}\hat{\beta}^{\text{LS}}_{M_i \cup l}).$$

After the first feature is selected, subsequent models are built having fixed that feature in the model. M_1 is the optimal size-1 model, but M_i for $i \geq 2$ is not guaranteed to be optimal, because M_i is forced to include the features identified at previous steps.

Conducting valid inference after selecting a model using an algorithm such as forward stepwise has become a topic of increasing concern. Recently, significant research has been focused on how to compute appropriate p-values for inference post model selection. Chapter 2 introduces a slightly different problem: *how can hypothesis testing be validly used to select*

a model? We want to use hypothesis testing to select one of the models identified by forward stepwise regression. This is a challenging task because the hypotheses being tested are suggested by the data and subsequent tests are only made if previous tests are rejected. Addressing the differences between these two challenges requires increased precision about the quantity of interest when using hypothesis testing for model selection. Our solution uses a sequential testing framework and demonstrates that multiple comparison methods can be adapted to this task. We provide three different procedures to control either the marginal false discovery rate or the family wise error rate. Furthermore, the resulting methods have much higher power than those from the conditional inference literature. This extends the critique of new p-value computations introduced in Brown and Johnson (2016).

Chapter 3 improves upon the methods of Chapter 2, providing several practical improvements which yield an algorithm, Revisiting Alpha-Investing (RAI), which is a fast approximation to forward stepwise. RAI performs model selection in $O(np\log(n))$ time while controlling both type-I and type-II errors. As an alpha-investing procedure, it controls mFDR and is proven to select a model that is a $(1-1/e)+\epsilon$ approximation to the best model of the same size. The algorithm is successful under the assumption of approximate submodularity, which is quite general and allows for highly correlated explanatory variables. We demonstrate the adaptability of RAI by using it to search complex interaction spaces in both simulated and real data.

Submodularity plays an important role in the theorems of Chapters 2 and 3 as it characterizes of the difficulty of the *search* problem of feature selection. The search problem is the ability of a procedure to identify an informative set of features as opposed to the performance of the optimal set of features. This is highly important because merely assuming that there is a true model which performs well does not entail that a modeler can find it. Chapter 4 provides a full discussion of submodularity in statistics. Submodular functions an important function class in optimization which are closely connected to greedy algorithms such as forward stepwise. In statistics, submodularity isolates cases in which

collinearity makes the choice of model features difficult from those in which this task is routine. Researchers often report the signal-to-noise ratio to measure the difficulty of simulated data examples. A measure of submodularity should also be provided as it characterizes an independent component of difficulty. Furthermore, it is closely related to other statistical assumptions used in the development of the lasso, Dantzig selector, and sure information screening.

The feature selection problem is fundamentally concerned with how features can interact in unexpected ways due to correlation. Here, "unexpected" means that the behavior of a feature in a model can be completely different depending on the other features included. One novel domain in which such interactions are highly important is fairness-aware data mining. This can be viewed as a reverse problem to feature selection in which the modeler wishes to remove the influence of a feature. Due to correlation between features, merely excluding the protected feature may leave discriminatory effects which permeate the data.

A simple example clarifies the issue of fairness. Consider a bank that wants to estimate the risk in giving an applicant a loan. The applicant has "legitimate covariates" such as education and credit history, that can be used to determine their risk. They also have "sensitive" or "protected" covariates such as race and gender that society does not want to be used to determine their risk. The bank's task is to model the credit risk or credit score C. To do so, they use historical data, estimate the credit worthiness of the candidate, then determine the interest rate of the loan. The question asked by FADM is whether or not the model the bank constructed is fair. This is different than asking if the data are fair or if the historical practice of giving loans was fair. It is a question pertaining to the estimates produced by the bank's model. This generates several questions. First, what does fairness even mean in this statistical model? Second, what is the role of the sensitive covariates in this estimate? Lastly, how do we constrain the use of the sensitive covariates in black-box algorithms?

The literature has come to an impasse as to what constitutes explainable variability as

opposed to discrimination. This stems from imprecise definitions of fairness in statistics. Chapter 5 provides a detailed account of fairness in statistics by accounting for different perspectives on the data generating process. This results in a tractable framework for understanding fairness in modeling as well as algorithms which are guaranteed to provide fair estimates. These algorithms can be tailored to post-process estimates from arbitrary models to achieve fairness. This effectively separates prediction and fairness goals, allowing the modeler to focus on generating highly predictive models without incorporating the constraint of fairness.

We conclude by discussing the host of new questions raised by this dissertation. As the sequential testing framework is not part of the statistical cannon, there are many lingering questions about its ability to solve broader problems. Similarly, fairness-aware data mining is still in its infancy, and our framework can be extended to many modeling paradigms.

CHAPTER 2 : VALID STEPWISE REGRESSION

Forward stepwise models can be selected in many ways. To illustrate the use of hypothesis testing for model selection, consider the prostate cancer data used to motivate the inference methods of (Taylor et al., 2014). The data set has 67 observations of 8 explanatory variables which will be used to predict the log PSA level of men who had surgery for prostate cancer. The traditional use of stepwise regression is summarized in Table 1. Each step of the procedure adds a feature to the model and assigns a p-value measuring the reduction in ESS using an F-test. Further information on the construction of the stepwise p-values is given in the Appendix. The second column of p-values in Table 1 are from Taylor et al. (2014) and adjust for selecting features using forward stepwise. These adjusted p-values are introduced in Section 2.1.2 and discussed at length in Section 2.3.

Our goal is to use the stepwise p-values in Table 1 to determine when to stop forward stepwise. For example, if it is claimed that the first 4 steps are significant but the 5th is not, the selected model will include lcavol, lweight, svi, and lbph. Such claims should be made solely on the basis of the p-values. That being said, attempting to test the addition of new features uses non-standard and complex distributions (Draper et al., 1971; Pope and Webster, 1972). Our goal is to provide a valid hypothesis testing framework in order to select a forward stepwise model.

Table 1: Stepwise Regression: Prostate Cancer Data

Step	Parameter	Stepwise p-value	Adjusted p-value
1	lcavol	0.0000	0.000
2	lweight	0.0003	0.006
3	svi	0.0424	0.425
4	lbph	0.0468	0.168
5	ppg45	0.2304	0.423
6	lcp	0.0878	0.273
7	age	0.1459	0.059
8	gleason	0.8839	0.156

2.0.1. Notation

We use notation from the multiple comparisons literature given its connection to our solution. Consider m null hypotheses, $H[m]$: $H_1, \ldots, H_m$, and their associated p-values, $p[m]$: $p_1, \ldots, p_m$. The hypotheses can be considered as H_i: $\beta_i = 0$. Define the statistic $R_i = 1$ if H_i is rejected and $R_i = 0$ if not. Similarly, let $V_i^\beta = 1$ if $R_i = 1$ is a false rejection (H_i is true) and $V_i^\beta = 0$ if not. The dependence of V_i^β on β indicates that it is an unknown quantity which depends on the parameter of interest. For simplicity, the definition below suppresses this notation. Define

$$R(m) = \sum_{i=1}^{m} R_i, \text{ and}$$
$$V(m) = \sum_{i=1}^{m} V_i^\beta$$

as the total number of rejections and false rejections in the m tests, respectively.

One object of concern when testing multiple hypotheses is the family wise error rate (FWER), which is the probability of making more than one false rejection regardless of the number of hypotheses tested:

$$\text{FWER} = \mathbb{P}(V(m) \geq 1).$$

If many hypotheses are tested, controlling the FWER may be too strict. Instead, it is often more instructive to control the proportion of false discoveries. Our method controls the marginal false discovery rate (mFDR) which is similar to the more common false discovery rate (FDR):

Definition 1 (Measures of the Proportion of False Discoveries)**.**

$$mFDR(m) = \frac{\mathbb{E}(V(m))}{\mathbb{E}(R(m)) + 1}$$
$$FDR(m) = \mathbb{E}\left(\frac{V(m)}{R(m)}\right), \textit{ where } \frac{0}{0} = 0.$$

In some respects, FDR is preferable to mFDR because it controls a property of a *realized* distribution. While not observed, the ratio $V(m)/R(m)$ is the realized proportion of false rejections in a given use of a procedure. FDR controls the expectation of this quantity. In contrast, $\mathbb{E}(V(m))/\mathbb{E}(R(m))$ is not a property of the distribution of $V(m)/R(m)$. That being said, FDR and mFDR behave similarly in practice, and mFDR yields a powerful and flexible martingale (Foster and Stine, 2008). This martingale provides the basis for proofs of type-I error control in a variety of situations.

2.0.2. Contributions

Our first contribution is an elucidation of the effects that must be considered when using hypothesis testing for model selection. Standard inference tools are invalidated due to two selection effects: the *ranking* effect and the *testing* effect. The ranking effect is the result of testing hypotheses that are suggested by the data and the testing effect is the result of only conducting future tests if previous tests have been rejected. The impacts of these effects are explained via example in Section 2.1.

In Section 2.2.1, we demonstrate that the sequential testing approach to multiple comparisons yields an approximate forward stepwise algorithm that controls for the selection effects. Our procedure, Revisiting-Holm (RH), is a threshold approximation to stepwise regression (Badanidiyuru and Vondrák, 2014). At each step, forward stepwise sorts the p-values of the m' remaining features, $p_{(1)} < \ldots < p_{(m')}$, and selects the feature with the minimum p-value, $p_{(1)}$. Instead of performing a full sort, threshold approximations use a set of increasing rejection thresholds, and hypotheses are rejected when their p-value falls below a threshold. A feature merely needs to be significant *enough*, and not necessarily the *most* significant. The initial rejection threshold conducts a strict test for which only highly significant features are added to the model. Subsequent thresholds perform less stringent tests. As such, the final model is built from a series of approximately greedy choices.

Proofs for the type-I error control of our procedures are conservative under an assumption

of submodularity. While not often discussed in the statistics literature, submodularity has important statistical implications and is closely related to more commonly used assumptions (Johnson et al., 2015c). Submodularity requires that features do not become more significant when included in a multiple regression than in a simple regression. More generally, consider a feature $\mathbf{X}_i$ orthogonal to those in a model, $\mathbf{X}_M$. This is referred to as adjusting $\mathbf{X}_i$ for $\mathbf{X}_M$. The projection operator (hat matrix), $\mathbf{H}_M = \mathbf{X}_M(\mathbf{X}_M^T\mathbf{X}_M)^{-1}\mathbf{X}_M^T$, computes the orthogonal projection of a vector onto the span of the columns of $\mathbf{X}_M$. Therefore, $\mathbf{X}_i$ adjusted for $\mathbf{X}_M$ is denoted $\mathbf{X}_{i.M^\perp} = (\mathbf{I} - \mathbf{H}_{\mathbf{X}_M})\mathbf{X}_i$. A suppressor variable is one which, once adjusted for, *increases* the observed significance of another feature. Submodularity is equivalent to the absence of conditional suppressor variables, implying that $\forall M \subset \{1, \ldots, m\}$ and $i, j \notin M$

$$|\text{Corr}(Y, \mathbf{X}_{i.(M\cup j)^\perp})| \leq |\text{Corr}(Y, \mathbf{X}_{i.M^\perp})|.$$

When this holds, stepwise p-values are non-decreasing as p-values are smaller when features are considered in smaller models. Clearly the prostate cancer data does not satisfy this; however, the first several steps have p-values which are non-decreasing. The assumption of submodularity can be relaxed as discussed in Johnson et al. (2015c), but doing so here unnecessarily complicates our discussion.

Our first result is proven in Section 2.2.3 by demonstrating that RH is an alpha-investing procedure (Foster and Stine, 2008). RH is presented independently of alpha-investing so that the algorithm and proof method are not conflated.

Theorem 1. *If the data (*Y*,*$\mathbf{X}$*) are submodular and* M^* *is the model chosen by Revisiting-Holm with user defined parameter* α*, then*

$$\frac{\mathbb{E}(V(m))}{\mathbb{E}(R(m)) + 1} \leq \alpha$$

In some cases, the approximation of RH may be unsatisfactory. Section 2.2.2 provides two relaxations of RH that can be used to stop forward stepwise. The first relaxation uses

the rejection thresholds used by RH as rejection levels for the true stepwise path. This procedure, "Approximate Revisiting-Holm" (aRH), is conjectured to control FDR under submodularity. While the martingale proofs of Foster and Stine (2008) are distorted, the simulations of Section 2.3.2 support this claim.

Conjecture 1. *If the data (Y,***X***) are submodular and M^* is the model chosen by Approximate Revisiting-Holm with user defined parameter α, then*

$$\frac{\mathbb{E}(V(m))}{\mathbb{E}(R(m)) + 1} \leq \alpha$$

Our construction of RH motivates one final relaxation: using the Holm significance levels as the rejection thresholds. This is introduced in Section 2.1.2. The resulting procedure, Stepwise-Holm (SH), is closely related to the Max-|t| procedure of Buja and Brown (2014) and controls the FWER under submodularity.

Theorem 2. *If the data (Y,***X***) are submodular and M^* is the model chosen by Stepwise-Holm with user defined parameter α, then*

$$\mathbb{P}(V(m) \geq 1)) \leq \alpha$$

As noted in Taylor et al. (2014), the Max-|t| procedure can be highly conservative, which is expected as it controls the FWER. That being said, it performs extremely well in the simulations of Section 2.3.2.

Our framework clearly shows the shortcoming of other procedures recently recommended in the literature (Taylor et al., 2014). This is discussed at length in Section 2.3, where we demonstrate that classically motivated methods such as RH are preferred. Further evidence is provided in Section 2.3.2, where RH and its relaxations are shown to have much higher power than competing methods. This extends the discussion of Brown and Johnson (2016).

2.1. Inference for Model Selection

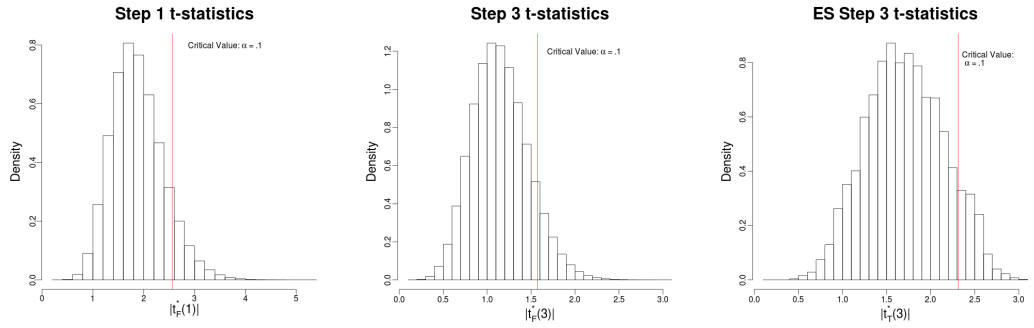
Attempting to use inference for model selection poses significantly different challenges than merely performing inference after a model is selected. Inference will be conducted multiple times based on the result of previous inferential claims. Section 2.1.1 describes two separate issues raised by such procedures, while Section 2.1.2 demonstrates that current solutions do not address both issues.

2.1.1. Selection Effects

We use a simple simulation to demonstrate that it is difficult to provide a valid stepwise procedure even in the orthogonal case. This separates questions about the statistical validity of forward stepwise from its ability to approximate the sparse regression problem (1.2). The model identified at the kth step of forward stepwise exactly solves (1.2) under orthogonality. The assumption of submodularity guarantees that forward stepwise is both a reasonable approximation to (1.2) (Nemhauser et al., 1978) and that our methods provide statistical guarantees. This is discussed in Section 2.2.

Suppose the data contain 10 orthogonal explanatory features, $\beta_1 = \ldots = \beta_{10} = 0$, and σ^2 is known. In this case, the test statistics are iid $N(0, 1)$ variables but will be called t-statistics for consistency with data applications. The t-statistics for $H_1, \ldots, H_{10}$ are $t_1, \ldots, t_{10}$ with corresponding p-values $p_1, \ldots, p_{10}$. The feature selection problem is equivalent to determining an order for testing $H[m]$ while controlling false rejections at level α. Since our goal is model selection, a feature is "included" or "added" to the model when the corresponding null hypothesis is rejected. Sort the hypotheses by their absolute t-statistics as $|t_{(1)}| > \ldots > |t_{(10)}|$ (equivalently $p_{(1)} < \ldots < p_{(10)}$). At step i, forward stepwise tests $H_{(i)}$. In the orthogonal setting, test statistics and p-values do not change depending on the order in which hypotheses are tested, because coefficients do not change due to the model in which they are estimated.

As expected, the distributions of the absolute order statistics are significantly different than



(a) (b) (c)

Figure 1: Illustration of selection and sequential effects under the global null hypothesis.

the naive $|N(0,1)|$. Figure 1a and 1b show the distributions of $|t_{(1)}|$ and $|t_{(3)}|$. Informally, the difference between these distributions and the distribution of $|N(0,1)|$ is the ranking effect. This name is motivated as the difference between the test of a rank statistic and a randomly chosen one. A more precise definition follows shortly.

Since our goal is not to estimate the correct distribution but to perform a valid test, we desire a critical value yielding a level-α test. The nominal $\alpha = .1$ critical value is $t = 1.645$, whereas the simulated threshold is $t = 2.58$. This value can be easily computed using the Bonferroni correction, and the expected size of further rank statistics can be computed in the orthogonal case (George and Foster, 2000). The .1-critical value for $|t_{(3)}|$ is approximately 1.57, which is lower than the naive level-.1 significance threshold. This is intuitive as $|t_{(3)}|$ is constrained to be less than $|t_{(2)}|$ by definition.

To be consistent with the standard use of hypothesis testing during forward stepwise, we propose the procedure "Exact Stepwise (traditional)" (ES-t), that terminates on the first step in which a hypothesis fails to be rejected using traditional stepwise p-values.[1] Such a perspective is often necessary to allow early termination of an algorithm in large feature spaces. ES-t only tests $H_{(i)}$ if $H_{(1)}, \ldots, H_{(i-1)}$ were rejected. On the subset of cases in which this occurs, $|t_{(i)}|$ is much less constrained, because all of $|t_{(1)}|, \ldots, |t_{(i-1)}|$ were large

[1] A similar procedure is discussed in Section 2.3 that uses the methodology of Taylor et al. (2014) which corrects p-values for ranking.

enough to be rejected. The distribution of $|t_{(3)}|$ considered by ES-t is only realized on the subset of cases in which $H_{(3)}$ is actually tested. Figure 1c shows the distribution of $|t_{(3)}|$ on the subset of cases in which $H_{(1)}$ and $H_{(2)}$ were rejected using $\alpha = .1$ under the Holm method. The Holm method is explained in Section 2.1.2 and entails that $p_{(1)} < \alpha/m$ and $p_{(2)} < \alpha/(m-1)$. Informally, the difference between the distributions in Figures 1b and 1c is the testing effect. The testing effect increases the simulated critical value from 1.57 to 2.32.

In order to provide precise definitions of the ranking and testing effects, it is necessary to define different distributions of statistical interest. Classical inference procedures are interested in controlling the *nominal type-I error* from a test specified prior to seeing the data:

$$\mathbb{P}_{M,H_0}(\text{reject } H_0) \leq \alpha. \tag{2.1}$$

Both the model M and null hypothesis H_0 are specified ex-ante and the test is conducted assuming that the data originate from the model M. If M is misspecified, inference is still possible though the object of inquiry is the best approximation of the true mean of Y in the model M (Buja et al., 2014). Orthogonality breaks the dependence of H_0 on M, as the presence of other variables does not change coefficient estimates. Most situations are similar to forward stepwise, in which M and H_0 are chosen together, reintroducing dependence. Therefore, the orthogonal case is still non-trivial.

Often both M and H_0 are the result of exploratory data analysis which invalidates the assumption of pre-specified inference goals. In this case, a reasonable alternative is to control the *selective type-I error* (Fithian et al., 2015):

$$\mathbb{P}_{M,H_0}(\text{reject } H_0 | (M, H_0) \text{ selected}) \leq \alpha. \tag{2.2}$$

Given the use of algorithms to identify models, we will separate two different types of selective type-I errors. The first arises when M is selected by a *fixed* algorithm $\mathcal{A}$. M is

still random as it depends on the data, but the algorithm $\mathcal{A}$ does not contain a random component such as statistical tests. For example, one could test the hypothesis selected on the third step of forward stepwise. The resulting error is the *fixed-selective type-I error*:

$$\mathbb{P}^{F}_{M,H_0}(\text{reject } H_0 | (M, H_0) \text{ selected by } \mathcal{A}) \leq \alpha. \tag{2.3}$$

The parameters M and H_0 are understood to be dependent on the algorithm, $\mathcal{A}$, used in their identification: $M = M(\mathcal{A})$ and $H_0 = H_0(\mathcal{A})$. For brevity, this dependence is included as a superscript in the probability notation.

The testing effect and the discussion in Brown and Johnson (2016) lead us to be more pedantic in the definition of the selective type-I error rate. We explicitly note the dependence of the algorithm $\mathcal{A}$ on test statistics $\mathcal{T}$ in the definition of the *total-selective type-I error*:

$$\mathbb{P}^{T}_{M,H_0}(\text{reject } H_0 | (M, H_0) \text{ selected by } (\mathcal{A}, \mathcal{T})) \leq \alpha. \tag{2.4}$$

As before, the parameters M and H_0 are understood to be dependent on both the fixed algorithm, $\mathcal{A}$, and the hypothesis tests, $\mathcal{T}$, used in their identification: $M = M(\mathcal{A}, \mathcal{T})$ and $H_0 = H_0(\mathcal{A}, \mathcal{T})$. For brevity, this notation is included as a superscript in the probability notation. Acknowledging that test statistics are used to select M does not change the concept of selective type-I error as the set $(\mathcal{A}, \mathcal{T})$ merely produces a meta-algorithm for performing selection. The difficulty in requiring a completely specified algorithm motivates the post-selection inference methods of Berk et al. (2013).

Using the nominal, fixed-selective, and total-selective type-I errors, we can clearly define the two selection effects which arise when using hypothesis testing for model selection. The ranking effect is defined as the lack of equivalence between the nominal and fixed-selective type-I error, and the testing effect is defined as the lack of equivalence between the fixed-selective and total-selective type-I error:

Definition 2 (Selection Effects)**.**

Ranking Effect:

$$\mathbb{P}_{M,H_0}(\textit{reject } H_0) \neq \mathbb{P}^F_{M,H_0}(\textit{reject } H_0 | (M, H_0) \textit{ selected by } \mathcal{A})$$

Testing Effect:

$$\mathbb{P}^F_{M,H_0}(\textit{reject } H_0 | (M, H_0) \textit{ selected by } \mathcal{A}) \neq \mathbb{P}^T_{M,H_0}(\textit{reject } H_0 | (M, H_0) \textit{ selected selected by } (\mathcal{A}, \mathcal{T}))$$

Both selection effects are the result of a selection procedure but are given different names to separate important distinctions in *types* of selection. The distinction can be seen by considering two separate methods to identify a forward stepwise model. The fixed algorithm perspective runs forward stepwise a specified number of steps, then tests the feature being added. If a stopping condition is used such as selecting the model with minimum cross-validated error, this must be specified and included in the conditioning. Loftus (2015) represents such a procedure as a set of constraints on Y in order to condition on the selection event. Therefore, $\mathbb{P}^F_{M,H_0}(\text{reject } H_0 | (M, H_0) \text{ selected by } \mathcal{A})$ is the appropriate object of inquiry and coefficients in the final model can be tested via their methods. Alternatively, ES-t can select the forward stepwise model using the p-values in Table 1. The model chosen via this method is the *result* of repeated hypothesis testing and the correct object of inquiry is $\mathbb{P}^T_{M,H_0}(\text{reject } H_0 | (M, H_0) \text{ selected by } (\mathcal{A}, \mathcal{T}))$. ES-t requires hypotheses to be tested *sequentially* and future tests are influenced by the results of past tests.

2.1.2. Problems with Previous Solutions

Broadly speaking, there are two perspectives on how to account for the selection effects. The first computes a p-value that corrects for selection. This is a challenging task and is impossible in some cases. If M is identified through a model selection procedure, $\mathbb{P}_{M,H_0}(\text{reject } H_0)$ cannot be estimated (Leeb and Pötscher, 2006). The hypothesis H_0 may be specified prior to data analysis, but M is often selected from a large set of candidate models based on per-

formance measures such as AIC. While this is a common practice, the inability to estimate $\mathbb{P}_{M,H_0}(\text{reject } H_0)$ renders control of the nominal type-I error impossible.

Estimation is made possible by conditioning on the selection event. Taylor et al. (2014) control the fixed-selective type I error rate, $\mathbb{P}^F_{M,H_0}(\text{reject } H_0|(M, H_0) \text{ selected})$, by specifying the choice of M via algorithm $\mathcal{A}$ as constraints on the response Y. For example, if $\mathbf{X}_1$ is chosen on the first step of forward stepwise, then the t-statistic of $\hat{\beta}_1$ is larger than that of $\hat{\beta}_i$, for $i \neq 1$. This implies a set of linear restrictions on Y. The algorithm $\mathcal{A}$ is fixed and is purely used for optimization as no decisions depend on the result of statistical tests. Their calculations result in statistics with a uniform distribution under H_0, and hence are called "exact p-values." The second column of p-values in Table 1 are the "exact p-values" computed using their procedure when forward stepwise is run on the prostate data.

While Taylor et al. (2014) do not explicitly advocate using the p-values as a way to select models, both the tacit discussion of modeling and the corresponding R package encourage such a use. One might think that improved p-values would lead to improved model selection, at least in some circumstances; however, the formulation in Taylor et al. (2014) involves a serious paradox. One needs to begin with a well-specified model selection algorithm and construct a model independent of the exact p-values described in the paper. The exact p-values can be constructed only after the model has been chosen; they cannot validly be used to select the model. If one tries to use them in this way, they become invalid, because such tests are not incorporated into the constraints on Y. While this does not invalidate the methodology, it both changes their p-values and significantly hinders computation. This paradox is also raised in Brown and Johnson (2016) and will be described fully in Section 2.3.

Furthermore, the corresponding package, *selectiveInference*, suggests selecting models using procedures from G'Sell et al. (2015). The independence between between p-values assumed by G'Sell et al. (2015) is of less concern than the fact that the p-values produced by *selectiveInference* do not account for the testing effect. To have valid conditional p-values,

selectiveInference must account for the influence of the G'Sell et al. (2015) selection procedure; however, these procedures are not valid stopping rules, and thus the conditioning event $M(\mathcal{A}, \mathcal{T})$ must encode the *entire* selection path. The adjusted p-value at step i depends on the result of calculations from step 1 to the maximum step m'. We provide more details and a conservative procedure in Section 2.3.2 which allows for early stopping. The computational cost of incorporating even the simpler constraints implied by the conservative procedure likely renders it impractical.

The second potential solution to inference for model selection uses traditional, stepwise p-values but changes the rejection threshold. Multiple comparison procedures such as Holm (Holm, 1979) and Benjamini-Hochberg (Benjamini and Hochberg, 1995) are of this form. Bonferroni is a classical, conservative method for controlling for multiple comparisons by bounding the FWER. Bonferroni changes the rejection threshold from α to α/m, so that H_i is only rejected if $p_i \leq \alpha/m$. This controls the FWER by Boole's inequality. The second relevant procedure is the Holm step-down method, which proceeds as:

1. Sort the p-values: $p_{(1)} < \ldots < p_{(m)}$ and corresponding hypotheses $H_{(1)}, \ldots, H_{(m)}$.
2. Identify $k = \min_i p_{(i)} > \frac{\alpha}{m-i+1}$.
3. Reject $H_{(1)}, \ldots, H_{(k-1)}$.

Holm rejects $H_{(1)}$ if $p_{(1)}$ falls below the Bonferroni level with m hypotheses: α/m. If $H_{(1)}$ is rejected, only $m-1$ hypotheses are still being considered; hence, $p_{(2)}$ is compared to the Bonferroni threshold using $m-1$ hypotheses: $\alpha/(m-1)$. This increases the power of subsequent tests, particularly if many hypotheses are rejected. The final test of $H_{(m)}$ can even be carried out at the nominal level α. The FWER is controlled according to the closure principle.

Instead of controlling the probability of making any false rejections, consider controlling the proportion of false rejections. This provides higher power and may be more appropriate if many hypotheses are being tested. The Benjamini-Hochberg step-down procedure (BH)

Table 2: Comparison of Holm and BH p-value rejection thresholds.

	1	2	3	4	5	6	7	8	9	10
Holm	0.010	0.011	0.013	0.014	0.017	0.020	0.025	0.033	0.050	0.100
BH	0.010	0.020	0.030	0.040	0.050	0.060	0.070	0.080	0.090	0.100

controls FDR (Benjamini and Hochberg, 1995). BH proceeds similarly to Holm:

1. Sort the p-values: $p_{(1)} < \ldots < p_{(m)}$ and corresponding hypotheses $H_{(1)}, \ldots, H_{(m)}$.
2. Identify $k = \min_i p_{(i)} > i\frac{\alpha}{m}$.
3. Reject $H_{(1)}, \ldots, H_{(k-1)}$

As in Holm, sorted p-values are compared to a rejection threshold and the algorithm terminates when the threshold is exceeded. Both procedures test $p_{(1)}$ using the Bonferroni threshold α/m and test $p_{(m)}$ at the nominal level α. BH increases linearly between these endpoints, providing significantly higher power. Table 2 compares the rejection thresholds of the two methods when $m = 10$ and $\alpha = .1$.

Both the Holm and BH procedures can be improved in the sense that more hypothesis can be rejected while maintaining control of their respective error criteria. Instead of setting k as the first time $p_{(k)}$ exceeds the required threshold, set $k-1$ to be the *last* time that $p_{(k-1)}$ is *less* than the required threshold. We focus on step-down procedures as opposed to these step-up procedures, because they are similar to sequential methods in that they provide valid stopping rules. Step-up procedures require all p-values to be computed before a set of hypotheses can be rejected. This is often unsatisfactory for model selection if there are many features. For example, the full stepwise path, or at least a predefined maximum number of steps, must be computed before a model can be identified.

We measure the ranking and testing effects by the difference in critical values for a level-α test from the different error distributions. In the following display, the notation is simplified by excluding M. At step i, the relevant model is $M_i = \{H_{(1)} \cup \ldots \cup H_{(i)}\}$. For each i, define

the following critical values:

$$\begin{aligned}
\text{Nominal threshold:} \quad & t_N^*(i) \quad \text{s.t.} \quad && \mathbb{P}_{H_{(i)}}(|t_i| > t_N^*(i)) = \alpha \\
\text{Fixed-sequential threshold:} \quad & t_F^*(i) \quad \text{s.t.} \quad && \mathbb{P}^F_{H_{(i)}}(|t_i| > t_F^*(i) | H_{(i)}(\mathcal{A}) \text{ selected}) = \alpha \\
\text{Total-sequential threshold:} \quad & t_T^*(i) \quad \text{s.t.} \quad && \mathbb{P}^T_{H_{(i)}}(|t_i| > t_T^*(i) | H_{(i)}(\mathcal{A}, \mathcal{T}) \text{ selected}) = \alpha
\end{aligned}$$

The nominal threshold $t_N^*(i)$ ignores the selection effects and hence treats all observed statistics as $N(0,1)$. Thus the ranking effect caused by testing hypothesis which are suggested by the data is measured by $t_N^*(i) - t_F^*(i)$. For example, $|t_{(1)}|$ is the maximum of 10 t-statistics and is not distributed as $|N(0,1)|$. The first step does not demonstrate the testing effect as no tests have occurred: $t_F^*(1) = t_T^*(1)$. On subsequent steps, the testing effect is measured by $t_T^*(i) - t_F^*(i)$, where the latter accounts for the previous hypothesis tests and the former does not.

Table 3 compares simulated critical values to those generated by Holm and BH for $\alpha = .1$. The true fixed-selective critical value, $t_F^*(i)$, is merely the .9-quantile of the distribution of $|t_{(i)}|$. The true total-selective critical value, $t_T^*(i)$, is the .9-quantile of the distribution of $|t_{(i)}|$ *on the subset of cases in which* $H_{(1)}, \ldots, H_{(i-1)}$ *were rejected according to some procedure.* Therefore, $t_T^*(i)$ depends on the hypothesis testing rule. As identifying the appropriate rule is the a goal of this paper, we present a couple of different options. The "true" total-selective $t_T^*(i)$ is the critical value if all previous tests are conducted at the "true" total-selective threshold, where the initial test does not contain the testing effect. We also show $t_T^*(i)$ when hypotheses are tested using Holm, BH, and our proposal, Revisiting Holm.

First, it is clear that the nominal threshold does not account for the selection effects and is misleading. The critical value $t_N^*(i)$ is initially much smaller than the corrected values, but quickly becomes larger than $t_F^*(i)$ and larger than $t_T^*(i)$ for large i. Second, by the fifth step, $t_F^*(i)$ and "true" $t_T^*(i)$ differ by almost 1 or approximately a factor of 2. The

Table 3: Simulated critical values under global null.

Step	1	2	3	4	5
Nominal $t_N^*(i)$	1.65	1.65	1.65	1.65	1.65
Fixed-Selective $t_F^*(i)$	2.58	1.92	1.57	1.32	1.11
"True" Total-Selective $t_T^*(i)$	2.58	2.39	2.26	2.15	2.12
Holm Total-Selective $t_T^*(i)$	2.58	2.40	2.32	2.29	-
Holm	2.58	2.54	2.50	2.45	2.39
BH Total-Selective $t_T^*(i)$	2.58	2.40	2.24	2.08	1.99
BH	2.58	2.33	2.17	2.05	1.96
RH Total-Selective $t_T^*(i)$	2.58	2.30	2.12	1.97	1.87
RH	2.58	2.31	2.13	1.99	1.86

difference is the testing effect and it quickly dominates the rank effect. The intuition for the magnitude of the testing effect was provided previously: on the subset of cases in which $H_{(i)}$ is tested by ES-t, $|t_{(i-1)}|$ does not place a strong constraint on $|t_{(i)}|$. The relative sizes of the selection effects is troubling as the conditional methods of Taylor et al. (2014) ignore the testing effect.

Third, there are large differences between the simulated critical values and those produced using the corresponding multiple comparison methods. While the procedures control very different error measures, it is instructive that neither method estimates the corresponding total-selective critical value correctly. This demonstrates the need for a new method accounting for the ranking and testing effects. Lastly, the bottom two rows show the critical values produced by our approximate stepwise procedure Revisiting Holm. The computed values match the simulated critical values $t_T^*(i)$.

As we demonstrate in Section 2.2, sequential testing fits somewhere between the two potential solutions discussed in this section. We use stepwise p-values such as those from Table 1 to craft an algorithm which looks like a multiple comparison procedure; however, an important update occurs between tests which changes the "effective" testing level. This update also accounts for the differences between the RH total-selective $t_T^*(i)$ and the "true" total-selective $t_T^*(i)$ seen in Table 3.

2.2. Sequential Testing

The Revisiting Holm procedure (RH) is motivated by controlling multiple comparisons in a sequential testing framework. Sequential testing assumes the hypotheses $H[m]$ arrive sequentially. As such, the current hypothesis must be tested before observing subsequent hypotheses. This leads to a new mindset for multiple comparison control as well as corrected rejection thresholds for inference for model selection. While the corrections are only exact in the orthogonal case, we demonstrate that they are robust to certain deviations from orthogonality. Furthermore, the simulations in Section 2.3.2 demonstrate that RH controls FDR in nonorthogonal cases and has much higher power than competitors.

2.2.1. Approximating Stepwise Regression

At each iteration, forward stepwise sorts the stepwise p-values of all remaining features in order to select the feature with the minimum p-value $p_{(1)}$. Instead of performing a full sort, consider using increasing rejection thresholds, where hypotheses are rejected when their p-value falls below a threshold. For now, set aside the concern about type-I error control to focus on the order in which covariates are selected by RH. Consider thresholds determined by the Holm step-down procedure. The approximate stepwise algorithm is:

1. Test $H_{(1)}, \ldots, H_{(m)}$ at level α/m.
2. If $p_{(1)} > \alpha/m$, stop with no rejections, else reject $H_{(1)}$. This was the first "pass" through the features or "round" of testing. For now, assume that only one rejection is made on this pass, ie, $p_{(2)} > \alpha/m$.
3. Test $H_{(2)}, \ldots, H_{(m)}$ at level $\alpha/(m-1)$, following the rejection procedure as before. Again, assume only one rejection.
4. Continue making testing passes using the Holm thresholds until all remaining hypotheses fail to be rejected in a round, then terminate. The selected model includes the variables corresponding to the rejected hypotheses.

While this looks identical to the original Holm procedure, there is an important distinction: hypotheses are formally tested multiple times. Therefore, the procedure must condition on the results of previous tests.

To introduce the implications of this conditioning, consider a level-α test with rejection threshold p^*. In general, α and p^* are not discussed of as two separate parameters, because they are equal when the p-value is uniformly distributed and unconstrained. When conditioning on the result of previous tests, however, they are different.

Definition 3 (Level-α test, rejection threshold p^*)**.**

$$\mathbb{P}_{H_0}(p \leq p^*) = \alpha \quad \Rightarrow \quad \textit{Standard case:} \quad p^* = \alpha$$

When testing a hypotheses on the second round, one must account for the failed test in the first round. In our simulation example with $m = 10$ and $\alpha = .1$, the second pass performs a level-.01/9 test conditional on the p-value being greater than the first pass threshold of .1/10. Under the null hypotheses, the sequential p-value is uniformly distributed; hence the threshold p^* can be computed as

$$\begin{aligned} .1/9 &= \mathbb{P}_{\text{null}}(p_2 \leq p^* | p_2 > .1/10) \\ &= \frac{p^* - .1/10}{1 - .1/10} \\ \Rightarrow p^* &= 0.021. \end{aligned} \tag{2.5}$$

Intuitively, the Holm testing level is an allocation of error probability. The rejection threshold with error probability .1/9 is not .1/9 given the failed test on the first pass. Revisiting Holm uses rejection thresholds which account for testing hypotheses multiple times. It formally revisits hypotheses and is named for this characteristic. (Foster and Stine, 2008) note that this procedure produces thresholds similar to BH, while this paper extends their discussion to model selection and points to additional benefits. The practical benefits of this approximation algorithm and subsequent improvements are discussed in Johnson et al.

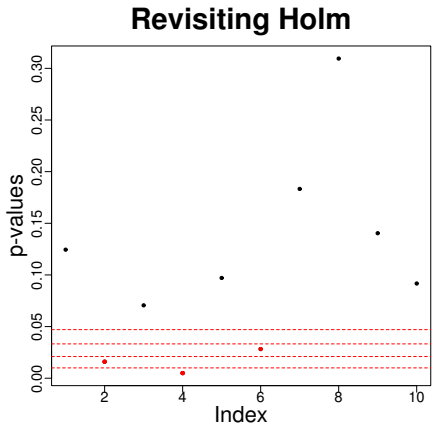


Figure 2: Example of "Revisiting" Holm procedure.

(2015b).

For clarity, Figure 2 shows a few steps of RH. Our hypothetical data follows the simple simulation example with 10 orthogonal explanatory variables and $\alpha = .1$. The rejection thresholds during the first four passes are the horizontal, dashed lines. The first step of the procedure tests all p-values at level .1/10. As one p-value falls below this threshold, its hypothesis is rejected and the procedure continues. Step 2 tests the remaining hypotheses at level-.1/9 which leads to a rejection threshold of .021. One p-value is below this threshold, so its hypothesis is rejected and the procedure continues. Step 3 tests the remaining hypotheses at level-.1/8 and rejection threshold .033, which leads to a third rejection. Step 4, however, fails to make any rejections using a rejection threshold of .047. Therefore the algorithm terminates, resulting in the model selected during the first 3 steps: features 2, 4, and 6.

If only one hypothesis is rejected per round, then RH exactly replicates the forward stepwise selection path. To relax this assumption, suppose that both $p_{(1)} < \alpha/m$ and $p_{(2)} < \alpha/m$ such that both hypotheses would be rejected on the first pass. Since $H_{(1)}$ was rejected, $H_{(2)}$ could be tested at level-$\alpha/(m-1)$. Such a test has higher power, but was ultimately unnecessary; the conservative test made in the first round successfully rejected $H_{(2)}$.

This early rejection results in two effects. First, the early rejection changes the reference distribution for the second testing pass. RH is not attempting to test all possible $H_{(2)}$ on the second pass, but merely those that were not rejected on the first pass. Therefore, the simulated total-selective critical value for RH should only be computed from the subset of cases in which $H_{(i)}$ was not already rejected in a previous pass. This excludes those cases where $p_{(i)} < t^*_T(i-1)$. The result of this change is seen in Table 3. RH produces critical values that are effectively identical to the simulated total-selective values $t^*_T(i)$.

Second, if multiple hypotheses are rejected in a round, RH is not guaranteed to have selected the most significant feature first. Since the features were not truly sorted, it is unknown which of the two hypotheses rejected in the first pass actually had a smaller p-value. Both p-values were merely smaller than α/m. In this case, the order in which the hypotheses are tested is influential. If $H_{(2)}$ is tested before $H_{(1)}$, then RH includes the corresponding features $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$ in the wrong order.

Selecting features in the wrong order is not of serious concern in the orthogonal case, because the same set of features will have been selected by the end of each testing pass. In nonorthogonal settings, however, test statistics change based on the model in which they are computed, so selecting features in a different order can lead to significantly different models. This is easiest to see by example. Table 4 gives the sequential p-values of all features in the prostate cancer data in different selected models. Two algorithms are compared: RH testing the features in sorted stepwise order 1-8 (RH-sort), and RH testing the features in the reverse order 8-1 (RH-rev). The reverse order provides a worst-case ordering for RH. In the table, hyphens indicate the features in the model.

Forward stepwise, RH-sort, and RH-rev consider the same p-values initially (step 0), as no features have been added to the model. These p-values are computed from simple regressions between the response and the feature of interest using an independent estimate of the error variance. While all features fall below the RH threshold, lcavol has the lowest p-value. Therefore, forward stepwise and RH-sort select the same feature on step 1. The

Table 4: Stepwise p-values after each step.

Feature (Step)	RH-0	RH-sort-1	RH-sort-2	RH-rev-1	RH-rev-2	RH-rev-3
lcavol (1)	0.0000	-	-	0.0000	0.0000	0.0000
lweight (2)	0.0000	0.0003	-	0.0000	0.0000	0.0000
svi (3)	0.0000	0.0410	0.0424	0.0000	0.0001	0.0000
lbph (4)	0.0006	0.0041	0.1506	0.0010	0.0001	-
pgg45 (5)	0.0000	0.1453	0.0758	0.0002	0.1330	0.1078
lcp (6)	0.0000	0.7300	0.9494	0.0000	-	-
age (7)	0.0027	0.7998	0.4649	0.1352	0.1331	0.7534
gleason (8)	0.0000	0.6516	0.3592	-	-	-

p-values in the column RH-sort-1 are the stepwise p-values given that lcavol is in the model. Again, RH-sort and forward stepwise select the same variable, lweight, at the second step. Adjusting the stepwise p-values for the model (lcavol, lweight) results in the column RH-sort-2. All of these p-values fall above the RH threshold for the third testing pass, so the procedure terminates. The correspondence between RH-sort and forward stepwise seen here is a general property: if RH tests variables in the order determined by stepwise, then RH selects variables in the same order as stepwise.

RH-rev behaves significantly differently than RH-sort and forward stepwise. The initial p-values it considers are identical, but RH-rev tests gleason first and the test is rejected. The p-values in the column RH-rev-1 condition on gleason being in the model. Proceeding in the reverse order, the test of age is not rejected, but the test of lcp is. Column RH-rev-2 updates the stepwise p-values given the model contains gleason and lcp. Using these p-values, lbph is also rejected, and the process continues. In fact, RH-rev rejects all 8 features. Given the ordering of the features this is at least justifiable. Each subsequent feature explains a significant reduction in ESS. Even after several features are in the model, lcavol provides unique information about the response. That being said, selecting all 8 features is clearly not desirable and motivates the relaxations of Section 2.2.2. Alternatively, by choosing a different set of rejection thresholds Johnson et al. (2015b) mimic stepwise regression very well. In this case, their method selects the RH-sort model of {lcavol, lweight} regardless of the ordering of the features.

In nonorthogonal data, one may object to the updating in equation (2.5) because sequential p-values are relevantly different between steps. While the same explanatory feature is being tested, the sequential p-value is a function of the other variables in the model. For example, if other hypotheses were rejected between two tests of H_i, there is, in general, no guarantee that the conditioning statement in equation (2.5) is accurate; a feature can become *more* significant in the presence of other features. This is seen in Table 6 and discussed at length in Johnson et al. (2015c). Three considerations alleviate this concern. First, the extent to which p-values can change in the presence of other variables is controlled by the approximate submodularity of the data Johnson et al. (2015b). In fact, the statement is conservative if the data are submodular. The degree of approximate submodularity can be bounded by the minimum eigenvalue of the covariance matrix of $\mathbf{X}$. Dependence is not measured by the correlation between variables because not all correlation is problematic for model selection. Second, the simulation in Section 2.3.2 demonstrates that deviations do not harm type-I error control. Third, the significance thresholds developed in Johnson et al. (2015b) render the critique obsolete, in that the correction in equation (2.5) makes a minuscule change to the effecting testing level. The correction can be ignored with a minimal reduction in power.

We provide two simulated examples using correlated data to demonstrate the behavior of RH in more difficult scenarios. RH provides a good approximation of true critical values under mild dependence, but extreme dependence degrades the approximation. As shown in Section 2.3.2, even conservative competitors can be broken in these cases. We directly simulate the distribution of 10 t-statistics as done in the original simulations of this section. The first simulation example has a benign correlation structure where the correlation between indices i and j is $.2^{|i-j|}$. The minimum eigenvalue of the corresponding data matrix is .68.

The second simulation example has a challenging correlation structure in which the correlation between indices i and j is $.8^{|i-j|}$. This results in the strange effects in Table 6, where $t^*_T(2) > t^*_T(1)$. This is counter-intuitive since forward stepwise selects the most significant

Table 5: Benign correlation structure with minimum eigenvalue .68.

	1	2	3	4	5
Fixed-selective $t_F^*(i)$	2.56	1.93	1.58	1.33	1.12
"True" Total-selective $t_T^*(i)$	2.56	2.43	2.30	2.20	1.98
R-Holm Total-selective $t_T^*(i)$	2.56	2.32	2.15	2.01	1.90
R-Holm	2.58	2.31	2.13	1.99	1.86

feature on the first step. Therefore, given this one feature model, it is possible for a feature to appear even more significant than the initially most significant feature. This reversal produces effects such as insignificant steps followed by significant ones. While RH does not approximate the total-selective critical value as well, the deviations are not extreme. This allows it RH to maintain performance measures in the simulations of Section 2.3.2.

Table 6: Challenging correlation structure with minimum eigenvalue .18.

	1	2	3	4	5
Fixed-selective $t_F^*(i)$	2.39	2.04	1.79	1.57	1.38
"True" Total-selective $t_T^*(i)$	2.39	2.85	3.17	3.39	3.41
R-Holm Total-selective $t_T^*(i)$	2.39	2.51	2.27	2.10	1.95
R-Holm	2.58	2.31	2.13	1.99	1.86

2.2.2. Relaxations

The dependence of RH on the order in which features are tested may be impractical or unsatisfactory for some applications. This subsection introduces two relaxations that can be validly used to identify a model using a stepwise table such as Table 1. The first relaxation makes the tacit assumption that only one hypothesis is rejected per round. In this case, RH selects the same features in the same order as forward stepwise. Furthermore, the RH critical values accurately describe the forward stepwise table. The approximate procedure, aRH, operates as follows:

1. Compute a stepwise table where the sequential p-value for step i is p_i. The p-values are not necessarily sorted: p_i need not be less than p_j if $i < j$.
2. Compute the set of RH critical values c_{rh}:

$c_{rh}(1) = \alpha/m$,

$c_{rh}(i) = c_{rh}(i-1) + \alpha/(m-i+1) - c_{rh}(i-1)\alpha/(m-i+1)$ for $i > 1$.

This follows from the conditioning in equation (2.5).

3. Identify $k = \min_i p_i > c_{rh}(i)$.
4. Reject $H_{(1)}, \ldots, H_{(k-1)}$.

Foster and Stine (2008) point out that c_{rh} are initially close to the Benjamini-Hochberg thresholds $i\alpha/m$. Conjecture 1 suggests that aRH controls mFDR and is substantiated by the simulations in Section 2.3.2. Clearly the claim is true if only one sequential p-value is rejected per round as aRH and RH coincide. The proof of this claim is less important as further revisions to RH yield a revisiting procedure, Revisiting Alpha-Investing, which is proven to closely mimic forward stepwise (Johnson et al., 2015b). The authors also explain and develop further practical benefits of using a precise threshold approximation to stepwise regression.

A final relaxation is of independent theoretical and practical interest. Given the discussion at the end of Section 2.2.1, one may question the update implied by revisiting in equation (2.5). In particular, if the data is not submodular, then there is no guarantee that a p-value does not decrease in the presence of other variables. A conservative statement is provided by ignoring the revisiting component of RH. Namely, merely use the Holm levels as the rejection thresholds for each testing pass on a given stepwise table. The resulting procedure, Stepwise-Holm (SH), proceeds as follows:

1. Compute a stepwise table where the sequential p-value for step i is p_i. The p-values are not necessarily sorted: p_i need not be less than p_j if $i < j$.
2. Identify $k = \min_i p_i > \alpha/(m-i+1)$.
3. Reject $H_{(1)}, \ldots, H_{(k-1)}$.

SH extends the intuition behind the Max-|t| procedure of Buja and Brown (2014). Under the assumption of submodularity, this procedure controls the FWER. The proof of Theorem

2 is given in the Appendix. While this controls false rejections in a conservative way, SH performs well in the simulations of Section 2.3.2.

As pointed out in Taylor et al. (2014), the conservatism of SH is in part due to the ranking of test statistics, ie $|t_{(3)}|$ is constrained to be less than $|t_{(2)}|$. That being said, their discussion is incomplete because it ignores the testing effect. Furthermore, as shown in Section 2.2.1, the t-statistics of subsequent steps need not be smaller than those of previous steps in non-orthogonal cases. This complicates the analysis because merely observing an insignificant variable is not sufficient indication that there is no signal left in the data. Such instances form the core problem cases for feature selection algorithms and are discussed at length in Johnson et al. (2015c).

2.2.3. Alpha-Investing

RH is closely connected to commonly used multiple comparison procedures which alludes to its type-I error control. This control is proven by demonstrating that the procedure is an alpha-investing rule (Foster and Stine, 2008). Alpha-investing rules are similar to alpha-spending rules in that they are given an initial amount of alpha-wealth which is spent on hypothesis tests. Wealth is the total allotment of error probability. Bonferroni allocates this error probability equally over all hypothesis, testing each one at level α/m. In general, the amount spent on tests can vary. If α_i is the amount of wealth spent on test H_i, FWER is controlled when

$$\sum_{i=1}^{m} \alpha_i \leq \alpha.$$

In clinical trials, alpha-spending is useful due to the varying importance of hypotheses. For example, many studies include both primary and secondary endpoints. The primary endpoint of a drug trial may be determining if a drug reduces the risk of heart disease. As this is the most important hypothesis, the majority of the alpha-wealth can be spent on it, providing higher power. There are often many secondary endpoints such as testing if the drug reduces cholesterol or blood pressure. Alpha-spending rules can allocate the

remaining wealth equally over the secondary hypotheses. FWER is controlled and the varying importance of hypotheses is acknowledged.

Alpha-investing rules are similar to alpha-spending rules except that alpha-investing rules earn a return, or contribution to their alpha-wealth, of $\omega \leq \alpha$ when tests are rejected. Therefore, the alpha-wealth after testing hypothesis H_i is

$$W_{i+1} = W_i - \alpha_i + \omega R_i$$

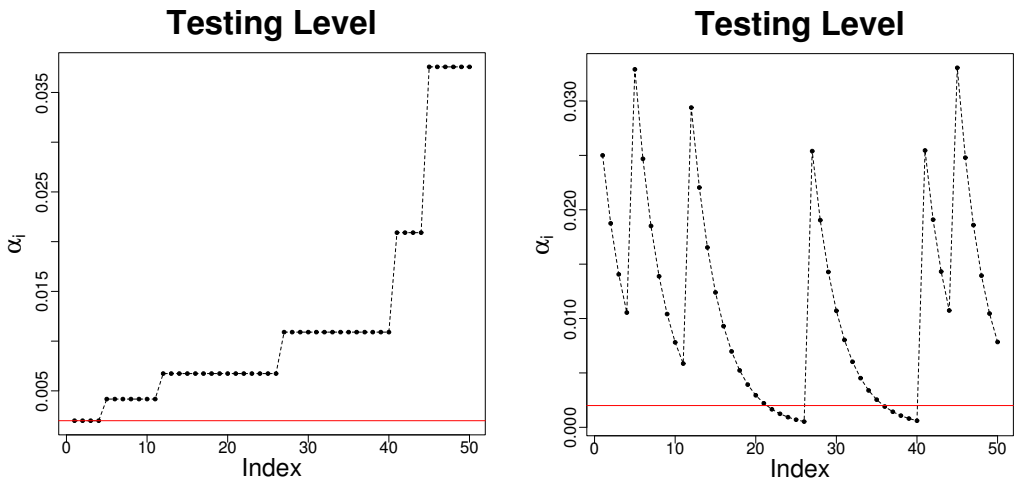
An alpha-investing strategy uses the current wealth and the history of previous rejections to determine which hypothesis to test and the amount of wealth that should be spent on it.

Intuitively, alpha-investing rules spend error probability in search of false null hypotheses to reject. Each false null that is rejected allows α more incorrect rejections in expectation. Alpha-investing rules merely need to spend more wealth (error probability) than the probability of error they incur. In some sense, this behavior is present in all procedures which control a proportion of false rejections. For example, if it is known that the first 9 rejections were of false hypotheses, then any 10th hypothesis can be rejected while controlling the proportion of false rejections at .1.

We provide two examples of potential spending rules. The first is similar to Bonferroni in that wealth is spent evenly over all remaining hypotheses. Note that this requires the number of hypotheses to be known in advance. Given current wealth W_i, this procedure spends

$$\alpha_i = \frac{W_i}{m - i + 1}$$

to test H_i. Figure 3a shows the testing levels of this rule with starting wealth $W_1 = .1$ when testing a set of 50 hypotheses in which rejections are made on tests 4, 11, 26, 40, and 44. The procedure begins by testing at the usual Bonferroni level with $m = 50$, but rejections allow more wealth to be spent, increasing the power of the tests.



(a) Bonferroni: $\alpha_i = \frac{W_i}{m-i+1}$ (b) Geometric: $\alpha_i = \frac{W_i}{4}$

Figure 3: Example alpha-investing rules with testing levels. Tests are ordered numerically and rejections are made at indices 4, 11, 26, 40, and 44.

The second example is a geometric spending rule that allows the total number of hypotheses to be unknown. In this case, the alpha-investing rule merely spends one quarter of its current wealth on the current test: $\alpha_i = W_i/4$. This spends wealth rapidly following a rejection and is sensible if false hypotheses are anticipated to arrive in groups. Figure 3b shows the testing levels in the same scenario considered for the Bonferroni strategy. One important difference between the procedures is that the Bonferroni rule spends all of its wealth by the end of the 50 tests whereas the geometric rule does not. The geometric procedure is also related to a universal strategy. If tests only arrive sequentially without being revisited, there exists a strategy that cannot be outperformed by a significant margin by any other strategy. This universal procedure is an adaptively weighted set of geometric strategies (Foster et al., 2001).

Viewing the Holm step-down procedure as an alpha-investing rule creates the Revisiting Holm procedure and proves its control of mFDR. Given initial alpha-wealth α and return $\omega = \alpha$, test all hypotheses at the Bonferroni level, α/m. This exhausts all alpha-wealth, so that the procedure terminates if no rejections are made. If a rejection is made, the

procedure earns a return equal to α and only $m-1$ hypotheses remain. The wealth is again split evenly among all remaining hypotheses, yielding the Bonferroni threshold over $m-1$ hypotheses of $\alpha/(m-1)$. While Holm controls FWER, conditioning on the result of previous tests improves power and controls mFDR. If any round is conducted without any rejections, then the procedure is out of wealth and terminates.

RH accounts for the effects of inference during model selection. It controls the rank effect by spending alpha-wealth on all hypotheses. This is an important change in perspective for multiple comparisons research: the effects of multiple comparisons are controlled by paying alpha-wealth to perform a given operation. For example, forward stepwise sorts the stepwise p-values and only tests the minimum; however, this sort operation requires considering all hypotheses, and is formally paid for using alpha-spending. Furthermore, RH controls the testing effect because the amount of alpha-wealth spent on tests is larger than the rejection threshold due to revisiting. If a hypothesis is tested twice, initially at level α_1 and then at level α_2, the rejection threshold is

$$\begin{aligned} \alpha_2 &= \mathbb{P}_{\text{null}}(p \leq p^* | p > \alpha_1) \\ &= \frac{p^* - \alpha_1}{1 - \alpha_1} \\ \Rightarrow p^* &= \alpha_1 + \alpha_2 - \alpha_1\alpha_2. \end{aligned} \tag{2.6}$$

This threshold is smaller than $\alpha_1 + \alpha_2$, which is the rejection threshold if the total wealth spent testing this hypothesis was used on one test instead of two.

2.3. Polyhedral Selection

One attempt so solve the problems brought up in Section 2.1.1 is provided by Taylor et al. (2014). The authors formalize the steps of a selection procedure as constraints on the response in order to accounts for the selection effect. For ease of exposition, we focus on the forward stepwise case, though the arguments are also applicable to LARS (Efron et al., 2004). The authors' propose an "Exact Forward Stepwise" procedure (eFS) that assigns

new, "exact" p-values to the variables in a standard forward selection algorithm. After a variable is added, it is assigned a "p-value" by this "exact" procedure. This is a numerical quantity that has a U(0,1) distribution conditional on the sign of the selected variable and the variables that have been previously chosen.

While forward stepwise and LARS operate independently of these p-values, one would expect the modeler to want to use the p-values to determine the step at which to "stop" the procedure and provide a final model. Consider the p-values given in Table 1, which compares their eFS p-values to naive stepwise p-values. Identifying a final model using such a table requires considering multiple p-values from separate steps of the procedure. Therein lies the problem: the set of exact p-values cannot be used to make decisions, else they are invalid. Even using these p-values as input into a secondary FDR-controlling procedure as in G'Sell et al. (2015) is inappropriate. Only one exact p-value can validly be used, testing one step of a much larger procedure. Similarly, if a model is selected through other means such as cross-validation, the inferential guarantees of related methods need not hold (Bachoc et al., 2014).

The conventional p-values are single-step values. They do not correct for the multiple testing nature of a stepwise procedure. Later in this section, we recommend the procedures from Section 2.2 which can be validly and directly used for stepwise selection. Alternatively, one could update the conditional inference logic of Taylor et al. (2014) to account for the testing effect. For reasons discussed below, however, we do not favor its use.

The paradox in using the eFS p-values is rather subtle, and is easiest to explain in the context of an example. Let Z_i be independently distributed $N(\theta_i, 1)$, for $i \in \{1, 2\}$. The forward selection problem is equivalent to determining an order for testing $H_{0,i}$: $\theta_i = 0$, while controlling false rejections at level α. Allowing correlated variables does not change our discussion, it merely complicates the exposition and graphs. Similarly, without loss of generality, let $Z_1 > Z_2 > 0$.

The authors' eFS significance thresholds are given as "eFS Step 1" and "eFS Step 2" in Figure 4. The conditioning set for both steps of the procedure is the same: $\{Z_1 > Z_2 > 0\}$. Values to the right of the curve "eFS Step 1" (in red) yield p-values below α when testing $H_{0,1}$ while values between "eFS Step 1" and the blue, 45° line yield p-values greater than α. Thus, values to the right of eFS Step 1 are those for which the statistician using eFS p-values would select Z_1 with a positive sign at the first step of the selection process. During the second step, values above the curve "eFS Step 2" (in gold) are significant at level α, while values below are not. Note that the calculation at the second step does not change depending on whether or not $H_{0,1}$ was rejected.

In order to use the eFS p-values as Table 1 would imply, testing $H_{0,2}$ must account for rejecting $H_{0,1}$ (the testing effect). Following the methodology of the authors' paper, this requires updating the conditioning set. Our corrected procedure, "Exact Stepwise (conditional)" (ES-c), terminates on the first step in which a corrected, conditional p-values is above α. If $H_{0,2}$ is only tested when $H_{0,1}$ is rejected by eFS, then the conditioning set is the region to the right of eFS Step 1. Those points to the right of eFS Step 1 and outside the convex, parabolic region whose boundary is the curve "ES-c Step 2" (in green) are those for which the new ES-c procedure selects Z_1 at the first step (with a positive sign) and Z_2 at the second step (with positive sign). It is clear that this correction does not invalidate the authors' methodology, but it does yields different p-values. Furthermore, the new conditioning sets are not polyhedral and need not be convex.

2.3.1. Stopping Procedures Using ES-c P-values

We are also concerned with the counter-intuitive results given when using conditional p-values, even when they are corrected as above. The problem has obvious symmetries such as relabeling variables 1 and 2 or changing their signs. While our new proposal, ES-c, preserves those symmetries, it does not preserve the natural monotonicity of the problem. For example, there exist values (z_1, z_2) and (z_1', z_2') for which $z_1 \leq z_1'$ and $z_2 \leq z_2'$, but for which ES-c selects both variables at (z_1, z_2) and no variables at (z_1', z_2'). The authors' eFS

procedure does not produce as extreme an example since $H_{0,2}$ is tested regardless of the result of testing $H_{0,1}$; however, the significance of the test of $H_{0,1}$ depends on the value of Z_2. This is particularly troubling given that Z_1 and Z_2 are independent.

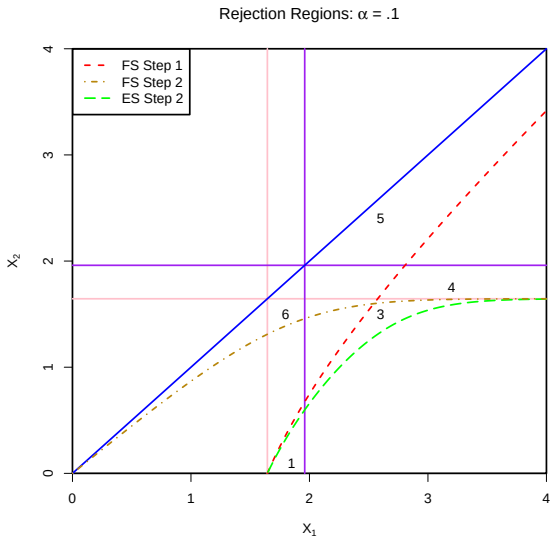


Figure 4: Stepwise rejection regions at $\alpha = .1$. The full picture is symmetric around the x- and y-axes. A corresponding image would be drawn if $Z_2 > Z_1 > 0$, in which case the graph would be rotated 90° and maintain its symmetries.

It is also instructive to compare the rejection regions of the eFS and ES-c procedures to those of more traditional methods (again, see Figure 4). The conventional procedure controls the nominal type-I error rate, ignoring the selection of M and H_0. It first adds Z_1 if $|Z_1| > |Z_2|$ and $|Z_1| > \Phi^{-1}(1-\alpha/2)$. It then adds Z_2 if, also, $|Z_2| > \Phi^{-1}(1-\alpha/2)$. If $|Z_2| > |Z_1|$ and $|Z_2| > \Phi^{-1}(1-\alpha/2)$ the first step adds Z_2, etc. Relevant portions of the lines in pink form the boundaries of this region. There are multiple, classically-constructed stepwise regions to control for selection and multiple comparisons:

a) Conventional: The pink lines are at $Z_1 = 1.645$ and $Z_2 = 1.645$. To the right of the blue 45° line, they show regions where the fully classical stepwise procedure would first choose Z_1 (to the right of $Z_1 = 1.645$) and then Z_2 (above $Z_2 = 1.645$). This region is not adjusted for multiple comparisons, and hence has an error probability of choosing non-empty models under the null hypothesis that exceeds the nominal level α.

b) Bonferroni (B): Similarly, the purple lines at $Z_1 = 1.96$ and $Z_2 = 1.96$ bound regions which provide conservative multiple-comparison adjustments based on conventional single-coordinate p-values. This uses the Bonferroni approximation to control for multiple-comparisons. The exact numerical value can be computed from a maximum modulus calculation and is 1.948.

c) Stepwise Holm (SH): A better stepwise procedure controlling for multiple comparisons can be constructed as follows: at each step, choose among the remaining k variables using a p-value threshold such that the null probability of choosing any model is less than or equal to α. As in b), Bonferroni yields the conservative threshold α/k, though an exact calculation is possible when k is small. On the figure, one would include Z_1 to the right of $Z_1 = 1.96$ (or 1.948 for an exact calculation) and then would include Z_2 when Z_2 is above $Z_2 = 1.645$. At each step of the procedure, the conditional probability under the null hypothesis of continuing with an incorrect rejection is α. This type of procedure was briefly proposed in Buja and Brown (2014) as Max-|t|. If Bonferroni is used to pick a conservative threshold at each step, then the resulting procedure is Stepwise Holm from Section 2.2.2. If the goal is to preserve mFDR, then Revisiting Holm or Approximate Revisiting Holm can be used to achieve substantially higher power.

Many interesting comparisons can be made between eFS, ES-c, and the more conventionally motivated, multiplicity-corrected procedures B and SH. These regions are labeled (numerically) in Figure 4.

1. Consider the triangular region to the right of eFS Step 1 and to the left of $Z_1 = 1.96$. This is where the ES-c procedure selects Z_1 and the conventionally motivated procedures choose no variables. Heuristically, this seems to be a success for the ES-c procedure.
2. Within the region described in 1, there is a sliver between eFS Step 1 and ES-c Step 2. Here, ES-c selects both Z_1 and Z_2, while the conventional procedures select neither. While this maintains a significance guarantee, this may not be an advantage. These points do have conventional (2-sided) p-values for Z_1 that are below α, but the

conventional p-value for Z_2 is quite large. Selecting Z_2 appears to be a mistake.

3. There is a more noticeable triangular region bounded by eFS Step 1, ES Step 2, and $Z_2 = 1.645$. In this region, the ES-c procedure selects both Z_1 and Z_2 but B and SH select only Z_1. For reasons similar to those in 2, the second step of ES-c appears undesirable. The disadvantage is not as clear, however, since Z_2 can have a 2-sided p-value as small as $\alpha = .1$ in this region. The uncorrected FS p-value yields intuitively more satisfactory results in this region.
4. Consider the region between $Z_2 = 1.645$ and $Z_2 = 1.96$, and to the right of eFS Step 1. ES-c and SH select both variables, but the simpler, conservative procedure B does not include Z_2. The advantage here goes to ES and SH.
5. The area between the 45° line and eFS Step 1 and to the right of $Z_1 = 1.96$ is where B and SH have a clear advantage in power relative to ES-c or to a procedure based on eFS. In those regions, ES-c and eFS have first step p-values above α and hence do not select any variable, while B and SH always select Z_1 and often select Z_2.
6. In the region above eFS Step 2 and below $Z_2 = 1.645$, Z_2 has a significant eFS p-value even though its conventional p-value can be close to 1 near the origin.

In summary, B or SH seem preferable to the ES-c procedure. The latter does better if the data fall in the small, but not negligible region 1; however, B and SH produce much more reasonable models in the more noticeable area 5. Procedure SH is preferred to B because of the difference noted in region 4. The regions 2 and 3 are quite small and nearly negligible in probability. While the ES-c procedure seems undesirable on these regions, the concern is not important.

As a further comparison, consider the point (4, 3.8). The ES-c p-values for step 1 and 2 are $\approx .44$ and $\approx .0001$ respectively, while the naive p-values are approximately .0001 for both variables. The decision of ES-c to stop at step 1 and declare an empty model might well be viewed as embarrassing and subjectively undesirable. This is consistent with the claimed ES-c p-values though. Similarly, while methods of G'Sell et al. (2015) are not required to

stop at step 1, the penalty for continuing is extremely large given the unconventionally large p-value.

In the correlated setting, the interesting simulation in Taylor et al. (2014) strongly suggests that the p-values used in procedures B and SH can be extremely conservative. The conservatism is the result of different objects of inquiry: both B and SH control the FWER as opposed to the FDR control provided by G'Sell et al. (2015). Furthermore, the simulations of Section 2.3.2 demonstrate that SH performs well and has uniformly higher power than the G'Sell et al. (2015) procedures. This is particularly interesting since SH provides stronger control over false rejections.

For all considered testing methods, when the regressors are correlated the values of regression coefficients depend on which other coefficients are in the current model. Hence a coefficient may have a non-zero value within the currently active set of variables; and so be correctly included into that model at that step. Within a later active model it might then have a value of 0. Thus a correct selection at a given step may become "incorrect" as the process proceeds, and vice-versa. The related phenomenon of suppression can yield a series of insignificant steps followed by highly significant steps (Johnson et al., 2015c). These issues have important consequences for interpretation of p-values produced in a stepwise routine. Such issues do not occur in the simple model at hand involving independent variables with fixed mean values.

2.3.2. Performance Comparisons

While we have raised some concerns about the accuracy of the "exact" description of the eFS p-values, one could still use them for model selection. This is particularly salient since the package reports a statistic that does precisely that. Furthermore, as our proposed alternative, Revisiting Holm, is only exact under orthogonality, we are interested in its performance when that assumption is violated. The simulations below indicate that eFS p-values do, in most cases, yield FDR control when coupled with appropriate model selection

methods. Given the regions where the eFS p-values are surprisingly small, we conjecture that a reasonable example could violate the FDR guarantees; however, the lack of power seen by the eFS p-values in region 5 as well as the conservative nature of the G'Sell et al. (2015) procedures provide some protection against this.

Given a full set of p-values as those in Table 1, selecting a model using hypothesis testing requires rejecting an initial contiguous set of hypotheses. If hypotheses are ordered numerically, H_2 and H_4 cannot be the only rejections. The sets $\{H_1, H_2\}$ or $\{H_1, \ldots, H_4\}$ are possible rejection sets that identify forward stepwise models. As the eFS p-values are not necessarily sorted by size as required by the step-up BH procedure, controlling FDR under this constraint is nontrivial. G'Sell et al. (2015) transform p-values such that they are ordered and uses step-up BH on the transformed p-values. The model selection criteria we consider is Forward Stop, which rejects hypotheses $H_1, \ldots, H_{\hat{k}}$ where

$$\hat{k} = \max_{k \in \{1,\ldots,m\}} -\frac{1}{k} \sum_{i=1}^{k} \log(1 - p_i) \leq \alpha.$$

Our simulated data has 200 observations, $m \in \{20, 50\}$ covariates with correlation between covariates i and j given by $\rho \in |i - j|$, $\rho \in \{0, .2, .8\}$. We consider 10 nonzero coefficients in both high- and low-signal cases. The high-signal case sets $|\beta_i| = i/6$ for $1 \leq i \leq 10$ and the low-signal case sets $|\beta_i| = \sqrt{2\log(p)}/\sqrt{n}$ for $1 \leq i \leq 10$. True covariates in the high-signal case have t-statistics in the true model in the range $[3, 24]$. The low-signal case produces t-statistics in the true model in the range in the range $[1, 3.5]$, which is close to the RIC threshold (Foster and George, 1994). As a final complication, the signs of the nonzero coefficients are either all positive or equal to $(-1)^i$. This is noted in Table 7 as "+" or "+/-". Comparisons are made between forward stepwise models selected via four procedures: Holm (equivalently Max-t or SH), Approximate Revisiting Holm (aRH), Revisiting Holm (RH), and Forward Stop (FStop). The SH, aRH, and RH procedures use traditional, stepwise p-values, while FStop uses the eFS p-values. Since RH depends on the order in which features

Table 7: Simulation results for Holm, Revisiting Holm, and Forward Stop selection rules.

				FDR				Power			
m	β	Sign	ρ	SH	aRH	RH	FStop	SH	aRH	RH	FStop
20	High	+	0	0.004	0.035	0.037	0.000	0.921	0.950	0.951	0.557
			.2	0.001	0.005	0.040	0.000	0.900	0.919	0.953	0.553
			.8	0.001	0.004	0.321	0.000	0.812	0.846	0.920	0.572
		+/-	0	0.000	0.005	0.036	0.000	0.895	0.921	0.954	0.560
			.2	0.001	0.004	0.039	0.000	0.873	0.907	0.949	0.583
			.8	0.003	0.004	0.288	0.001	0.561	0.626	0.767	0.212
	Low	+	0	0.006	0.007	0.029	0.000	0.143	0.178	0.364	0.099
			.2	0.001	0.005	0.030	0.000	0.254	0.318	0.539	0.102
			.8	0.005	0.008	0.222	0.001	0.293	0.307	0.811	0.168
		+/-	0	0.004	0.006	0.029	0.000	0.143	0.176	0.367	0.099
			.2	0.009	0.009	0.031	0.001	0.101	0.104	0.123	0.093
			.8	0.008	0.008	0.045	0.005	0.065	0.065	0.009	0.065
50	High	+	0	0.000	0.003	0.049	0.000	0.870	0.897	0.940	0.331
			.2	0.000	0.001	0.052	0.000	0.874	0.893	0.936	0.361
			.8	0.001	0.002	0.313	0.000	0.769	0.801	0.903	0.368
		+/-	0	0.000	0.001	0.048	0.000	0.862	0.884	0.934	0.347
			.2	0.000	0.002	0.049	0.000	0.830	0.861	0.931	0.328
			.8	0.003	0.006	0.302	0.001	0.475	0.531	0.688	0.158
	Low	+	0	0.004	0.006	0.044	0.000	0.124	0.139	0.328	0.098
			.2	0.004	0.005	0.048	0.000	0.214	0.256	0.535	0.101
			.8	0.003	0.005	0.222	0.000	0.294	0.306	0.826	0.148
		+/-	0	0.004	0.005	0.039	0.000	0.125	0.141	0.326	0.098
			.2	0.009	0.009	0.049	0.001	0.093	0.095	0.100	0.091
			.8	0.004	0.004	0.048	0.002	0.042	0.042	0.005	0.042

are tested, we provided a worst case ordering. Features are tested from $\mathbf{X}_m$ to $\mathbf{X}_1$. This not only provides the greatest chance for false rejections, but makes $\mathbf{X}_{11}, \ldots, \mathbf{X}_m$ as significant as possible in high-signal, high-correlation settings. Both power and FDR is measured over 1000 simulation repetitions. Results are given in Table 7.

There are many messages conveyed by Table 7. First, while SH is conservative, it still performs very well. This is especially telling as Taylor et al. (2014) demonstrate that the test statistic is highly conservative after multiple steps. It has significantly higher power than FStop, by as much as a factor of three, as is never outperformed by FStop. It may be surprising that in many cases there is not a large difference between SH and RH.

Their similarity is due to the small number of non-zero coefficients. The corrections due to revisiting in equation (2.5) can be quite small. In some low-signal cases, however, RH achieves much higher power. This is expected as it controls mFDR instead of the FWER.

The RH procedure has the highest power of all of the methods by a significant margin, even in cases in which its corrections are not exact. In low signal cases, this comes at a cost of having the highest FDR of the considered algorithms. In most highly correlated cases, RH is significantly higher than the desired FDR bound. This is large due to the worst-case ordering of the hypotheses, which can be improved using different thresholds (Johnson et al., 2015b). There is a necessary trade-off between power and FDR in this setting since all algorithms are considering the stepwise model: the only way to increase power is to create larger models, which in turn may result in more false rejections. That being said, in the majority of cases FDR is still controlled at level-.05, demonstrating the similarity between mFDR and FDR. Positively correlated variables with direct effects of mixed sign creates the most challenging cases for feature selection algorithms as the data is highly non-submodular. For a complete discussion of this, see Johnson et al. (2015c).

While FDR does provide a measure of false rejections in this setting, it is somehow inappropriate. It is difficult to disentangle the meaning of false rejections given that the stepwise model is being tested. If forward stepwise identified a highly significant variable on the first step which is actually not in the true model, then the false rejection is the fault of stepwise, not the hypothesis testing procedure. Furthermore, one may be inclined to claim that it is not a false rejection at all! Given the high correlation between features in real data, the tested feature may account for a statistically significant portion of the variability in Y in the model in which it is considered. Addressing this requires a difference in perspective on false rejections. Alternative definitions motivated by such considerations have been proposed by (G'Sell et al., 2013).

The conservativeness of FStop warrants further discussion. The transformation performed by FStop can be interpreted through the lens of a sequential revisiting procedure. Suppose

that we repeatedly test a hypothesis H_0 at level $\delta > 0$. Consider the amount of wealth spent to reject H_0 if its p-value is p_0. Each failed rejection implies that p_0 is in the upper $(1-\delta)$ portion of its feasible region, which is initially $[0,1]$. If H_0 is rejected after q tests, a Taylor approximation provides

$$\begin{aligned}(1-\delta)^q &= 1-p_0 \\ \Rightarrow q\delta &\approx -log(1-p_0).\end{aligned}$$

While H_0 could have been rejected by spending p_0, $-log(1-p_0) > p_0$ was spent on the rejection. If p_0 is small, the amount of alpha-wealth wasted by revisiting is minor, but larger p-values waste significant wealth. This wastefulness is one explanation for the conservative behavior of FStop. Given the eFS p-values are conservative in large regions of the parameter space, it is not surprising that the combined procedure is highly conservative. Less wasteful alpha-investing procedures can be given to rectify this conservativeness, and we are currently working on a paper to describe how to do so.

One may be interested in rectifying the concerns we have raised about the eFS p-values. If a model is selected via FStop, one suggestion would be to include this as a constraint on Y. Even ignoring the computational difficulties of such a task, the enterprise itself is unjustified. FStop is not a stopping rule in the sense of providing a stopping time. As it mimics step-up BH, it is forward looking and cannot be computed until all p-values are known. If one desires a proper stopping rule, FStop can be modified to mimic step-down BH. The revised step-down Forward Stop rejects hypotheses $H_1, \ldots, H_{\hat{k}-1}$ where

$$\hat{k} = \min_{k \in \{1,\ldots,m\}} -\frac{1}{k}\sum_{i=1}^{k} \log(1-p_i) \geq \alpha.$$

2.4. Appendix

The sequential p-values were constructed using data downloaded from Robert Tibshirani's website. The p-values computed in Table 1 are computed from the standard F-test with 1 and 58 = 67 - 9 degrees of freedom. As some additional numerical details, note that the MSE from the full model is $\hat{\sigma}^2 = 0.5074$. Thus, for example, the sequential F-value for testing "svi" is $2.1841/.5074 = 4.305$ with a t-value of $2.075 = \sqrt{4.305}$. This has a p-value with 58 degrees of freedom of 0.0426.

2.4.1. Proof of Theorem 2

While the assumption of submodularity is uncommon in the statistics literature, an equivalent concept has been discussed in the social sciences: the absence of conditional suppressor variables. A full discussion of this connection is given in Johnson et al. (2015c), and only the required implication is provided here. If data is submodular, then the p-values in a stepwise table are necessarily non-decreasing. Clearly, the prostate data from Table 1 is not submodular, however, the majority of steps have non-decreasing p-values. Deviations from submodularity can be captured using notions discussed in Johnson et al. (2015c).

The main challenge in applying multiple comparison proof methods in sequential testing is that test statistics can change between rounds of stepwise regression as seen in Table 4. Under submodularity, however, the sign of the change is fixed. Consider the initial sequential p-values for all features considered marginally. Since this is prior to the first step of stepwise, the p-values will be written as p_i^0, to index the 0th round. Submodularity guarantees that $p_i^0 \leq p_i^j$ for all i and $j > 0$. Consider testing the hypotheses H_i: $\beta_i = 0$ for all i.

Proof of Theorem 2. Holm provides a valid procedure for controlling the FWER in this scenario. Let $I(\beta)$ be the set of true null hypotheses for a given parameter vector β. For all $i \in I(\beta)$, $p_i^0 < p_i^j$ where j is the step at which forward stepwise would select the ith variable

for inclusion into the model. Holm is guaranteed to control FWER when using p_i^0 and the probability of making a false rejection on the jth step is smaller. Therefore, SH controls FWER under submodularity. In fact, it is clear from this discussion that the absence of conditional suppressor variables only needs to hold for the member of $I(\beta)$. This mirrors the use of positive regression dependence in proofs of FDR control Benjamini and Yekutieli (2001). □

2.4.2. Computations for Updated Polyhedral Methods

FS Step 1 (red curve): If Z_1 is chosen before Z_2 with a positive sign, the observation lies in the cone $R_1 = \{Z_1 > Z_2 > 0\}$. In order to have a level α test of H_0: $\theta_1 = 0$ conditional on $(z_1, z_2) \in R_1$ one must have

$$\theta = \mathbb{P}(Z_1 > \tau_1 | (z_1, z_2) \in R_1, Z_2 = z_2) \quad \forall z_2.$$

This entails choosing the point via

$$\alpha = \frac{1 - \Phi(z_1)}{1 - \Phi(|z_2|)}. \tag{2.7}$$

This defines $z_1 = z_1(z_2)$ for the red curve.

FS Step 2 (yellow curve): The conditioning region is the same, so the level α test of H_0: $\theta_2 = 0$ conditional on $(z_1, z_2) \in R_1$ requires

$$\theta = \mathbb{P}(Z_2 > \tau_2 | (z_1, z_2) \in R_1, Z_1 = z_1) \quad \forall z_1.$$

This entails choosing the point via

$$\alpha = \frac{\Phi(z_1) - \Phi(Z_2)}{\Phi(z_1) - 1/2}. \tag{2.8}$$

ES-c Step 2 (green curve): Given H_0: $\theta_1 = 0$ has been rejected, possible values of (Z_1, Z_2)

lie to the right of eFS Step 1. Denote this region as R_2. Now the test $H_0 : \theta_2 = 0$ must satisfy

$$\theta = \mathbb{P}(Z_2 > \tau_2 | (z_1, z_2) \in R_2, Z_1 = z_1)$$

for all z_1 for which the conditioning region is non-empty. The only change from eFS Step 2 is that the conditioned region is a function of z_2. This entails choosing the point $z_2 = z_2(z_1)$ for which

$$\alpha = \frac{\Phi(z_2^*) - \Phi(Z_2)}{\Phi(z_2^*) - 1/2}, \tag{2.9}$$

where z_2^* denotes the value for which $z_1(z_2^*) = z_1$. The computation in equation (2.9) is facilitated by noting that equation (2.7) implies

$$\Phi(z_2^*(z_1)) = 1 + \frac{\Phi(z_1) - 1}{\theta}. \tag{2.10}$$

A final model from the forward stepwise path is often identified using cross-validation or minimizing a criterion such as AIC. The classical rules to stop forward stepwise such as F-to-enter do not control any robust statistical quantity, because attempting to test the addition of such a feature uses non-standard and complex distributions (Draper et al., 1971; Pope and Webster, 1972). Chapter 2 approximates forward stepwise in order to provide valid statistical guarantees. We improve upon this method, ensuring the risk of the selected model is close to that of the stepwise model, even in cases when the models differ (Theorem 3). In this way, our results can be interpreted as type-II error control. We also provide the speed and flexibility to use forward stepwise in modern problems.

There are few provable bounds on how well greedy statistical algorithms perform. Zhang (2008) provides bounds for a forward-backward selection algorithm, FoBa, but the algorithm is slower than forward stepwise, limiting its use. Guaranteeing the success of greedy methods is important as l_1-relaxations of (1.3) such as the Lasso (Tibshirani, 1996) can introduce estimation bias that generates infinitely greater relative risk than l_0-methods (Johnson et al., 2015a). While convex regularizers are computationally convenient, they are less desirable than their non-convex counterparts (Breheny and Huang, 2011). Furthermore, l_1-based methods over-estimate the support of β (Zou, 2006). One potential solution is to use blended, non-convex regularizers such as SCAD (Fan and Li, 2001), to which we will compare out solution in Section 3.2.

Our solution, given in Section 3.1, provides performance enhancing modifications to the valid stepwise procedure identified in Johnson et al. (2016). The resulting procedure, Revisiting Alpha-Investing (RAI), has a performance guarantee that can be interpreted as type-II error control as the procedure is guaranteed to find signal. The guarantees of Section 3.1.2 are closely related to Das and Kempe (2008, 2011) and the classical result of Nemhauser et al. (1978) which states that, under suitable assumptions, greedy algorithms provide a $(1 - 1/e)$

approximation of the optimal solution to (1.2).

Section 3.1 describes how RAI leverages statistical testing to choose the order in which features are added to the model. RAI is a "streaming" procedure that sequentially considers each feature for addition to the model, instead of performing a global search for the best one. A feature merely needs to be significant *enough*, and not necessarily the *most* significant. It is a thresholding approximation to the greedy algorithm (Badanidiyuru and Vondrák, 2014). RAI makes multiple, fast passes over the features. No more than $\log_2(n)$ passes are required, but in practice we find that 5-7 passes are sufficient regardless of sample size. Each testing pass identifies those features that meet a required level of statistical significance. The initial testing pass conducts a strict test for which only extremely significant features are added to the model. Subsequent passes perform a less stringent test. RAI is not guaranteed to pick the most significant feature, only one that is significant enough to pass the test. As such, the final model is built from a series of approximately greedy choices.

The sequential testing framework of RAI allows the order of tested hypotheses to be changed as the result of previous tests. This allows for directed searches for data base queries or identifying polynomials. Section 3.2 leverages this flexibility to greedily search high-order interactions spaces. We provide simulations and real data examples to demonstrate the success of our method.

RAI enjoys three key properties:

1. It is guaranteed to find signal.
2. It does not over-fit the data by controlling type-I errors; few spurious features are selected.
3. It is computationally efficient.

By leveraging Variance Inflation Factor Regression (Lin et al., 2011), if the final model is of size $q << \min(n, p)$, the computational complexity of RAI grows at $O(np\log(n))$. Using the full data requires computing $\mathbf{X}'y$, which takes $O(np)$ time. Therefore, RAI merely adds

a log factor to perform valid model selection.

3.0.3. Notation

We use notation from the multiple comparisons literature given its connection to RAI. Consider m null hypotheses, $H[m]$: $H_1, \ldots, H_m$, and their associated p-values, $p[m]$: $p_1, \ldots, p_m$. The hypotheses can be considered to be H_i: $\beta_i = 0$. Forward stepwise provides an ordering for testing $H[m]$. Since our goal is model selection, a feature is "included" or "added" to the model when the corresponding null hypothesis is rejected. Define the statistic $R_i = 1$ if H_i is rejected and the random variable $V_i^\beta = 1$ if this was a false rejection. The dependence of V_i^β on β indicates that this is an unknown quantity which depends on the parameter of interest. Define

$$R(m) = \sum_{i=1}^{m} R_i, \text{ and}$$
$$V(m) = \sum_{i=1}^{m} V_i^\beta.$$

RAI approximates forward stepwise by making approximately greedy choices of features. At each step, forward stepwise sorts the p-values of the m' remaining features, $p_{(1)} < \ldots < p_{(m')}$, and selects the feature with the minimum p-value $p_{(1)}$. Instead of performing a full sort, consider using increasing significance thresholds, where hypotheses are rejected when their p-value falls below a threshold. Chapter 2 used thresholds determined by the Holm step-down procedure. The resulting procedure, Revisiting Holm (RH), is well-motivated for type-I error control, but can fail to accurately approximate forward stepwise in some cases.

3.0.4. Outline

Section 3.1 improves the Holm thresholds by controlling the risk incurred from making false selections. From the perspective of alpha-investing (Section 2.2.3), this has many other benefits as well. Section 3.1.2 discusses the main performance result in which RAI is shown to produce a near-optimal approximation of the performance of the best subset of the data.

RAI, and alpha-investing procedures in general, have additional flexibility beyond merely performing forward stepwise. This flexibility is demonstrated in Section 3.2 by adaptively searching high-order interaction spaces using RAI.

3.1. Better Threshold Approximation

Threshold approximations to forward stepwise can select different features than stepwise, reducing performance guarantees. To motivate our new thresholds, consider a threshold approximation that is guaranteed to mimic forward stepwise exactly. Suppose the set of thresholds is $i\delta$, $\delta > 0$ and $i = 1, 2, \ldots$. On the first "pass" through the hypotheses or "round" of testing, all m p-values are compared to δ. If none fall below this threshold, a second pass or round is conducted, in which all m p-values are compared to 2δ. The threshold is increased by δ each round until a hypothesis is rejected and the corresponding feature is added to the model. The stepwise p-values are recomputed according to their marginal reduction in ESS given the new model, and the process continues.[1] For sufficiently small δ, this procedure selects variables in the same order as forward stepwise.

3.1.1. Revisiting Alpha Investing

Instead of focusing on identifying the identical features as forward stepwise, we develop a set of thresholds to control the *additional risk* incurred from selecting different features than forward stepwise. These thresholds must initially correspond to stringent tests so that only extremely significant features are selected. Our significance thresholds correspond to selecting features which reduce the ESS by $\text{ESS}(\bar{Y})/2^i$, where i indicates the round of testing. The resulting procedure is provided in Algorithm 1 and is called Revisiting Alpha-Investing (RAI) because it is an alpha-investing strategy that tests hypotheses multiple times. This results in both practical and theoretical performance improvements while maintaining type-I error guarantees. The details of the required calculations are given in the Appendix.

RAI is well defined in any model in which it is possible to test the addition of a single feature

[1] If $\mathbf{X}$ is orthogonal, p-values do not change between steps. If p-values do change, a conservative approach restarts testing with threshold δ.

Algorithm 1 Revisiting Alpha-Investing (RAI)

Input: data Y and $\mathbf{X}$. Without loss of generality, let Y and $\mathbf{X}$ be centered and $\|Y\|_2 = \|X_i\|_2 = 1$, $\forall i$.
Initialize: $W = .25$, $\omega = .05$, $s = 1$, $S = \emptyset$
Output: selected model with coefficients
Set: $\mathbf{r} = \mathbf{y}$
repeat
  tlvl $= \sqrt{n} * 2^{-s/2}$ // testing level per testing pass
  α_s = 2*Φ(-tlvl) // alpha spent per test
  for $j \notin S$ **do** // consider features not in the model
    Compute t-statistic for $\mathbf{X}_j$: $\hat{t}_j$
    $W = W - \alpha_s$ // wealth lost from test
    if $|\hat{t}_j|$ >tlvl and $W > \alpha_s$ **then**
      $S = S \cup \{j\}$; $W = W + \omega$
      $\mathbf{r} = (\mathbf{I} - \mathbf{H}\mathbf{x}_S)\mathbf{y}$ // make new residuals
    end if
    if $W < \alpha_s$ **then** // If run out of alpha-wealth, end
      Output S
    end if
  end for
  $s = s + 1$
until max CPU time

such as generalized linear models. The testing thresholds ensure that the algorithm closely mimics forward stepwise, which provides performance guarantees. A precise statement of this comparison is given in Section 3.1.2.

Approximating stepwise using these thresholds has many practical performance benefits. First, multiple passes can be made without rejections before the algorithm exhausts its alpha-wealth and terminates. The initial tests are extremely conservative but only spend tiny amounts of alpha-wealth. Tests rejected in these stages still earn the full return ω. This ensures that wealth is not wasted too quickly when testing true null hypotheses. Furthermore, false hypotheses are not rejected using significantly more wealth than is required. An alternative construction of alpha-investing makes this latter benefit explicit and is explained in Foster and Stine (2008). Taken together, this improves power in ways not addressed by the theorem in the next section. By earning more alpha-wealth, future tests can be conducted at higher power while maintaining the type-I error guarantee.

RAI performs a sequential search for sufficient model improvement as opposed to the global search for maximal improvement performed by forward stepwise. Most sequential, or online,

algorithms are online in the observations, whereas RAI is online in the *features*. This allows features to be generated dynamically and allows extremely large data sets to be loaded into RAM one feature at a time. As such, RAI is trivially parallelizable in the MapReduce setting, similar to (Kumar et al., 2013). For example, many processors can be used, each considering a disjoint set of features. Control need only be passed to the master node when a significant feature is identified or a testing pass is completed. Parallelizing RAI will be particularly effective in extremely sparse models, such as those considered in genome-wide association studies. Online feature generation is beneficial when features are costly to generate and can be used for directed exploration of complex spaces. This is particularly useful when querying data base or searching interaction spaces and is described in Section 3.2.

Using additional speed improvements provided by variance inflation factor regression (VIF) (Lin et al., 2011), RAI performs forward stepwise and model selection in $O(np\log(n))$ time as opposed to the $O(np^2q^2)$ required for traditional forward stepwise, where q is the size of the selected model. The log term is an upper bound on the number of passes through the hypotheses performed by RAI. This is significantly reduced for large n by recognizing when passes may be skipped. This is possible whenever a full pass is made without any rejections, as all of the sequential p-values are known. The control provided by alpha-investing is maintained, because RAI must pay for all of the skipped tests. Using this computational shortcut, only 5-7 passes are required to select a model using RAI.

3.1.2. Performance Guarantee

This subsection bounds the performance of RAI and requires additional notation. Let $[m] = \{1, \ldots, m\}$. For a subset of indices $S \subset [m]$, we denote the corresponding columns of our data matrix as $\mathbf{X}_S$, or merely S when the overloaded notation will not cause confusion. Most of our discussion concerns maximizing the model fit as opposed to minimizing loss. Our measure of model fit for a set of features $\mathbf{X}_S$ is the coefficient of determination, R^2 ,

defined as

$$R^2\ (S) = 1 - \frac{ESS(\mathbf{X}_S\hat{\beta}_S)}{ESS(\bar{Y})}$$

where $\bar{Y}$ is the constant vector of the mean response and $\hat{\beta}_S$ is the least squares estimate of β_S. Maximizing an in-sample criterion such as R^2 is known to over-fit the data, worsening out-of-sample performance. To prevent over-fitting, a practical implementation of forward stepwise requires selecting a model size via cross-validation or criteria such as AIC. RAI bypasses this problem by controlling mFDR. Without loss of generality, we assume that our data is centered and normalized such that $\|Y\|_2 = \|X_i\|_2 = 1,\ \forall i$.

We will often need to consider a feature $\mathbf{X}_i$ orthogonal to those currently in the model, $\mathbf{X}_S$. This will be referred to as adjusting $\mathbf{X}_i$ for $\mathbf{X}_S$. The projection operator (hat matrix), $\mathbf{H}_{\mathbf{X}_S} = \mathbf{H}_S = \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T$, computes the orthogonal projection of a vector onto the span of the columns of $\mathbf{X}_S$. Therefore, $\mathbf{X}_i$ adjusted for $\mathbf{X}_S$ is denoted $\mathbf{X}_{i.S^\perp} = (\mathbf{I} - \mathbf{H}_{\mathbf{X}_S})\mathbf{X}_i$. This same notation holds for sets of variables: $\mathbf{X}_A$ adjusted for $\mathbf{X}_S$ is $\mathbf{X}_{A.S^\perp} = (\mathbf{I} - \mathbf{H}_{\mathbf{X}_S})\mathbf{X}_A$.

RAI is proven to perform well if the improvement in fit obtained by adding a set of features to a model is upper bounded by the sum of the improvements of adding the features individually. If a large set of features improves the model fit when considered together, this constraint requires some subsets of those features to improve the fit well. Consider the improvement in model fit by adding $\mathbf{X}_S$ to the model $\mathbf{X}_T$:

$$\Delta_T(S) := R^2\ (S \cup T) - R^2\ (T).$$

Letting $S = A \cup B$, we bound $\Delta_T(S)$ as

$$\Delta_T(A) + \Delta_T(B) \geq \Delta_T(S). \tag{3.1}$$

If $A \cup B$ improves the model fit, equation (3.1) requires that either A or B improve the fit. Therefore, signal that is present due to complex relationships among features cannot

be completely hidden when considering subsets of these features. Equation (3.1) defines a submodular function:

Definition 4 (Submodular Function)**.** *Let* $F : 2^{[m]} \to \mathbb{R}$ *be a set function defined on the the power set of* $[m]$*.* F *is submodular if* $\forall A, B \subset [m]$

$$F(A) + F(B) \geq F(A \cup B) + F(A \cap B) \tag{3.2}$$

This can be rewritten in the style of (3.1) as

$$\begin{aligned} F(A) - F(A \cap B) + F(B) - F(A \cap B) &\geq F(A \cup B) - F(A \cap B) \\ \Rightarrow \Delta_{A \cap B}(A) + \Delta_{A \cap B}(B) &\geq \Delta_{A \cap B}(A \cup B), \end{aligned}$$

which considers the impact of $A \backslash B$ and $B \backslash A$ given $A \cap B$. Given (3.1), it is natural to approximate the maximizer of a submodular function with a greedy algorithm. We provide a proof of the performance of RAI by assuming that R^2 is submodular or approximately so (made precise below).

In order for these results to hold even more generally, the definition of submodularity can be relaxed. To do so, iterate (3.1) until the left hand side is a function of the influences of individual features and only require the inequality to hold up to a multiplicative constant $\gamma \geq 0$. For additional simplicity, consider adding the set $A = \{a_i, \ldots, a_l\} \subset [m]$ to the model S. Hence $\Delta_S(a_i)$ is the marginal increase in R^2 by adding a_i to model S. When data is normalized, $\Delta_S(a_i)$ is the squared partial-correlation between the response Y and a_i given S: $\Delta_S(a_i) = \text{Cor}(Y, a_{i.S}^{\perp})^2$. Therefore, define the vector of partial correlations as $r_{Y,A.S^{\perp}} = \text{Cor}(Y, A.S^{\perp})$, then the sum of individual contributions to R^2 is $\|r_{Y,A.S^{\perp}}\|_2^2$. Similarly, if we define $C_{A.S^{\perp}}$ as the correlation matrix of $A.S^{\perp}$ then $\Delta_S(A) = r'_{Y,A.S^{\perp}} C_{A.S^{\perp}}^{-1} r_{Y,A.S^{\perp}}$.

Definition 5. *(Submodularity Ratio) The submodularity ratio,* γ_{sr}*, of* R^2 *with respect to a*

set S *and* $k \geq 1$ *is*

$$\gamma_{sr}(S,k) = \min_{(T:T\cap S=\emptyset, |T|\leq k)} \frac{r'_{Y,T.S^{\perp}} r_{Y,T.S^{\perp}}}{r'_{Y,T.S^{\perp}} C^{-1}_{T.S^{\perp}} r_{Y,T.S^{\perp}}}$$

The minimization identifies the worst case set T to add to the model S. It captures how much R^2 can increase by adding T to S (denominator) compared to the combined benefits of adding its elements to S individually (numerator). If S is the size-k set selected by forward stepwise, then R^2 is approximately submodular if $\gamma(S,k) > \gamma$, for some constant $\gamma > 0$. We will refer to data as being approximately submodular if R^2 is approximately submodular on the data. R^2 is submodular if $\gamma_(S,2) \geq 1$ for all $S \subset [m]$ (Johnson et al., 2015c). This definition is similar to that of Das and Kempe (2011).

Our main theoretical result provides a performance guarantee for RAI and is proven in the Appendix. The result is similar to the in-sample performance guarantees for forward stepwise provided by Das and Kempe (2011). Let s index the testing pass, with s_f denoting the first pass in which a hypothesis is rejected. The term $\gamma(S_l,k)$ is the submodularity ratio of the selected set of l features, denoted S_l, and S^*_k is the set of k features which minimizes equation (1.2).

Theorem 3. *Algorithm 1 (RAI) selects a set of features* S_l *of size* l *such that*

$$R^2(S_l) \geq \max\left\{c_1 R^2(S^*_k) - \sum_{j=1}^{l} e^{\frac{-(j-1)\gamma_{S_l,k}}{k}} 2^{j-(l+\xi_f)},\ c_2 R^2(S^*_k)\right\}$$

where $c_i = \left(1 - e^{\frac{-l\gamma_{S_l,k}}{ik}}\right)$.

The constant c_1 is the optimal constant for greedy approximation, yielding the standard (1-1/e) approximation for submodular function maximization. As RAI may deviate from true forward stepwise, the loss incurred can be summarized in two ways. The first term of the maximization incorporates a small, additive loss that is constructed by considering the additive cost of selecting a different variable than forward stepwise. It provides a better

bound when k is not small and $l \geq k$. The additive error is small, often less than .02. The cost of errors made during early testing passes are heavily discounted because the loss of the incorrect selections can be made-up in subsequent steps. The second term in the maximization incorporates the loss of selecting a different feature than forward stepwise as a multiplicative error and provides a better bound when l and k are small. Performance is often better than these bounds indicate, because performance is not a function of the order in which variables are added, but merely the set of variables in the final model. If RAI and forward stepwise select the same variables but in a different order, the bound is merely $c_1 R^2(S_k^*)$.

The bound from Theorem 3 does not require the linearity assumption of equation (1.1) to hold. It holds uniformly over the true functional relationship between Y and $\mathbf{X}$ because RAI is compared to the best linear approximation of Y given $\mathbf{X}$. The guarantee is also not a probabilistic statement. Therefore, the ability to compare a model of size l to a model of size k, where $l \neq k$, allows practitioners to trade computation time and model complexity for fit. For example, suppose a selection method selects k features with $R^2 = R^{2*}$. RAI can quickly identify a model that is guaranteed to achieve $.95R^{2*}$ by selecting $l = 3k$ features. RAI is designed to determine $l = k$ adaptively, however, which results in a stronger interpretation of the Theorem 3: RAI produces a near optimal approximation of the best size-k model, where k is chosen such that little signal remains in unselected features. The extent to which signal is hidden in the remaining features is a function of the approximate submodularity of the data (Johnson et al., 2015c).

3.2. Searching Interaction Spaces

As an application of RAI, we demonstrate a principled method to search interaction spaces while controlling type-I errors. In this case, submodularity is merely a formalism of the principle of marginality: if an interaction between two features is included in the multiple regression, the constituent features should be as well. This reflects a belief that an interaction is only informative if the marginal terms are as well. RAI can perform a greedy search

for main effects, while maintaining the flexibility to add polynomials to the model that were not in the original feature space. Therefore, we search interaction spaces in the following way: run RAI on the marginal data $\mathbf{X}$; for $i, j \in [m]$, if $\mathbf{X}_i$ and $\mathbf{X}_j$ are rejected, test their interaction by including it in the stepwise routine. This bypasses the need to explicitly enumerate the interaction space, which is computationally infeasible for large problems. Furthermore, as our data results indicate, it is highly beneficial to only consider relevant portions of interaction spaces, as the full space is too complex. This is addressed in detail below. To demonstrate the success of this routine we provide results on both simulated and real data.

3.2.1. Simulated Data

Simulated data is used to demonstrate the ability of RAI to identify polynomials in complex spaces. Our simulated explanatory variables have the following distribution:

$$\mathbf{X}_{i,j} \sim N(\tau_j, 1) \quad \text{where} \quad \tau_j \sim N(0, 4).$$

The true mean of Y, μ_Y, includes four terms which are polynomials in the first ten marginal variables:

$$\begin{aligned} Y &= \mu_Y + \epsilon \\ \mu_Y &= \beta_1 \mathbf{X}_1 \mathbf{X}_2 + \beta_2 \mathbf{X}_3 \mathbf{X}_4^2 + \beta_3 \mathbf{X}_5 \mathbf{X}_6^3 + \beta_4 \mathbf{X}_7 \mathbf{X}_8 \mathbf{X}_9 \mathbf{X}_{10} \\ \epsilon &\sim N(0, \mathbf{I}) \end{aligned}$$

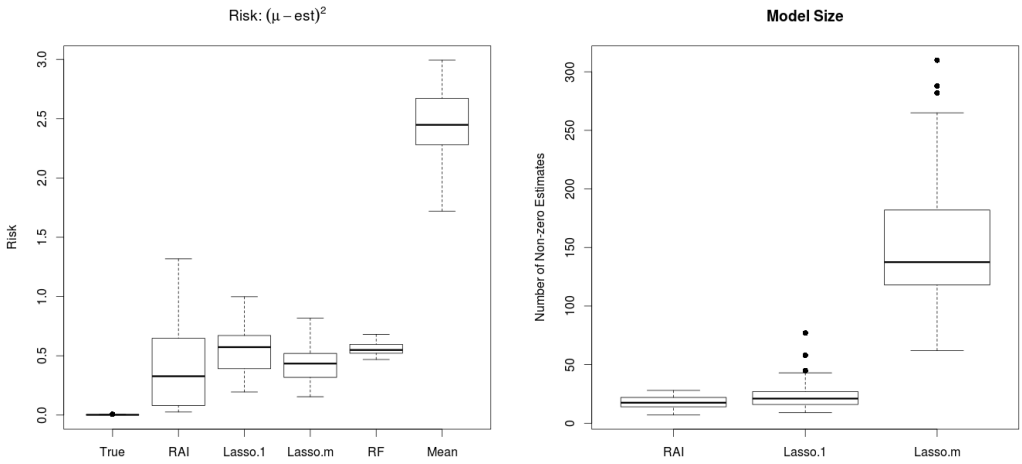
The coefficients $\beta_1, \ldots, \beta_4$, are equal given the norm of the interaction and are chosen to yield a true model R^2 of approximately .83. The t-statistics of features in the true model range between 25 and 40.

We first simulate a small-p environment: 2,000 observations with 350 explanatory features. While our features are simulated independently, the maximum observed correlation is ap-

proximately .14. While many competitor algorithms are compared on the real data, only two are presented here for simplicity. Our goal is to demonstrate the gains from searching complex spaces using feature selection algorithms. Five algorithms are compared: RAI searching the interaction space, the Lasso, random forests (Breiman, 2001), the true model, and the mean model. The mean model merely predicts $\bar{Y}$ in order to bound the range of reasonable performance between that of the true model and the mean model. Two Lasso models are compared: the one with minimum cross-validated error (Lasso.m) and the smallest model with cross-validated error within one standard deviation of the minimum (Lasso.1). Since the feature space is small, it is possible to compute the full interaction space of approximately 61,000 variables. Lasso is given this larger set, while RAI and random forests are only given the 350 marginal variables. Random forests is included such that comparison can be made to a high-performance, off-the-shelf procedure that also constructs its own feature space.

Figure 5 compares the risk of all procedures and the size of the model produced by the feature selection algorithms. The risk is computed using squared error loss from the true mean: $\|\mu_Y - \hat{Y}\|_2^2$. RAI often outperforms the competitors even though it is provided with far less information. The success of Lasso demonstrates the strength of correlation in this model. Even though Lasso can only accurately include the interaction $\mathbf{X}_1\mathbf{X}_2$, it is able to perform reasonably well in some cases. Figure 5 resamples the data 50 times, creating cases of varying difficulty. Often, difficult cases are challenging for all algorithms, such that the highest risk data set is the same for all procedures. RAI performs better than Lasso.m on the majority of cases and almost always outperforms Lasso.1. The overlapping box plots merely demonstrates the variability in the difficulty of data sets.

It is also worth comparing the size of the model selected by different procedures. The Lasso often selects a very large number of variables to account for its inability to incorporate the correct interactions. As we show more explicitly in the real data examples, this is a general problem even when the Lasso is provided the higher-order interactions. Using the model



(a) Risk

(b) Model Size

Figure 5: Small-p results.

identified by Lasso.1 dramatically reduces model size with a concomitant increase in loss. Contrast this with RAI, which selects a relatively small number of features even though its search space is conceptually infinite, as no bounds on complexity of interactions is imposed. Furthermore, RAI necessarily selects more than four features in order to identify the higher order terms. For example, in order to identify the term $\mathbf{X}_7\mathbf{X}_8\mathbf{X}_9\mathbf{X}_{10}$, all four marginal features need to be included as well.

While our results do not focus on speed, it is worth mentioning that RAI easily improves speed by a factor of 10-20 over the Lasso. This is notable since the Lasso is computed using glmnet (Hastie and Junyang, 2014), a highly optimized Fortran package, while RAI is coded in R and is geared toward conceptual clarity as opposed to speed. It also does not implement the improvement provided by VIF (Lin et al., 2011). Even outside of these considerations, RAI also does not have to compute the full interaction space.

Next, consider a comparatively large feature space case: 2,000 observations with 10,000 explanatory features. In this case, the maximum observed correlation between features is .177. Both RAI and the Lasso are only given the marginal variables because the full second-order interaction space has 50 million features. Traditional forward stepwise is also very

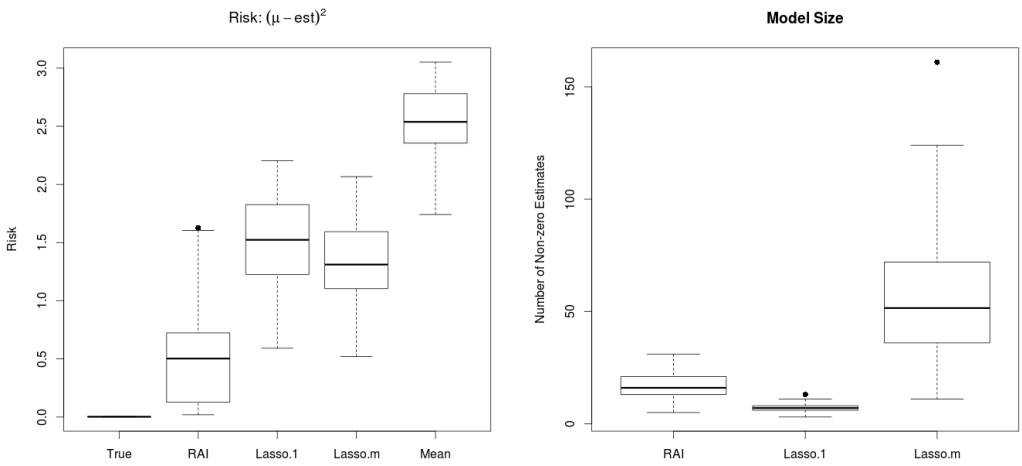


Figure 6: Large-p results.

time intensive to run on models of this size even without considering the interaction space. Therefore, an intelligent search procedure is required to identify signal. Random forests were excluded given the excessive time required to fit them off-the-shelf.

Figure 6 shows the risk and model size resulting from the algorithms fit to these data. When comparing the risk of the algorithms, RAI always outperforms both Lasso models, often by 55-90%. The overlapping region in the plot merely shows the variability in the difficulty of data cases.

RAI only identifies 1.6 true features on average, and rarely identifies all four. This is in large part due to the number of hypotheses that need to be rejected in order to identify an interaction such as $\mathbf{X}_7\mathbf{X}_8\mathbf{X}_9\mathbf{X}_{10}$. At least seven tests must be rejected, starting with the marginal terms, some second- and third-order interactions, and lastly the true feature. As the last test of $\mathbf{X}_7\mathbf{X}_8\mathbf{X}_9\mathbf{X}_{10}$ cannot even be conducted until all previous tests have been rejected, there is a high barrier to identifying such complex interactions. That being said, significant progress toward this true feature is made in all cases. For example, the model includes features such as $\mathbf{X}_7\mathbf{X}_8$ and $\mathbf{X}_8\mathbf{X}_9$ or $\mathbf{X}_7\mathbf{X}_9\mathbf{X}_{10}$. Therefore, in the small-m case, the true fitted space is not much larger than that considered by the Lasso. RAI performs

better in this case because it does not need to consider the full complexity of the 61,000 features in the interaction space.

3.2.2. Real Data

One may be concerned that the reason RAI outperforms the Lasso in the simulated scenarios is that RAI is able to search a more complex space. The simulated signal lies in higher-order polynomials of the features to which Lasso does not have access. While this itself is an important benefit of our method, we address this concern using a small, real data set. The results demonstrate that RAI is able to identify the appropriately complex interaction space. Searching unnecessarily complex spaces worsens performance of the competitor algorithms.

We use the concrete compressive strength data from the UCI machine learning repository (Yeh, 1998). This data set was chosen because the response, compressive strength, is described as a "highly nonlinear function of age and ingredients" such as cement, fly ash, water, superplasticizer, etc. It is also useful since it has approximately 1000 observations and only 8 features. A small number of features is needed so that a very large, higher-order interaction space can be generated. All interactions up to fifth order are provided to competitor algorithms, in which case there are 1,200 features.

We compare RAI to forward stepwise, Lasso, SCAD (Fan and Li, 2001), and the Dantzig selector (Candes and Tao, 2007). These are computed in R with the packages leaps, glmnet, ncvreg, and flare, respectively. Leaps and glmnet are both written in Fortran with wrappers for R implementation. SCAD uses a non-convex regularizer that attempts to blend the benefits of non-convex and convex regularizers. The stepwise model is chosen by minimizing AIC as this asymptotically selects a model that performs best among candidate models (Shao, 1997). The leaps package does not actually fit each model, so if selection with cross-validation is desired, computation time will increase concordantly. Lasso and SCAD use 10-fold cross-validation to determine the regularization parameter, since this is the default for their estimation functions. As before, both Lasso.m and Lasso.1 are considered. The

regularization parameter for the Dantzig selector is chosen via 5-fold cross-validation due to its slow run time.

To honestly estimate out-of-sample performance, we create 20 independent splits of the data into training and test sets. The training data is 5/6 of the full data, and the test set is the remainder. Each algorithm is fit using the training data; hence, cross-validation is conducted by splitting the training data again. The test set was only considered after the model was specified. We compare models using the predictive mean-squared error (PMSE) on the test set and average model size. Each row in Table 8 indicates the explanatory variables that the algorithms are provided. For example, the first row shows the performance results when all algorithms are only given marginal features, while in subsequent rows the competitor algorithms are given all second order interactions etc. Both the Stepwise and Dantzig models were excluded for the data set of fifth-order interactions. The Leaps package was unable to manage the complexity of the space and other implementations of stepwise proceed far too slowly to even consider being used on these data. Similarly, the Dantzig selector was too slow and performed too poorly on smaller feature spaces to warrant its inclusion.

Table 8: Concrete Compression Strength Results.

Set	Statistic	RAI	RAI.L	Step	Lasso.m	Lasso.1	SCAD	Dantzig
$\mathbf{X}$	MSE	37.33	39.75	112.93	112.63	116.78	112.80	204.08
$p = 8$	Size	37.25	36.30	6.60	8.60	7.85	7.90	3.05
$\mathbf{X}^2$	MSE	-	-	59.03	60.70	64.65	60.59	171.37
$p = 44$	Size	-	-	41.45	39.05	24.60	34.70	5.40
$\mathbf{X}^3$	MSE	-	-	38.58	38.02	40.25	38.50	174.29
$p = 164$	Size	-	-	135.10	125.55	73.15	68.15	6.35
$\mathbf{X}^4$	MSE	-	-	287.86	33.02	34.83	39.45	173.77
$p = 494$	Size	-	-	279.65	235.10	145.60	102.40	9.70
$\mathbf{X}^5$	MSE	-	-	-	47.65	51.32	61.03	-
$p = 1286$	Size	-	-	-	73.10	59.80	17.35	-

There are several important points in Table 8. As expected, RAI is superior to other feature selection methods when only considering marginal features; however, we can adjust for the information differences by giving the competitor algorithms a richer feature space. Other

methods need to be given all fourth-order interactions before they are competitive with RAI. As further information is provided, however, the performance of the competitor algorithms worsens. This demonstrates that RAI adaptively determines the appropriate feature space.

The number of features chosen by the algorithms is wildly different. For example, in the case of fourth-order interactions where Lasso.1 and Lasso.m are competitive with RAI, they select approximately 3.5-6.5 times as many features. As seen in the simulated data example, this may partially be due to the necessity of accounting for missing higher-order terms. These terms cannot be provided directly, however, because larger spaces, such as the fifth-order interaction space, become too complex for them to be found. The procedure RAI.L is included to further test the hypothesis that Lasso is merely not provided with the correct feature space. RAI.L first fits RAI to select the feature space, then selects a submodel via the Lasso. This second step rarely removes variables and performs statistically significantly worse than merely using the ordinary least squared model identified by RAI.

RAI can intelligently search the interaction space to identify complex signal while still performing valid feature selection. The performance stems from its simultaneous type-I and type-II error control.

3.3. Appendix

Before proving our main result, we give a detailed derivation of the spending and selection rules used by RAI. First, we give the sequence of cutoff values that RAI uses to select features. Second, we bound the maximum error made by the algorithm on each accepted feature. There exist both additive and multiplicative bounds on this error, each yielding a separate proof for the performance bound. These proofs are first given under submodularity. Lastly, we extend the proofs to allow for approximate submodularity. Without loss of generality, let Y and $\mathbf{X}$ be centered and $\|Y\|_2 = \|X_i\|_2 = 1$, $\forall i$.

RAI passes over the features several times, testing them against t-statistic thresholds that decrease with each pass. Each pass through the full data is indexed by s. Our algorithm

searches for features that result in an increase of $(1/2)^s$ in R^2 for the current model. This increase is upper bounded by $(1/2)^{(s+i)}$ in terms of R^2 on the original scale, where i is the number of features in the current model. Therefore, rejecting multiple hypotheses at the same level can result in an exponential decrease in residual variation. First, we must convert this increase in terms of R^2 to critical values. The maximum t-statistic is $(n-1)^{1/2}$, because

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} = (n-1)^{1/2} Cor(\mathbf{X}_i, Y).$$

R^2 of the simple regression of Y on $\mathbf{X}_i$ is merely $Cor(\mathbf{X}_i, Y)^2$. Therefore the desired cutoff is

$$t_s = (n-1)^{1/2} 2^{-s/2}.$$

Since RAI tests all of the features every pass, the largest difference between the R^2 of RAI and forward stepwise is incurred by selecting a feature whose hypothesis test was barely rejected, while the hypothesis test of the best feature barely failed being rejected in the previous testing pass. These choices increase R^2 by at least $(1/2)^{(s+i)}$ and $(1/2)^{(s+i-1)}$, respectively. The difference decreases by a multiplicative factor of $1/2$ whenever a feature is added or a testing pass is completed. Hence, the additive loss in terms of R^2 incurred when selecting a feature is bounded by

$$R^2_{opt} - R^2_{chosen} \leq 2^{-(s+i)}$$

We can plug this additive bound into the standard greedy proof of (Nemhauser et al., 1978), most of which stays the same. The proof is valid for any appropriately bounded submodular function. Therefore, we use a general f instead of R^2 . We denote the discrete derivative of f at S with respect to v as $\Delta(v|S) := f(S \cup \{v\}) - f(S)$. Let a_i be the feature chosen by RAI

at time i, s_f the testing pass in which the first feature is chosen, and $\delta_i = f(S_k^*) - f(S_i)$.

$$
\begin{aligned}
f(S_k^*) &\leq f(S_i \cup S_k^*) \\
&= f(S_i) + \sum_{j=1}^{k} \Delta(v_j^* | S_i \cup v_1, ...v_{j-1}) \\
&\leq f(S_i) + \sum_{v^* \in S_k^*} (f(S_i \cup v^*) - f(S_i)) \\
&\leq f(S_i) \\
&\quad + \sum_{v^* \in S_k^*} (2^{-(s+i)} + f(S_i \cup a_i) - f(S_i)) \\
&\leq f(S_i) + k(2^{-(s+i)} + f(S_{i+1}) - f(S_i)) \\
\Rightarrow \delta_{i+1} &\leq (1 - 1/k)\delta_i + 2^{-(s+i)} \\
&\leq e^{\frac{-(i+1)}{k}} f(S_k^*) + \sum_{j=0}^{i} (1 - 1/k)^j 2^{j-(s+i)} \\
\Rightarrow f(S_{i+1}) &\geq \left(1 - e^{\frac{-l}{k}}\right) f(S_k^*) - \sum_{j=1}^{l} e^{\frac{-(j-1)}{k}} 2^{j-(l+s)}
\end{aligned}
$$

Instead of considering the error bounded in additive terms, consider the multiplicative bound. Since the error is being cut in half, the worst error incurred is the remaining half. This is upper bounded by the observed reduction from adding the next feature. Changing

the proof accordingly yields,

$$
\begin{aligned}
f(S_k^*) &\leq f(S_i \cup S_k^*) && \text{by monotonicity} \\
&= f(S_i) + \sum_{j=1}^{k} \Delta(v_j^* | S_i \cup v_1, ...v_{j-1}) && \text{expanding sum} \\
&\leq f(S_i) + \sum_{v^* \in S_k^*} \Delta(v^* | S_i) && \text{by submodularity} \\
&= f(S_i) + \sum_{v^* \in S_k^*} (f(S_i \cup v^*) - f(S_i)) \\
&= f(S_i) + \sum_{v^* \in S_k^*} (f(S_i \cup v^*) - f(S_i \cup a_i) + f(S_i \cup a_i) - f(S_i)) \\
&\leq f(S_i) + \sum_{v^* \in S_k^*} 2(f(S_i \cup a_i) - f(S_i)) && \text{by the above discussion} \\
&= f(S_i) + \sum_{v^* \in S_k^*} 2(f(S_{i+1}) - f(S_i)) \\
&\leq f(S_i) + 2k(f(S_{i+1}) - f(S_i))
\end{aligned}
$$

Replacing k by $2k$ in the remainder of the proof yields

$$f(S_l) \geq \left(1 - e^{\frac{-l}{2k}}\right) f(S_k^*)$$

The above proofs do not leverage particular characteristics of the R^2 objective. The proof below is similar to that of (Das and Kempe, 2011), though somewhat cleaner and allows for $l \geq k$.

First, we need to compute the difference between the R^2 of adding a set of features and the sum of the changes in R^2 by adding the features one at a time. In what follows, T/S is the elements in T projected off of the elements in S and $T \backslash S$ is set difference. We write $\mathrm{R}^2(T/S)$ to be the contribution to R^2 of the features in T/S. $\mathbf{b}_T^S$ and C_T^S are defined in the introduction.

Lemma 1. *Given subsets S and T,*

$$R^2\ (S \cup T) = R^2\ (S) + R^2\ (S/T)$$

Proof. Let $\mathbf{X}_{S,S/T} = [\mathbf{X}_S, \mathbf{X}_{S/T}]$

$$\begin{aligned} R^2\ (S \cup T) &= R^2\ (S \cup S/T) \\ &= Y'\mathbf{X}_{S,S/T}(\mathbf{X}_{S,S/T}^T\mathbf{X}_{S,S/T})^{-1}\mathbf{X}_{S,S/T}^T Y \\ &= Y'\mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T Y + Y'\mathbf{X}_{S/T}(\mathbf{X}_{S/T}^T\mathbf{X}_{S/T})^{-1}\mathbf{X}_{S/T}^T Y \\ &= R^2\ (S) + R^2\ (S/T) \end{aligned}$$

The first line follows because the prediction space did not change and the second to last line follows because $(\mathbf{X}_{S,S/T}^T\mathbf{X}_{S,S/T})^{-1}$ is block diagonal. □

Lemma 2. *For simplicity let $T \cap S = \emptyset$, or define $\tilde{T} = T \backslash S$. Then,*

$$R^2\ (T/S) \leq \frac{\sum_{x \in T \backslash S} R^2\ (S \cup \{x\}) - R^2\ (S)}{\gamma(S, |T|)}$$

Proof.

$$\begin{aligned} R^2\ (T/S) &= (\mathbf{b}_T^S)'(C_T^S)^{-1}(\mathbf{b}_T^S) \\ &\leq \frac{(\mathbf{b}_T^S)'(\mathbf{b}_T^S)}{\gamma(S, |T|)} \\ &= \frac{\sum_{x \in T/S} R^2\ (\{x\})}{\gamma(S, |T|)}, \end{aligned}$$

where the inequality follows by the definition of $\gamma(S, |T|)$. Since each element in $\mathbf{b}_T^S$ is a correlation, squaring this gives the R^2 from the simple regression of Y on $\mathbf{x}/S$, giving the final equality. The lemma just rewrites the result of the projection off of S as a difference in observed R^2 . □

Proof of Theorems 1 & 2.

$$
\begin{aligned}
R^2\ (S_k^*) &\leq R^2\ (S_i \cup S_k^*) && \text{by monotonicity} \\
&= R^2\ (S_i) + R^2\ (S_k^*/S_i) && \text{Lemma 1} \\
&\leq R^2\ (S_i) + \frac{\sum_{x \in S_k^* \setminus S_i} R^2\ (\{x\})}{\gamma_{S_i, |S_k^* \setminus S_i|}} && \text{Lemma 2} \\
&\leq R^2\ (S_i) + \frac{k}{\gamma_{S_i, |S_k^* \setminus S_i|}} \max_{x \in S_k^* \setminus S_i} R^2\ (\{x\}) && \text{sum less than k*max} \\
&\leq R^2\ (S_i) + \frac{k}{\gamma_{S_i, |S_k^* \setminus S_i|}} (R^2\ (S_{i+1}) - R^2\ (S_i)) && \text{by greedy algorithm}
\end{aligned}
$$

Increasing the size of the set T by inclusion and increasing k can only decrease $\gamma(T, k)$. Therefore, $\gamma(S_i, |S_k^* \setminus S_i|) \geq \gamma(S_l, |S_k^*|)$. Making this replacement, this is the original proof with k replaced by $k/\gamma(S_l, k)$. Therefore,

$$R^2\ (S_l) \geq \left(1 - e^{\frac{-l\gamma(S_l,k)}{k}}\right) R^2\ (S_k^*)$$

The proofs for RAI using approximate submodularity can be added exactly as they were presented before. □

CHAPTER 4 : SUBMODULARITY IN STATISTICS

Subset selection problems are difficult because features can interact in unexpected ways. Here, "unexpected" means that the change in model fit when adding a feature can be completely different depending on the other features in the model. This paper characterizes the cases in which features produce such unexpected results.

A simple example from Miller (2002) clarifies this point. Suppose forward stepwise is run on the data in Table 9. The first step selects the feature that is maximally correlated with Y. For features $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, these are $r_{Y1} = .0$, $r_{Y2} = -.0016$, and $r_{Y3} = .4472$, respectively. Therefore, forward stepwise selects $\mathbf{X}_3$ on the first step. The second step chooses the feature with the maximum partial correlation. That is, the maximum correlation when features are considered orthogonally to $\mathbf{X}_3$. For $\mathbf{X}_1$ and $\mathbf{X}_2$ these are $r_{Y1.3^\perp} = .0$ and $r_{Y2.3^\perp} = -.0014$, respectively. Forward stepwise appears to find a significant features on the first step, but then no other features seem important. The true equation for the response, however, is $Y = \mathbf{X}_1 - \mathbf{X}_2$. This cannot be identified by forward stepwise because of the "incorrect" first step which includes $\mathbf{X}_3$. Furthermore, unless forward stepwise continues to select features even when they appear uninformative, the optimal set can never be found. Intuitively, the difficulty arises because $\mathbf{X}_1$ and $\mathbf{X}_2$ have large errors which cancel out.

Table 9: Simple data in which forward stepwise fails to identify the correct model.

Y	$\mathbf{X}_1$	$\mathbf{X}_2$	X_3
-2	1000	1002	0
-1	-1000	-999	-1
1	-1000	-1001	1
2	1000	998	0

Most of our discussion concerns maximizing the model fit as opposed to minimizing loss. Let $[m] = \{1, \ldots, m\}$. For a subset of indices $S \subset [m]$, we denote the corresponding columns of our data matrix as $\mathbf{X}_S$, or merely S when the overloaded notation will not cause confusion. Our measure of model fit for a set of features $\mathbf{X}_S$ is the coefficient of determination, R^2 ,

defined as

$$R^2\ (S) = 1 - \frac{\text{ESS}(\mathbf{X}_S\hat{\beta}_S)}{\text{ESS}(\bar{Y})}$$

where $\bar{Y}$ is the constant vector of the mean response and $\hat{\beta}_S$ is the least squares estimate of β_S.

Forward stepwise performs well when the improvement in fit obtained by adding a set of features to a model is upper bounded by the sum of the improvements of adding the features individually. If a set of features improves the model fit when considered together, a subset of those features must improve the fit as well. Consider the improvement in fit by adding $\mathbf{X}_S$ to the model $\mathbf{X}_T$:

$$\Delta_T(S) := R^2\ (S \cup T) - R^2\ (T).$$

Letting $S = A \cup B$, bound $\Delta_T(S)$ as

$$\Delta_T(A) + \Delta_T(B) \geq \Delta_T(S). \tag{4.1}$$

If $A \cup B$ improves the model fit, equation (4.1) requires that either A or B improve the fit when considered in isolation. Therefore, signal that is present due to complex relationships among features cannot be completely hidden when considering subsets of these features. Equation (4.1) defines a submodular function:

Definition 6 (Submodular Function)**.** *Let $F : 2^{[m]} \to \mathbb{R}$ be a set function defined on the the power set of $[m]$. F is submodular if $\forall A, B \subset [m]$*

$$F(A) + F(B) \geq F(A \cup B) + F(A \cap B) \tag{4.2}$$

The intuition in equation (4.1) is recovered by considering $A \cap B = \emptyset$. Alternatively,

Definition 6 can be rewritten as

$$F(A) - F(A \cap B) + F(B) - F(A \cap B) \geq F(A \cup B) - F(A \cap B)$$
$$\Rightarrow \Delta_{A \cap B}(A) + \Delta_{A \cap B}(B) \geq \Delta_{A \cap B}(A \cup B),$$

which considers the impact of $A \backslash B$ and $B \backslash A$ given $A \cap B$. The influence of the union of set differences is less than the impact of the sum of their marginal influences. We will refer to data as being submodular if R^2 is a submodular function on the data.

Forward selection is a natural algorithm under submodularity as it adds the feature to the model that yields the maximum marginal increase in fit. To fix notation, if S_i is the model at step i, feature X_j is added if

$$j = \underset{l \notin S_i}{\operatorname{argmax}} \Delta_{S_i}(X_l)$$

and $S_{i+1} = \{S_i \cup j\}$. Other greedy procedures have been proposed that change the criteria being maximized at each step. For different criteria, this yields orthogonal matching or orthogonal projection pursuit (Barron et al., 2008; Miller, 2002).

For all such methods, let $\hat{Y}^{(k)} = \mathbf{X}_{S_k} \hat{\beta}_{S_k}$ be the estimated response after k steps of the algorithm. Previous analyses determined the rate at which $\text{ESS}(\hat{Y}^{(k)})$ decreases as a function of k (Barron et al., 2008; Jaggi, 2013). Instead, we focus on identifying the data conditions that guarantee that $\text{ESS}(\hat{Y}^{(k)})$ is close to that of the optimal size k subset. If forward stepwise is used and R^2 is submodular, the classic result of Nemhauser et al. (1978) shows that $R^2\ (S_k) \geq (1 - 1/e) R^2\ (S_k^*)$, where S_k^* is the subset of features which solves the sparse regression problem in equation (1.2).

Instead of asking for an approximate solution to (1.3), one can relax the problem formulation. For example, the l_0 penalty can be relaxed to an l_1 penalty, yielding the Lasso (Tibshirani, 1996). Additionally, loss be measuring with the l_∞ norm, which yields the

Dantzig selector (Candes and Tao, 2007). While subset selection and greedy methods like forward stepwise are classically studied, these relaxations have been the primary focus of research in recent years. For cases where $\log(p) = O(n^c)$ for $c > 0$, the computational improvements from relaxing the constraint in equation (1.2) do not produce efficient algorithms. In these cases, a feature screening method can be used to reduce the dimensionality p to feasible ranges before performing model selection (Fan and Lv, 2008).

These define two classes of algorithms: the first maintains the problem formulation in (1.2) and provides approximate solutions, while the provides exact solutions to relaxed problem formulations. Given that both classes of algorithms can be used to answer the same question, it is natural to ask which style of approximation is preferred. A general framework comparing these as penalized regressions is given in Fan and Li (2001), and cases in which approximating (1.2) is preferable to solving the relaxations are discussed in Johnson et al. (2015a). We take a different approach and analyze the assumptions necessary to have performance guarantees for either class of methods.

Our main contribution is a characterization of the data situations which are difficult for feature selection algorithms. This characterization should provide statistical insight as well as a way to generalize insight gained from low-dimensional problems. Unfortunately these two are not accomplished in the same way, which necessitates providing multiple definitions of approximate submodularity.

Das and Kempe (2011) introduced a notion of approximate submodularity, measured by the submodularity ratio, which we will call "statistical submodularity" given its connection to performance guarantees of statistical algorithms. We provide a characterization of the data situations in which this criteria holds. While the submodularity ratio is statistically useful, it does not allow insight gained from low-dimensional problems to be generalized. We provide a stronger definition for approximate submodularity and show that it yields a lower bound on the submodularity ratio. In particular, we explain which data conditions yield approximate submodularity for all feasible two-dimensional regression problems. While this

is restrictive, it yields generalizable bounds and insights.

As submodularity is a function of model fit, it depends on the response Y. This allows for a broader understanding of problematic correlation structures and is highly relevant to many simulation settings. From this perspective, not all deviations from orthogonality are the same. Spectral measures of such deviations do not always account for this lack of symmetry. Often simulations are described by their signal to noise ratio without considering the relative difficulty of different functional forms of the response. Provide an honest measurement of the difficulty of simulated data case requires considering both the strength of the signal and the ease with which the signal can be found.

Lastly, we demonstrate that submodularity often appears in statistics literature, just not by that name. We discuss the restricted eigenvalue (Raskutti et al., 2010) and conditions for sure independent screening (SIS) (Fan and Lv, 2008). The discussion highlights the data situations in which the sparse regression problem (1.2) is solvable by either approximating the solution or relaxing the problem formulation. Essentially, achieving an approximate solution to the exact problem is successful is the same instances in which achieving an exact solution to the approximate problem is successful. Furthermore, counter-intuitive results from recent conditional testing literature on forward stepwise and Lasso (Taylor et al., 2014) are explained by deviations from submodularity.

Section 4.1 introduces submodularity and our definition of approximate submodularity. Section 4.2 provides a simple example with only two features to provide intuition about the constraint of approximate submodularity. Furthermore, it is shown how submodularity can influence the search path identified by a greedy procedure. We also demonstrate the effect of signal strength in conjunction with submodularity. If the signal is strong enough, deviations from submodularity are easier to tolerate because signal is harder to hide in complex relationships between features. Lastly, Section 4.3 discusses the connection between submodularity and more common assumptions in statistics.

4.1. Submodularity

Submodularity is a condition under which greedy algorithms perform well. In this section, submodularity is given a statistical interpretation which begins to reveal its relevance in statistics. We often need to consider a feature $\mathbf{X}_i$ orthogonal to those currently in the model, $\mathbf{X}_S$. This is referred to as adjusting $\mathbf{X}_i$ for $\mathbf{X}_S$. The projection operator (hat matrix), $\mathbf{H}_{\mathbf{X}_S} = \mathbf{H}_S = \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T$, projects a vector onto the span of the columns of $\mathbf{X}_S$. Therefore, $\mathbf{X}_i$ adjusted for $\mathbf{X}_S$ is denoted as residual $\mathbf{X}_{i.S^\perp} = (\mathbf{I} - \mathbf{H}_{\mathbf{X}_S})\mathbf{X}_i$. This same notation holds for sets of features: $\mathbf{X}_A$ adjusted for $\mathbf{X}_S$ is $\mathbf{X}_{A.S^\perp} = (\mathbf{I} - \mathbf{H}_{\mathbf{X}_S})\mathbf{X}_A$.

While assuming R^2 is submodular is uncommon in the statistics literature, an equivalent formulation has been discussed in the social science literature: the absence of conditional suppressor variables (Das and Kempe, 2008). It is often observed features that have positive marginal correlation with the response can have negative partial correlation in the presence of other features. Similarly, features can be more significant in the presence of others than they are in isolation. In these situations, "suppression" is said to have occurred. The words "suppression" and "suppressor variable" can be understood through the algebra of adjustment for multiple regression coefficients.

If $\mathbf{X}$ and Y are standardizes, the coefficient for a feature $\mathbf{X}_i$ in a simple regression is the correlation between $\mathbf{X}_i$ and Y: $r_{Y,i}$. Letting $C = S\backslash i$ be the other features in the model, the coefficient for $\mathbf{X}_i$ in a multiple regression is

$$\hat{\beta}_i = \frac{\langle Y, \mathbf{X}_{i,C^\perp}\rangle}{\langle \mathbf{X}_{i,C^\perp}, \mathbf{X}_{i,C^\perp}\rangle}.$$

Therefore suppression occurs when variability in the feature of interest that is unrelated to Y is *suppressed* by the other features in the model.

A suppressor variable is one which, once controlled for, *increases* the observed significance of another feature. The absence of a conditional suppressor implies that $\forall S \subset [m]$ and

$i, j \notin S$

$$|\text{Corr}(Y, \mathbf{X}_{i,(S\cup j)^\perp})| \leq |\text{Corr}(Y, \mathbf{X}_{i,(S)^\perp})|.$$

Suppression is fundamentally the same problem as Simpson's paradox and Lord's paradox. The distinction arises based on the type of features being considered. Given features $\mathbf{X}_1$ and $\mathbf{X}_2$, Simpson's paradox can occur when both features are categorical, Lord's paradox can occur when one is categorical and the other is numeric, and suppression can occur when both features are numeric. Any of these paradoxes create problems with interpreting the influence of features in a regression model.

If one is only interested in prediction, the interpretation of coefficients is often unimportant. The existence of a suppressor variable does not change the predictions from a model; however, suppression has significant consequences for the ability of an algorithm to identify an important feature. In extreme cases, important features can only be identified as such in the context of many other features. To extend the simple example given in the introduction, consider the following set of random variables:

$$\begin{aligned} \mathbf{Z} &= N_p(\mathbf{0}, \sigma_z \mathbf{I}_p) & \epsilon &= N_{p-1}(\mathbf{0}, \sigma_\epsilon \mathbf{I}_{p-1}) \\ X_{1:(p-1)} &= \mathbf{Z}_{1:(p-1)} + \epsilon & X_p &= \mathbf{Z}_p - \sum_i^{p-1} \epsilon_i \\ Y &= \sum_{i=1}^{p} X_i = \sum_i^p \mathbf{Z}_i \end{aligned}$$

Suppose that σ_ϵ/σ_z is large enough that the variability in ϵ hides any signal that is in $\mathbf{Z}_i$. In this example, any model with fewer than p features has an R^2 near 0, while using all p features yields an R^2 of 1. The improvement in fit by adding any single variable is approximately 0 or 1, depending on which other variables are in the model. This clearly harms any procedure that solves isolated subproblems. Given the equivalence between lack of suppression and submodularity, we will use these term interchangeably. Similarly,

subsets of features which violate Definition 6 are instances of supermodularity. Therefore suppression situations are also supermodular.[1] Further implications of the submodularity of R^2 are understood by considering equivalent definitions of submodular functions. Definition 6 provides the classical definition of submodularity, and two refinements can be made that merely specify the sets under consideration in increasing detail. For completeness, all three formulations are provided in Definition 7 and are ordered in terms of specificity.

Definition 7 (Submodularity)**.** *Let $F : 2^{[m]} \to \mathbb{R}$ be a set function defined on the the power set of $[m]$. F is submodular iff*

1. *(Definition)* $\forall A, B \subset [m]$

$$\begin{aligned} F(A) + F(B) &\geq F(A \cup B) + F(A \cap B) \\ \Rightarrow F(A) - F(A \cap B) &\geq F(A \cup B) - F(B) \\ \Rightarrow \Delta_{A \cap B}(A) &\geq \Delta_B(A) \end{aligned}$$

2. *(First-order difference)* $\forall A, B$ *such that* $A \subset B \subset [m]$ *and* $i \in [m] \backslash B$

$$\begin{aligned} F(A \cup \{i\}) - F(A) &\geq F(B \cup \{i\}) - F(B) \\ \Rightarrow \Delta_A(i) &\geq \Delta_B(i) \end{aligned}$$

3. *(Second-order difference)* $\forall A \subset [m]$ *and* $i, j \in [m] \backslash A$

$$\begin{aligned} F(A \cup \{i\}) - F(A) &\geq F(A \cup \{i, j\}) - F(A \cup \{j\}) \\ \Rightarrow \Delta_A(i) &\geq \Delta_{A \cup j}(i) \end{aligned}$$

The definition in terms of first-order differences shows that submodular functions are similar to concave functions in that they exhibit diminishing marginal returns. The marginal impact or discrete derivative of adding a feature to A is larger than that of adding it to B since

[1] Given that $R^2 \geq 0$, submodular function are also subadditive. Similarly, supermodular ones are superadditive. While we do not use this terminology, it may be encountered elsewhere.

$A \subset B$. In terms of optimization, however, they behave like convex functions and can be efficiently minimized. See Bach (2011) for a survey of this viewpoint. One further simplification is possible by specifying $B = A \cup \{j\}$, which yields the definition in terms of second-order differences. This provides the most granular, well-specified definition of submodularity, and it is the easiest to verify in practice. The proofs of the equivalence of these definitions are standard and can be found in many places, for example Bach (2011). Furthermore, when showing the equivalence of definitions for approximate submodularity, we will be using proofs of essentially the same form.

In statistical terms, the first- and second-order difference definitions capture the intuitive notion that correlated features *share* information. Suppose $\mathbf{X}_S$ is a highly positively correlated set of features where $\beta_i \geq 0$, $\forall i \in S$. If only $\mathbf{X}_j$, $j \in S$, is added to the model, it produces a larger marginal improvement in fit than if the entire set $\mathbf{X}_S$ is included: $\Delta_\emptyset(\mathbf{X}_j) \geq \Delta_{S \setminus j}(\mathbf{X}_j)$. This claim does not hold in general, but does in this example because it is submodular. Correlation structures which violate this intuitive notion of shared information are described in Section 4.2.1.

The above discussion follows from elementary decompositions of simple and multiple regression coefficients. Let $S = \{i \cup j\}$ and consider the following models, where subscripts m and s index the model coefficients and error terms:

	Multiple Regression	Simple Regression
Model	$Y = \beta_{0,m} + \mathbf{X}_i\beta_{i,m} + \mathbf{X}_j\beta_{j,m} + \epsilon_s$	$Y = \beta_{0,s} + \mathbf{X}_i\beta_{i,s} + \epsilon_m$
Estimated Coefficients	$\hat{\beta}_{0,m}$, $\hat{\beta}_{i,m}$ and $\hat{\beta}_{j,m}$	$\hat{\beta}_{0,s}$ and $\hat{\beta}_{i,s}$

The simple regression coefficient can be decomposed into direct and indirect effects:

$$\hat{\beta}_{i,s} = \underbrace{\hat{\beta}_{i,m}}_{\text{direct}} + \underbrace{\hat{\alpha}_j\hat{\beta}_{j,m}}_{\text{indirect}}. \tag{4.3}$$

where $\hat{\alpha}_j$ is estimated from

$$\mathbf{X}_i = \alpha_0 + \mathbf{X}_j\alpha_j + \epsilon.$$

By construction, all terms are positive in equation (4.3) and the simple regression coefficient $\hat{\beta}_{i,s}$ is larger than $\hat{\beta}_{i,m}$. Therefore, the marginal impact of adding $\mathbf{X}_i$ is larger in isolation than in conjunction with $\mathbf{X}_j$. While this is a simplistic example, it introduces the general insight gained in later sections. In the simplest case, submodular data requires positively correlated features to have correlations with the response of the same sign. For example, both must be negative or positive. Similarly, if features are negatively correlated, their correlations with the response need to be of opposite sign.

The conditions provided in Definition 7 need to be relaxed in order to capture the continuum of possible scenarios. This will provide a measure of how signal can "hide" in sets of features while not being visible marginally. This measure is closely connected to assumptions more commonly discussed in statistics (see Section 4.3). There are two conflicting interests when providing an approximate definition of submodularity. First, it needs to be statistically meaningful. Such a definition should characterize a relevant statistical problem that needs to be addressed by many algorithms. Second, understanding submodularity in spaces with few features should provide generalizable insight into higher-dimensional problems. Unfortunately, both goals are not accomplished in the same way. Therefore, two notions of approximate submodularity are developed and their relationships are described.

Forward stepwise works better if the influence of a set S can be bounded by the sum of the margin influences of the elements in it. This can be achieved by applying Definition 6 multiple times to reduce the left hand side to a sum of individual elements. If $A = \{a_1, \ldots, a_l\} \subset [m]$ and $B = \{b_1, \ldots, b_m\} \subset [m]$, this yields

$$\sum_{i=1}^{l} \Delta_{A \cap B}(a_i) + \sum_{i=1}^{m} \Delta_{A \cap B}(b_i) \geq \Delta_{A \cap B}(A \cup B). \tag{4.4}$$

Note that for elements $a_i \in A \cap B$ or $b_i \in A \cap B$ that $\Delta_{A \cap B}(a_i) = \Delta_{A \cap B}(b_i) = 0$.

Das and Kempe (2011) propose a definition of approximate submodularity that requires equation (4.4) to hold approximately by including a constant $\gamma_{sr} > 0$ on the right hand

side. This is different than incorporating the same constant into Definition 6 as multiple applications of the definition are required to produce equation (4.4). For additional simplicity, consider adding the set $A = \{a_i, \ldots, a_l\} \subset [m]$ to the model S. Hence $\Delta_S(a_i)$ is the marginal increase in R^2 by adding a_i to model S. In this simple case, $\Delta_S(a_i)$ is the squared partial-correlation between the response Y and a_i given S: $\Delta_S(a_i) = \text{Cor}(Y, a_{i.S}^{\perp})^2$. Therefore, define the vector of partial correlations as $r_{Y,A.S^{\perp}} = \text{Cor}(Y, A.S^{\perp})$, then the left hand side of (4.4) is $\|r_{Y,A.S^{\perp}}\|_2^2$. Similarly, if we define $C_{A.S^{\perp}}$ as the correlation matrix of $A.S^{\perp}$ then $\Delta_S(A) = r'_{Y,A.S^{\perp}} C^{-1}_{A.S^{\perp}} r_{Y,A.S^{\perp}}$.

Definition 8. *(Das and Kempe, 2011) The submodularity ratio, γ_{sr}, of R^2 with respect to a set S and $k \geq 1$ is*

$$\gamma_{sr}(S,k) = \min_{(T:T\cap S=\emptyset, |T|\leq k)} \frac{r'_{Y,T.S^{\perp}} r_{Y,T.S^{\perp}}}{r'_{Y,T.S^{\perp}} C^{-1}_{T.S^{\perp}} r_{Y,T.S^{\perp}}}$$

The minimization identifies the worst case set T to add to the model S. It captures how much R^2 can increase by adding T to S (denominator) compared to the combined benefits of adding its elements to S individually (numerator). R^2 is submodular if $\gamma_{sr} \geq 1$ for all $S \subset [m]$ and $k = 2$. Only checking $k = 2$ is sufficient due to the second-order difference definition of submodularity and is clear from the proofs later in this section.

To not conflate the different notions of approximate submodularity introduced in this section, γ_{sr} will be referred to as the submodularity ratio or statistical submodularity. It can be used in proofs of the performance of greedy algorithms (Johnson et al., 2015b; Das and Kempe, 2011) and is lower bounded by a sparse eigenvalue (Das and Kempe, 2011). The connection to spectral quantities is obvious as γ_{sr} is an inverted Rayleigh quotient of the covariance matrix $C_{T.L^{\perp}}$. As $C_{T.L^{\perp}}$ is the Schur complement of $C_{T\cup L}$, Corollary 2.4 from Zhang (2006) proves that γ_{sr} is lower bounded by the minimum eigenvalue of $C_{T\cup L}$. The minimum sparse eigenvalue merely removes the dependence on the selected sets L and T. The connections to other algorithms that depend on spectral quantities are discussed in Section 4.3.

The submodularity ratio is not appealing from the perspective of submodularity. It is redefined for different cardinalities k and does not allow information gained for fixed k to percolate to larger k. We now provide a refined construction of approximate submodularity that produces generalizable insights. The definitions of approximate submodularity should mirror those of Definition 7, so that knowledge gained from restrictive, two-dimensional cases can generalize to higher-dimensional cases. These equivalent definitions, however, consider submodular functions in a slightly different context than the submodularity ratio γ_{sr}. The distinction is due to bounding the minimum of a set of differences versus the sum of a set of differences. Clearly bounding the minimum is stronger.

Approximate submodularity is constructed by starting with the second-order differences definition as it is the most granular and well-specified. Ideally, the sum of the marginal impact of features considered individually would be approximately greater than their impact considered jointly. Namely, for some constant $\gamma > 0$,

$$\Delta_A(i) + \Delta_A(j) \geq \gamma \Delta_A(i, j). \tag{4.5}$$

This is $\gamma_{A,2}$ after fixing the sets being minimized, but is unfortunately too weak to generalize to the larger sets considered in Definition 7. Instead, we must maintain the type of comparisons considered in the standard definitions.

Definition 9 (Approximate Submodularity)**.** *F is approximately submodular if there exists constants γ_s γ_{s2}, where $\gamma_{s2} \geq \gamma_s > 0$, such that any of the following hold*

1. *(Second order difference) $\forall A \subset [m]$ and $i, j \in [m] \backslash A$*

$$F(A \cup \{i\}) - F(A) \geq \gamma_{s2}(F(A \cup \{i, j\}) - F(A \cup \{j\}))$$
$$\Rightarrow \Delta_A(i) \geq \gamma_{s2} \Delta_{A \cup j}(i)$$

2. *(First order difference)* $\forall A, B$ *such that* $A \subset B \subset [m]$ *and* $i \in [m] \backslash B$

$$F(A \cup \{i\}) - F(A) \geq \gamma_s (F(B \cup \{i\}) - F(B))$$
$$\Rightarrow \Delta_A(i) \geq \gamma_s \Delta_B(i)$$

3. *(Definition)* $\forall A, B \subset [m]$

$$F(A) - F(A \cap B) \geq \gamma_s (F(A \cup B) - F(B))$$
$$\Delta_{A \cap B}(A) \geq \gamma_s \Delta_B(A)$$

One difference between the definitions for submodularity and approximate submodularity is that the constant will not be the same in all three cases, as indicated by our use of γ_s and γ_{s2}; however, if either is strictly greater than 0, then they both are. We are most interested in γ_s, which considers large sets, instead of γ_{s2}, which only holds for second order differences. We are able to provide a full account for γ_{s2} though, which yields a conservative lower bound on γ_s. Therefore, understanding approximate submodularity in two dimensions gives generalizable insights. The equivalence of these definitions is proved in the Appendix.

It is easy to see that $\gamma_{sr} \geq \gamma_{s2}$ in the relevant region in which $\gamma_{s2} \leq 1$. In this region, forward stepwise can perform poorly. The second-order characterization of γ_{sr} in equation (4.5) can be constructed using γ_{s2}.

$$\Delta_A(i) \geq \gamma_{s_2} \Delta_{A \cup j}(i)$$
$$F(A \cup i) - F(A) \geq \gamma_{s2}(F(A \cup \{i, j\}) - F(A \cup \{j\}))$$
$$F(A \cup i) - F(A) + \gamma_{s2}(F(A \cup \{j\}) - F(A)) \geq \gamma_{s2}(F(A \cup \{i, j\}) - F(A))$$
$$\Rightarrow \Delta_A(i) + \Delta_A(j) \geq \gamma_s \Delta_A(i, j).$$

Where the last line follows since $\gamma_{s2} \leq 1$. The submodularity ratio fixes the base set; hence the above rearranges the definition of γ_{s2} such that the marginal impact of all features is

relative to the same base set A. This yields a bound on the sum of marginal effects, whereas the γ_{s2} is a bound on the minimum of the marginal effects. As expected, the minimum can yield much worse bounds than the sum; however, as seen in Section (4.2.1), not all steps can be taken at this worst case bound.

4.2. Submodularity in 2 Dimensions

Attempting to classify types of suppression led Tzelgov and Henik (1991) to graph suppression situations that are possible with only two features. These graphs have unintuitive dimension, double-count data instances, and show impossible configurations. We analyze the same case, but provide graphs that fully characterize the set of possible regression problems. This clearly displays the regions in which γ_{s2} and γ_{sr} are bounded.

4.2.1. Graphing Approximate Submodularity

We parameterize possible regression problems using angles derived from projecting the response onto individual features. Our data consists of Y, $\mathbf{X}_1$, and $\mathbf{X}_2$ and $\hat{Y}_i$ as the response Y projected onto $\mathbf{X}_i$. See Figure 7 for an illustration. Since all features have been normalized, the distance from the origin to $\hat{Y}_i$ is the correlation between Y and $\mathbf{X}_i$, r_{Yi}. The correlation between explanatory features is parameterized as $\cos(\theta)$, where θ is the angle between $\mathbf{X}_1$ and $\mathbf{X}_2$. The relative predictive power of the features is measured by τ, the angle between $\hat{Y}_1$ and $\hat{Y}_2$. Lastly, the strength of the signal is a function of the length of b, the side between $\hat{Y}_1$ and $\hat{Y}_2$.

Figures 8, 9, and 10 only display $\theta \in [0, \pi]$ and $r_{Yi} \geq 0$ because of the symmetries in submodularity. $\theta > \pi$ is equivalent to $\theta' = (2 - \theta)\pi \in [0, \pi]$ and $r_{Yi} = -r_{Yi}$ for some i. The vertical axis has units $(\tau + \theta/2)\pi$ so that the contour plots are symmetric around $.5\pi$. Figure 7 is an isosceles triangle when $\tau + \theta/2 = .5\pi$, meaning that both features have the same marginal significance. Therefore, deviations correspond to one feature being marginally more significant than the other. Similarly, $\theta = .5\pi$ is the orthogonal case and represents one line of symmetry on the horizontal axis.

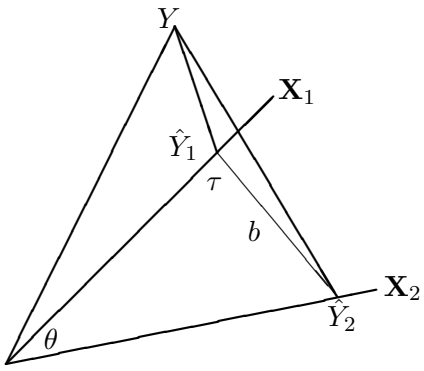


Figure 7: Characterization of possible two-dimensional regression problems: our data consists of Y, $\mathbf{X}_1$, and $\mathbf{X}_2$. $\hat{Y}_i$ is Y projected on $\mathbf{X}_i$. The side length from the origin to $\hat{Y}_i$ is r_{Yi}.

To completely specify the derived triangle in Figure 7, fix a measure of the signal to noise ratio as this does not represent a meaningful distinction between models for submodularity. Higher signal just means that the effects will be larger. This has the practical impact of being making it easier to identify a significant effect, but this discussion is delayed until Section 4.2.2. For convenience, we fix R^2 under the full model: $R^2_{\ full} = .5$. All figures are identical for any value of $R^2_{\ full} \in (0, 1]$. The length of b, the side between $\hat{Y}_1$ and $\hat{Y}_2$, is $\sqrt{(1 - r_{12}^2)R^2_{\ full}}$.

Figure 8 is a contour plot of γ_{s2} over the set of feasible regression problems. It demonstrates that submodularity ($\gamma_{s2} \geq 1$) is only possible when $\text{sign}(r_{12}r_{Y1}r_{Y2}) = 1$. This is the intuitive case introduced in Section 4.1: if features have opposing relationships with the response, we expect them to be negatively correlated. Since Figure 8 displays $r_{Y1} > 0$ and $r_{Y2} > 0$, submodularity only occurs when the features are positively correlated. Furthermore, for fixed r_{12} the maximum γ_{s2} occurs when both features have equal marginal effect. As this is a only a two-feature problem, the joint effects are also equal. Therefore, the common simulation setting that sets all non-zero coefficients to the same value maximizes the worse-case step, improving the performance of feature selection algorithms.

Figure 8 demonstrates that while submodularity holds in a large area, relaxing the definition does not increase the set of problems in a dramatic way; however, this is because γ_{s2} is the

single worst case step. Let γ_i be $\Delta(\mathbf{X}_i)/\Delta_{\mathbf{X}_j}(\mathbf{X}_i)$, $i \neq j$, $i, j \in \{1, 2\}$, then γ_{s2} is calculated by

$$\begin{aligned}
\gamma_1 &= \frac{r_{Y1}^2}{(r_{Y1}^2 - 2r_{Y1}r_{Y2}r_{Y2} + r_{Y2}^2 r_{12}^2)/(1 - r_{12}^2)} \\
\gamma_2 &= \frac{r_{Y2}^2}{(r_{Y2}^2 - 2r_{Y1}r_{Y2}r_{12} + r_{Y1}^2 r_{12}^2)/(1 - r_{12}^2)} \\
\gamma_{s2} &= \min(\gamma_1, \gamma_2).
\end{aligned}$$

γ_i is not symmetric in $\mathbf{X}_1$ and $\mathbf{X}_2$, though given our interest is in the true model containing both features, it is only important that one feature appears marginally significant. Importantly, both features cannot attain the minimum level γ_{s2} simultaneously.

To illustrate this, consider bounding the marginal impact of both $\mathbf{X}_1$ and $\mathbf{X}_2$ using γ_{s2}. Summing these two inequalities produces

$$\begin{aligned}
&\frac{\Delta_A(i) + \Delta_A(j)}{\Delta_{A\cup j}(i) + \Delta_{A\cup i}(j)} \geq \gamma_{s2} \\
\Rightarrow &\frac{\Delta_A(i) + \Delta_A(j)}{2\Delta_A(i,j) - \Delta_A(i) - \Delta_A(j)} \geq \gamma_{s2},
\end{aligned} \tag{4.6}$$

where the second line just rewrites the first such that the base set is constant. Figure 9 is a contour plot of the left hand side of equation (4.6). Clearly γ_{s2} is a poor bound on this function, demonstrating that if signal is contained in the joint distribution of the features, it cannot be hidden from both marginal distributions simultaneously. It demonstrates that useful properties of submodularity obtain in much larger region than indicated by γ_{s2} due to its conservativeness.

Lastly, Figure 10 is a contour plot of the submodularity ratio γ_{sr}. It behaves similarly to the bound on the sum in Figure 9, though more regularly. There are several interesting features that can be seen from this graph. First, γ_{sr} can be larger than 1. These are data situations in which forward stepwise achieves a better bound than the usual $(1 - 1/e)$ factor off of the optimal. This region corresponds to cases when the features are highly correlated

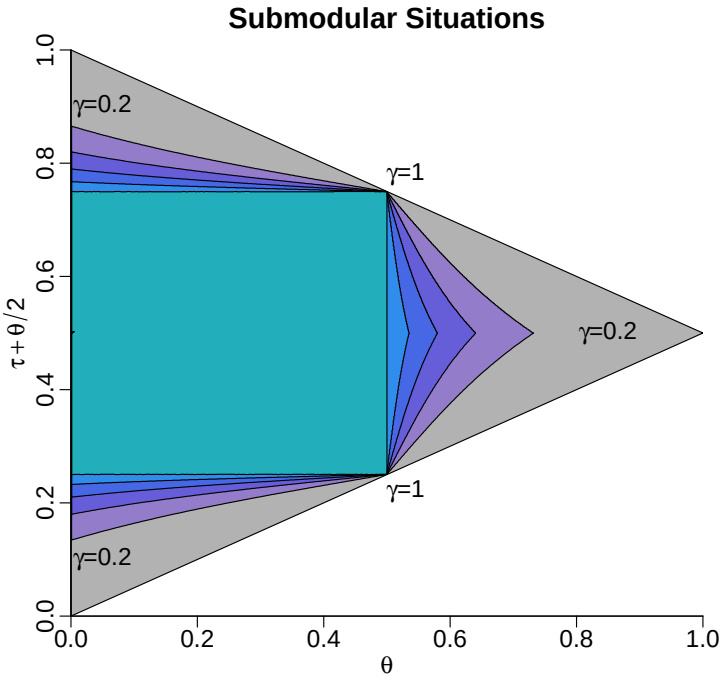


Figure 8: Contour plot of approximate submodularity using second order differences (γ_{s2}). Level sets are given for $\gamma_{s2} \in \{.2, .4, \ldots, 1\}$.

and have similar marginal relationships with Y. In this case, there is redundancy in our data and selecting appropriate features is less difficult.

Second, the dependence of γ_{sr} on Y is captured by the vertical axis via τ. Only orthogonal data, $\theta = .5\pi$, is submodular regardless of Y. In this case, the definition of submodularity, equation (4.2), holds with equality. This defines a modular function, and it is well known that the greedy algorithm produces the optimal answer when maximizing a modular function (Fujishige, 2005). Due to this dependence on Y, γ_{sr} is not symmetric around the orthogonal case. Obviously the feasible region is not symmetric, but we consider symmetry in terms of the contours of γ_{sr}. The minimum γ_{sr} along any vertical strip is achieved at the boundary of the feasible region. Along the boundaries, submodularity decays at the same rate when orthogonality is violated with by either positive or negative correlation. In this way, submodularity is symmetric around the orthogonal case. This demonstrates the result

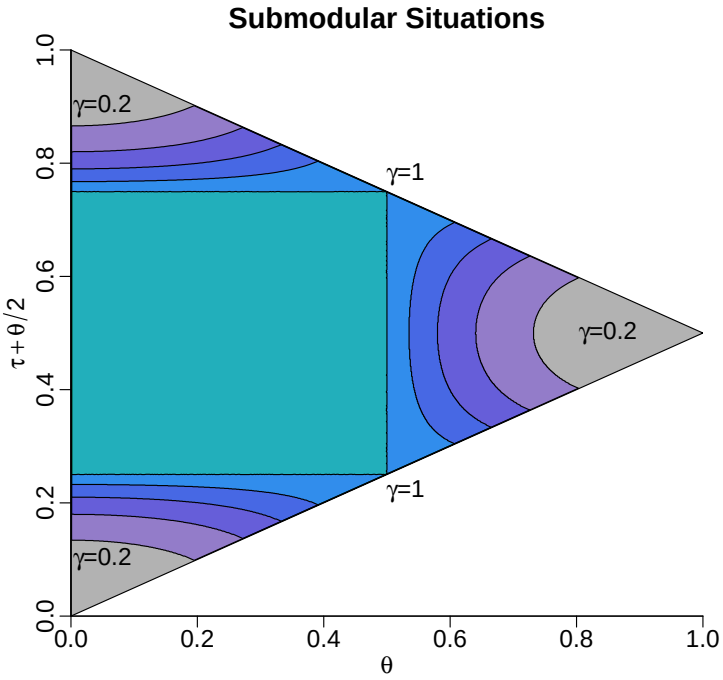


Figure 9: Contour plot of the left hand side of equation (9). The level sets are $\{.2, .4, \ldots, 1\}$.

of Das and Kempe (2011), that γ_{sr} is lower bounded by minimum eigenvalues, which occur on the boundary of the feasible region.

4.2.2. Graphing Change in t-Statistics

We now address the issue of *significant* suppression. As the deviation from submodularity grows, the greedy search path can deviate from the optimal path; however, slight suppression does not mean that the true model will not be found. For example, even suppressed features may still be marginally significant enough to be identified. In this case, the greedy search procedure has not been harmed.

To analyze these cases, the submodularity ratio can be related to differences in t-statistics. As in Figure 10, consider the contours of the percentage change in t-statistics caused by

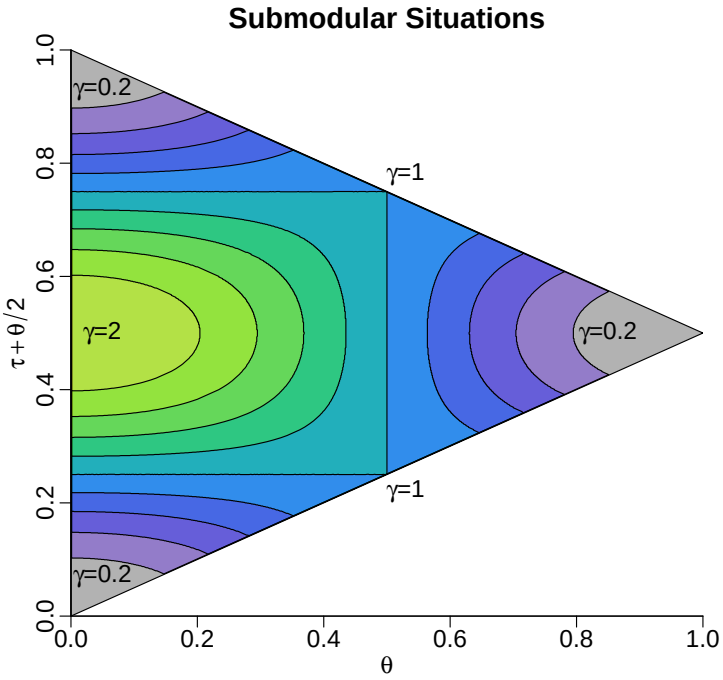


Figure 10: This is a contour plot of the submodularity ratio over the set of feasible regression problems. Level sets are given for $\gamma_{sr} \in \{.2, .4, \ldots, 2\}$.

different correlation structures. For clarity, consider the following statistics:

$$
\begin{aligned}
\beta_{1m} &= r_{y1} && \text{m for marginal} \\
\beta_{1j} &= \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2} && \text{j for joint} \\
t_{1m} &= \frac{r_{y1}}{\sigma_{im}} \\
\sigma_{1m}^2 &= \frac{1 - r_{y1}^2}{\sqrt{n-1}} \\
t_{1j} &= \frac{(r_{y1} - r_{y2}r_{12})}{(1 - r_{12}^2)^{1/2}\sigma_j} \\
\sigma_j^2 &= \frac{1}{\sqrt{n-1}} - \frac{r_{y1}^2 - 2r_{y1}r_{y2}r_{12} + r_{y2}^2}{\sqrt{n-1}(1 - r_{12}^2)}
\end{aligned}
$$

Submodularity requires $t_{1m}^2 \geq t_{1j}^2$. This is a conservative statement since $\frac{\sigma_{1m}^2}{\sigma_j^2} > 1$. If the features are jointly highly significant, this becomes very conservative because the ratio is

much larger than 1.

$$
\begin{aligned}
t_{m1}^2 = \frac{r_{y1}^2}{\sigma_m^2} &\geq \frac{r_{y1}^2 - 2r_{y1}r_{y2}r_{12} + (r_{y2}r_{12})^2}{(1-r_{12}^2)\sigma_j^2} = t_{j1}^2 \\
\Rightarrow r_{y1}^2 &\geq \frac{r_{y1}^2 - 2r_{y1}r_{y2}r_{12} + (r_{y2}r_{12})^2}{1-r_{12}^2} \\
\Rightarrow r_{y1}^2 + r_{y2}^2 &\geq \frac{r_{y1}^2 - 2r_{y1}r_{y2}r_{12} + r_{y2}^2}{1-r_{12}^2}
\end{aligned}
$$

Some algebra and incorporating γ_{sr} yields the following bound on the difference between the squared t-statistics:

$$
\Rightarrow t_{j1}^2 - t_{m1}^2 \leq \frac{(1-\gamma_{sr})(r_{y1}^2 - 2r_{12}r_{y1}r_{y2} + r_{y2}^2)}{1-r_{12}^2}
$$

The previous display ignores the symmetry of the problem: it is not of concern which of $\mathbf{X}_1$ or $\mathbf{X}_2$ is the suppressed feature, merely that there exists one. As such, add the corresponding equation for $\mathbf{X}_2$ and divide by the sum of the marginal t-statistics. This treats $\mathbf{X}_1$ and $\mathbf{X}_2$ symmetrically, and yields

$$
\begin{aligned}
\frac{t_{j1}^2 + t_{j2}^2}{t_{m1}^2 + t_{m2}^2} &\leq 1 + \frac{2(1-\gamma_{sr})(r_{y1}^2 - 2r_{12}r_{y1}r_{y2} + r_{y2}^2)}{(1-r_{12}^2)(r_{y1}^2 + r_{y2}^2)} \\
&= 2\gamma_{sr}^{-1} - 1. \qquad (4.7)
\end{aligned}
$$

Since equation (4.7) is conveniently written in terms of the γ_{sr}, we provide its contour plot in Figure 11. Equation (4.7) is always positive since $\gamma_{sr} \leq 2$.

The contours of Figure 11 are similar to those in Figure 10, but the contours change at different rates. If $\gamma_{sr} > .8$, then the ratio of squared t-statistics cannot be greater than 1.5. In this case, if a greedy procedure stops because all remaining features have a marginal t-statistic less than 2 in absolute value, neither feature can have a t-statistic larger than 3.46 when considered jointly. This upper bound is attained when one feature has a joint t-statistic of 0. If the joint information is split evenly between the two features, the maximum joint t-statistics are 2.44. Again, it is important to note that R^2 is not involved in this

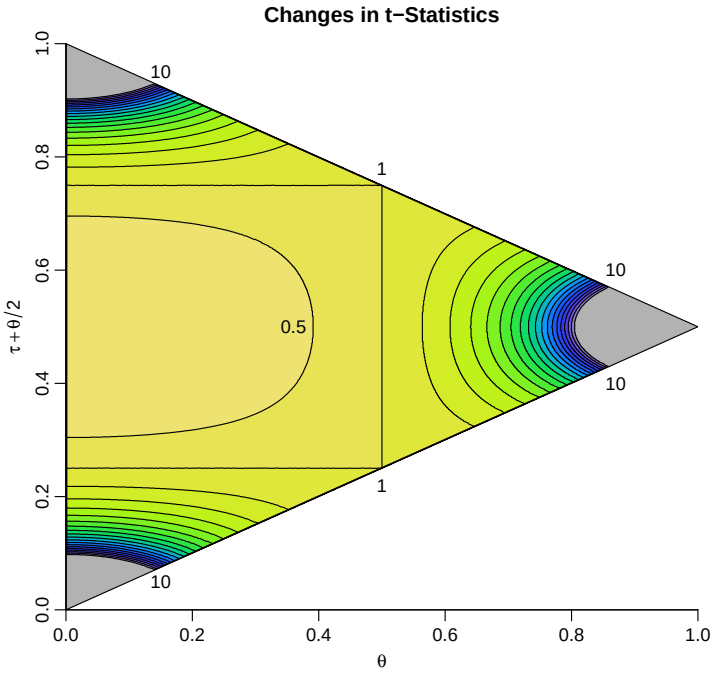


Figure 11: Contour plot of equation (4.7). The contours interpolate between .5 and 10 with a step-size of .5.

equation. Therefore submodularity is measuring a fundamentally different component than the signal-to-noise ratio.

4.3. Connection to Other Assumptions

Some algorithms that leverage assumptions similar to submodularity are the Lasso, Dantzig selector, and sure independent screening (SIS). It should not be surprising that the Lasso and forward stepwise are closely connected as the LARS procedure demonstrates the approximate greedy nature of the Lasso (Efron et al., 2004). This similarity extends to the Dantzig selector given that the same assumption guarantees success of the Lasso and Dantzig selector (Bickel et al., 2009). Lastly, SIS needs guarantees that information learned from marginal correlations is sufficient for model selection (Fan and Lv, 2008). This section describes these procedures, the assumptions used to demonstrate their success, and their close connection to submodularity.

4.3.1. Lasso and Dantzig

Relaxing the constraint from problem (1.3) from $\|\beta\|_{l_0}$ to $\|\beta\|_{l_1}$ yields the Lasso problem (Tibshirani, 1996).

$$\hat{\beta}_l = \text{argmin}_\beta \{\text{ESS}(\mathbf{X}\beta) + \lambda\|\beta\|_{l_1}\} \tag{4.8}$$

This is a convex program and can be efficiently solved using a variety of algorithms (Efron et al., 2004; Hastie and Junyang, 2014).

The Dantzig selector (Candes and Tao, 2007) optimizes the following linear program

$$\hat{\beta}_d = \text{argmin}_\beta \{\|Y - \mathbf{X}\beta\|_\infty + \lambda\|\beta\|_{l_1}\} \tag{4.9}$$

Our discussion of these procedures focuses on the assumptions required to provide bounds on their prediction loss. Many properties have been defined such as the restricted isometry

property (Candes and Tao, 2005), the restricted eigenvalue constant (Bickel et al., 2009; Raskutti et al., 2010), or the compatibility condition (van de Geer, 2007). For a review of these and related assumptions, see van de Geer and Bühlmann (2009). For our purposes, the most important of these is the restricted eigenvalue, which is defined over a restricted set of vectors that contain $\hat{\beta}_l$ and $\hat{\beta}_d$. Consider a subset $S \subset \{1, \ldots, p\}$ and constant $\alpha > 1$. Define the set

$$C(S; \alpha) := \{\beta \in \mathbb{R}^p | \|\beta_{S^c}\|_1 \leq \alpha \|\beta_S\|_1\}$$

The restricted eigenvalue of the $p \times p$ sample covariance matrix $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}/n$ is defined over S with parameter $\alpha \geq 1$.

$$\gamma_{re}^2(\alpha, S) := \min \left\{ \frac{\beta' \hat{\Sigma} \beta}{\|\beta_S\|_2^2} : \beta \in C(S; \alpha) \right\}$$

If γ_{re} is uniformly lower-bounded for all subsets S with cardinality k, $\hat{\Sigma}$ satisfies a restricted eigenvalue condition of order k with parameter α.

The restricted eigenvalue is effectively the submodularity ratio tailored to the Lasso and Dantzig selectors and generalized to hold for all response vectors Y. Previous work has demonstrated the connection between γ_{sr} and sparse eigenvalues (Das and Kempe, 2011). A sparse eigenvalue with parameter $k < p$ is

$$\lambda_{\min}(k) = \min_{\delta \in \mathbb{R}^k : 1 \leq \|\delta\|_0 \leq k} \frac{\delta^T \mathbf{X}^T \mathbf{X} \delta}{n \|\delta\|_2^2}.$$

In order to remove the dependence on Y in the definition of γ_{sr}, both the model S of size k and the comparison set L of size k need to be arbitrary. Therefore, $\gamma_{sr}(S, k) \geq \lambda_{\min}(2k)$. As discussed in Bickel et al. (2009), bounding the restricted eigenvalue bounds the minimum $2k$-sparse eigenvalue. Thus the data conditions under which the Lasso and Dantzig selector are guaranteed to be successful are stronger than those under which forward stepwise is. Granted, the form of the guarantees are significantly different, but of interest is the similarity

of the assumptions required.

The Lasso and Dantzig selector are known to over-estimate the support of β (Zou, 2006), and thus should not be compared to a sparse vector with k non-zero entries. The estimates β_l and β_d are elements of $C(S;\alpha)$ with probability close to 1 (Bickel et al., 2009). Therefore, the bound corresponding to submodularity needs to minimize over $C(S;\alpha)$ instead of truly sparse vectors. Given the looseness of $\lambda_{\min}(2K)$ as a lower bound on $\gamma_{sr}(S,k)$, we expect a similar looseness exists between the restricted eigenvalue and the corresponding Y-dependent bound. While it is useful to provide guarantees that do not depend on Y, the potential to produce a better estimate of the crucial constant at runtime may provide stronger practical performance guarantees. This development could mirror (Bertsimas et al., 2015).

4.3.2. SIS: Sure Independent Screening

SIS is a correlation learning method in which the marginal correlations between the response and all features are computed and the features with the largest d correlations are kept. This can be coupled with subsequent feature selection algorithms such as SCAD, Dantzig, or Lasso to select a final model from these d features. As an additional step, this process can be iterated in much the same way as stepwise regression: all remaining features are projected off of the selected set, and the process continues using the residuals from the first model. Therefore, iterated SIS is similar to a batch greedy method.

Fan and Lv (2008) split the assumptions for the asymptotic analysis into two groups: one focuses on parameters of the true regression function and the second focuses on the sampling distribution of the data. The assumptions on the true function are stronger than submodularity and the sampling distribution does not distort this. The most relevant assumption the authors make is the following:

Assumption 1. *Fan and Lv (2008) For some* κ, $0 \leq \kappa < 1/2$, *and* c_2, $c_3 > 0$,

$$\min_{i \in M_*} |\beta_i| \geq \frac{c_2}{n^\kappa} \quad \text{and} \quad \min_{i \in M_*} |Cov(\beta_i^{-1} Y, X_i)| \geq c_3.$$

This is of the same form as submodularity by:

$$\begin{aligned} |\text{cov}(\beta_i^{-1} Y, \mathbf{X}_i)| &= |\beta_i^{-1}||\text{cov}(Y, \mathbf{X}_i)| \\ &= |\beta_i^{-1}||r_{Yi}|, \end{aligned}$$

where the second line follows because $\mathbf{X}_i$ and Y are standardized. As r_{Yi} is the coefficient estimate when $\mathbf{X}_i$ considered marginally, Assumption 1 assures that features with non-zero coefficients in the true model have marginal correlations which are large enough to fall above the noise level. If S is the true model, this can be written in a similar form as submodularity as $\Delta(\mathbf{X}_i) \geq c_3 \Delta_{S \setminus i}(\mathbf{X}_i)$. This is the first order difference definition of submodularity when $A = \emptyset$. Furthermore, this is more restrictive than statistical submodularity since $\gamma_{sr} > 0$ merely requires that there exists at least one feature which increases the model fit when considered in isolation. Assumption 1 requires that all true features increase model fit when considered in isolation. Therefore, all relevant joint information is visible from correlations. It is impossible to hide signal in even two-dimensional subproblems such as those considered in Section 4.2.

4.4. Appendix

Proof of equivalence of Definition 7. Implications 3. $\Rightarrow$ 2. $\Rightarrow$ 1. are clear by appropriately defining the sets of interest as done when introducing the definitions of submodularity. To prove the reverse implications, we write lower-level definitions multiple times using nested sets. Summing these inequalities and simplifying gives the result.

To prove the first-order definition from the second-order definition, consider $B = A \cup \{b_1, \ldots, b_k\}$, and apply the second-order definition to sets $A_i' = A \cup \{b_1, \ldots, b_i\}$. This yields

a set of inequalities

$$
\begin{aligned}
\Delta_A(i) &\geq \gamma_{s2}\Delta_{A'_1}(i) \\
\Delta_{A'_1}(i) &\geq \gamma_{s2}\Delta_{A'_2}(i) \\
&\vdots \\
\Delta_{A'_{k-1}}(i) &\geq \gamma_{s2}\Delta_B(i) \\
\Rightarrow \Delta_A(i) &\geq \gamma_{s2}\Delta_B(i) + (\gamma_{s2}-1)\Delta_{A'_1}(i) + \ldots + (\gamma_{s2}-1)\Delta_{A'_{k-1}}(i) \\
&\geq \gamma_{s2}\Delta_B(i) + \frac{\gamma_{s2}-1}{\gamma_{s2}}\Delta_A(i) + \ldots + \frac{\gamma_{s2}-1}{\gamma_{s2}^{k-1}}\Delta_A(i) \\
&\geq \left(\gamma_{s2} + (1-\gamma_{s2})\frac{1-\gamma_{s2}^{-k}}{1-\gamma_{s2}^{-1}}\right)^{-1}\Delta_B(i)
\end{aligned}
\tag{4.10}
$$

where the second to last line follows from applying the second order definition repeatedly to convert $\Delta_{A'_i}$ to Δ_A. The constant in the last line provides a lower bound on γ_s and is always strictly positive if γ_{s2} is. It assumes that all of the individual steps are worst-case steps. As seen in Section 4.2.1, there are constraints on the number of steps that can be taken at this worst case level.

Similarly, to prove the standard definition from the first-order definition, apply the latter multiple times and sum the inequalities to produce $\Delta_A(C) \geq \gamma_s\Delta_B(C)$. Here $C = \{c_1, \ldots, c_k\}$ and $C \cap A = \emptyset$. Again, let $A'_i = A \cup \{c_1, \ldots, c_i\}$. Note that since $A \subset B$ this implies that $B'_i = B \cup \{c_1, \ldots, c_i\}$. This yields a set of inequalities

$$
\begin{aligned}
\Delta_A(c_1) &\geq \gamma_s\Delta_B(c_1) \\
\Delta_{A'_1}(c_2) &\geq \gamma_s\Delta_{B'_1}(c_2) \\
&\vdots \\
\Delta_{A'_{k-1}}(c_k) &\geq \gamma_s\Delta_{B'_{k-1}}(c_k) \\
\Rightarrow \Delta_A(C) &\geq \gamma_s\Delta_B(C)
\end{aligned}
$$

Where the last line follows by summing the previous lines, canceling most terms. $\forall S, T \subset$

$[m]$, set $A = S \cap T$, $C = S \backslash T$, and $B = T$. This yields the result. $\square$

CHAPTER 5 : ENSURING FAIRNESS IN ARBITRARY MODELS

Machine learning has been a boon for improved decision making. The increased volume and variety of data has opened the door to a host of data mining tools for knowledge discovery; however, automated decision making using vast quantities of data needs to be tempered by caution. In 2014, President Obama called for a 90-day review of big data analytics. The review, "Big Data: Seizing Opportunities, Preserving Values," concludes that big data analytics can cause societal harm by perpetuating the disenfranchisement of marginalized groups House (2014). Fairness aware data mining (FADM) aims to address this concern.

Broadly speaking, the goal of this project is to allow increasingly complex methods to be widely used without fear of infringing upon individuals' rights. This will be beneficial in all domains that have the potential for discrimination on the basis on data. Applications abound in both the private sector and academics. Companies will be able to justify the use of partially automated decision making in areas as diverse as loan applications, employment, and college admissions. There will be clear fairness criteria to guide the construction of fair models, thus reducing unintentional discrimination and litigation. A proper understanding of fairness will inform regulatory agencies and policy makers such that they can promote fairness and understand its statistical implications. In legal disputes, a set of fairness models provides a baseline from which detrimental impact can be assessed.

To clarifies the issue of fairness we reiterate the example from the introduction. Consider a bank that wants to estimate the risk in giving an applicant a loan. The applicant has "legitimate covariates" such as education and credit history, that can be used to determine their risk. They also have "sensitive" or "protected" covariates such as race and gender, which society does not want to be used to determine their risk. The bank's task is to model the credit risk or credit score C. To do so, they use historical data, estimate the credit worthiness of the candidate, then determine the interest rate of the loan. The question asked by FADM is whether or not the model the bank constructed is fair. This is different

than asking if the data are fair or if the historical practice of giving loans was fair. It is a question pertaining to the estimates produced by the bank's model. This generates several questions. First, what does fairness even mean in this statistical model? Second, what is the role of the sensitive covariates in this estimate? Lastly, how do we constrain the use of the sensitive covariates in black-box algorithms?

Multiple authors have raised doubts that the legal requirement of removing race prior to fitting a model is sufficient to achieve fairness Kamishima et al. (2012); Kamiran et al. (2013). Due to the relationships between race and other covariates, merely removing race can leave lingering discriminatory effects that permeate the data and potentially perpetuate discrimination. These equate this type of statistical discrimination to redlining. As we will demonstrate, their discussion is incomplete and conflates two different effects. We answer all of these questions and provide a post-processing method that corrects estimates from arbitrary models to achieve fairness. This work is primarily foundational as the literature is still debating the definition of fairness in statistics.

The use of historical data raises further concerns. First, if loans were provided in a discriminatory manner, the data set may not contain sufficient data on all relevant subpopulations for some analyses. Two effects can result: unfair estimates could perpetuate discriminatory lending practices, or missing information could lead to lower quality estimates of default rates for unobserved groups. The standard FADM problem, introduced by Pedreschi et al. (2008), often focuses on similar issues, where the data are the *result* of discrimination.

Historical data can lead to a second issue in which sufficient data is available for assessment but there is an observed difference in default rates based on a sensitive covariate. This possibility makes our loan example fundamentally different from, and more challenging than, the standard problem statement in FADM. The response is not the decision variable of an institution, but the result of individuals' behavior. Race may be an informative predictor in the sense that it improves the predictive ability of the model both in- and out-of-sample. As an example of this, Ridgeway (2016) identifies race of the officer as an

informative risk factor associated with police shootings.

Suppose the response in the loan example was the indicator of an applicant being given a loan instead of the risk of the applicant. If the bank used this data to estimate the probability that a loan was given, and only gave loans based on this measure, historical discrimination would bias the results. This case is easier to analyze since it posits that certain observable differences in the data are the result of discrimination, whereas they can be informative in our model. A more plausible business problem may be the automation of salary decisions. A statistical procedure derived from discriminatory, historical data may predict salaries that discriminate against women. Therefore, the standard FADM problem is a simple, special case of our framework. Furthermore, we answer the socially relevant discussion surrounding the use of sensitive information in general prediction tasks.

The problem is fundamentally about what constitutes "explainable variation." That is, what differences between groups are explainable due to legitimate covariates, and what differences are due to discrimination. More precisely, there are important distinctions between statistical discrimination and redlining. Statistical discrimination is defined as a sufficiently accurate generalization. In many ways, this is the statistical enterprise. For example, it is a sufficiently accurate generalization that individuals with good repayment history are likely to repay future loans. Therefore, such applicants are considered to be of lower risk and receive lower interest rates. The Equal Credit Opportunity Acts of 1974 and the amendments in 1975 and 1976 allow such "discrimination" if it is "empirically derived and statistically valid." It is clear that "discrimination" in this case refers to distinguishing good and bad risks.

The distinction between legal and illegal forms of statistical discrimination primarily arises due to which covariates are being used to make generalizations. The concept of a "sensitive" or "protected" characteristic in some ways prohibits its use for generalizations. For example, in the United States there are unfortunate relationships between incarceration and race; black males are significantly more likely to have been imprisoned at some point in their

lives than white males. Actions based on such heuristics are often illegal, though they may be economically rational. Risse and Zeckhauser (2004) provide a richer account of these cases, addressing concerns surrounding racial profiling. They separate the notion of statistical discrimination from the larger setting of societal discrimination. The debate often centers on what is a "disproportionate" use of sensitive information Banks (2001). In Section 5.1, we provide a detailed account of statistical discrimination and describe how proportionality can be measured.

Statistical discrimination is contrasted with redlining, which is a negative consequence of the ability to estimate sensitive covariates using legitimate ones. This can be used to discriminate against a protected group without having to see group membership. In this case, "discrimination" is used to describe prejudicial treatment and is a normatively negative concept. While often clear from context, in the interest of avoiding confusion between a legitimate type of statistical discrimination and redlining, "discrimination" will primarily be used in a normative, prejudicial sense. The exception is in the phrase "statistical discrimination," in which case we will be more precise. General statistical uses of "discrimination" will be described as "differentiating" individuals etc. This will separate the normative and statistical uses of the term "discrimination."

The term "redlining" originated in the United States to describe maps that were color-coded to represent areas in which banks would not invest. Figure 12 shows one such map from Philadelphia in the 1930s. It is marked in different colors to indicate the riskiness of neighborhoods. For example, red indicates hazardous areas and blue indicates good areas for investment. Denying lending to hazardous areas appears "facially neutral" because it is "race blind:" the bank need not consider racial information when determining whether to provide a loan. These practices, however, primarily denied loans to black, inner-city neighborhoods. This was used as a way to discriminate against such borrowers without needing to observe race. This clearly demonstrates that merely excluding sensitive information does not remove the possibility for discrimination. Conceptually, the core issue is the *misuse* of

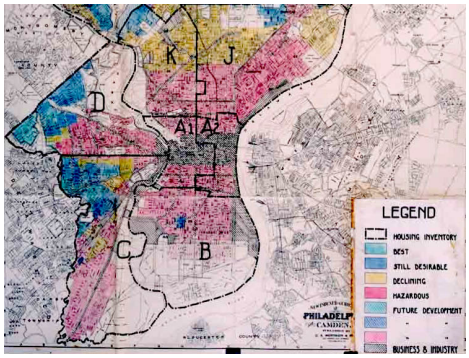


Figure 12: A 1936 map of Philadelphia marking high and low-risk areas.

available information. This will be an often-cited example in the remainder of the paper.

Our contributions fall into two main categories: conceptual and algorithmic. First, we provide a *statistical* theory of fairness. The literature is lacking a serious discussion of the philosophical components of fairness and how they should be operationalized in statistics. Doing so will require a spectrum of models to be defined, because fairness is a complicated philosophical topic. Furthermore, such a spectrum is necessary to capture varying notions of fairness present in multiple cultures. Second, after providing this framework it will be clear how to both construct fair estimates using simple procedures as well as correct black-box estimates to achieve fairness. It is important to note that these corrections can only be made by using all of the data, including the sensitive covariates. This is intuitively clear because guaranteeing that discrimination has not occurred requires checking the estimates using sensitive characteristics. Having a spectrum of fairness models also yields two practical applications. First, we can quantify the cost of government programs by considering the different incentives between a profit-maximizing firm and a government-owned, welfare-maximizing one. Second, we can quantify disparate impact, which is an important component of redlining litigation.

A few remarks need to be made about the sensitive nature of the topic at hand. Sensitive covariates such as race and gender are not neutral concepts. It is the plight of the data analyst that these categories are taken as given. We assume that data are provided in

which someone else has determined group membership. Our questions are about the types of protection that can be offered given such a data set. Furthermore, this project is descriptive, not normative. Our goal is to provide scientific, data-generating scenarios that elucidate philosophical nuances in statistical modeling. Each scenario gives rise to a different fair estimate; however, determining which scenario is accurate for a given culture is outside of the scope of this project. An important consequence of this is that different scenarios are accurate in different societies. For example, a society with more social mobility or equal opportunity can use different estimation techniques than those where discrimination is pervasive. For a further discussion of race in data analysis, interested readers are referred to Holland (2003).

The main body of the paper is organized as follows: Section 5.1 defines fairness in statistics. This requires careful consideration of the philosophical and legal underpinnings of fairness and is motivated by a long history of literature in ethics. This discussion constructively generates fair estimates using multiple regression. We also compare estimates on an individual level in Section 5.1.4. This has never been done in the literature but is crucially important if we intend to justify our models as fair. Often social discussions revolve around a individual being treated fairly, so it is ironic that the only discussion of fairness in the literature is about global properties. Section 5.2 uses the methods generated in Section 5.1 to correct estimates from black-box models to achieve fairness. In light of Section 5.1, this is a straight-forward task. We also test our methods on a data example to not only elucidate the conceptual difficulties that the literature has been having with fairness, but also to demonstrate that our method achieves superior results.

5.1. Defining Fairness

Fairness has been discussed extensively in the philosophy as it is both a highly complex and socially relevant topic. Any meaningful treatment of fairness-aware data mining needs to address a variety of viewpoints, as there is no consensus as to what "fairness" means. Colloquially fairness is considered as similar people are treated similarly. This posits the

need for a metric by which we can measure the similarity of individuals. FADM is motivated by the requirement that sensitive covariates such are not relevant measures of similarity in many applications; they are not a meaningful source of variability. As such, sensitive information needs to be "ignored;" however, the naive method of merely excluding the information, is obviously insufficient due to redlining. In this section, we describe a way to ignore the *influence* of these covariates.

More formally, the discussion of fairness-aware data mining revolves around the literature on equality of opportunity Arneson (2015). The philosophical literature on equality of opportunity analyses the way in which benefits are allocated in society. A benefit can be anything from a home loan or high salary to college admission and political office. One viewpoint is formal equality of opportunity (FEO), which requires an open-application for benefits (anyone can apply) and that benefits are given to those of highest merit. Merit will of course be measured differently depending on the scenario or benefit in question.[1] Therefore, the most productive employee receives the high salary, while the least-risky borrower receives a low interest rate loan. There is cause for concern if discrimination exists in either the ability of some individuals to apply for the benefit or in the analysis of merit. A constraint in FADM is the belief that sensitive features are not a relevant criteria by which to judge merit.[2]

Substantive equality of opportunity (SEO) contains the same strictures as above, but questions whether everyone has a genuine opportunity to be of high merit. In particular, suppose only those who received benefits in the past are of high merit. For clarity, consider a rigid caste system, where only the upper caste has the time and financial resources to educate and train their children. Only children born to upper-caste parents will be of high quality and receive the benefits. This is true even though lower-caste individuals can apply for the benefits and benefits are given based on merit. In this case, proponents of SEO claim

[1] Here we will ignore the randomness in estimates of quality.

[2] There are important legal exceptions such as business necessity allowed in the doctrine of disparate treatment.

that true equality of opportunity has not been achieved. While the United States is not a caste system, some may argue that the cycle of poverty may lead to a similar regress in the reasons for the disparity between races. It is important to draw the distinction between these two viewpoints, because it represents a large distinction in the philosophical literature as well as captures a significant component of the social debate around fairness. Concerns with SEO will be captured through the use of a new set of "suspect" covariates, that have not been discussed in the literature. This is explained in detail in Section 5.1.3.

In the United States, legal cases on equality of opportunity are based on two theories of discrimination outlined under Title VII of the U.S. Civil Rights Act. Disparate treatment is direct discrimination on the basis of a protected trait, and disparate impact is discrimination on the basis of another covariate which disproportionately effects a protected class. While initially introduced to govern employment, they have been expanded to other domains. For example, in June, 2015 the U.S. Supreme Court ruled that the Fair Housing Act extends these definitions to apply to housing dis (a). These concepts also govern legal cases in Europe, Australia, and New Zealand, though by other names such as "discrimination by subterfuge" or "indirect discrimination."

According to the U.S. Supreme Court, disparate treatment occurs when action is taken that "simply treats some people less favorably than others because of their race, color, religion, sex, or national origin" dis (b). It requires justification of the intent to discriminate based on the protected trait. An easy solution to prevent disparate treatment is merely to hide the information. Kamishima et al. (2012) termed this direct prejudice, providing the mathematical definition of its presence as conditional dependence of the response and sensitive covariates given the legitimate covariates.

Disparate impact occurs when "practices that are facially neutral in their treatment of different groups ... fall more harshly on one group than another and cannot be justified by business necessity" dis (c). Under this tenant, a policy is not discriminatory by definition (in that it does not codify treating groups differently) but is discriminatory in practice.

Kamishima et al. (2012) called this indirect prejudice, but incorrectly defined its presence as dependence of the response and sensitive covariates. Defining disparate impact requires a more refined notion of fairness, one that is able to capture the distinction between explainable variability and discrimination, see Section 5.1.1.

The canonical example of disparate impact is redlining. The bank treats all races equally within each neighborhood; however, it decides to build offices and provide loans in only select regions. While race is irrelevant in the statement of the policy, the racial homogeneity of many neighborhoods reveals this practice to be potentially discriminatory. While redlining occurs increasingly less often, two large cases were settled in Wisconsin and New Jersey in 2015. In Section 5.1.4, we provide a detailed numerical example that measures redlining precisely.

5.1.1. Mathematical Models of Fairness

All parties involved in our loan example have a vested interest in the model. The bank wants the best estimate of credit risk, society is interested in equality of opportunity, and the loan applicant wants to be treated fairly as an individual, not merely as a member of a group. All parties can be satisfied by acknowledging that "skin color," in its own right, has nothing to do with credit risk. If sensitive features appear to be informative, it indicates there are important excluded covariates. For example, there are unfortunate discrepancies between races in incarceration rates, income, and education. Such covariates may be legitimately predictive of credit risk, and their exclusion from the model will lead to the apparent importance of race.

Fairness assumptions will be explained via directed acyclic graphs (DAGs), which are also referred to as Bayesian or Gaussian networks or path diagrams. DAGs will be used to conveniently represent conditional independence assumptions. While often used as a model to measure causal effects, we are explicitly not using them for this purpose. As previously stated, our goal is to create fair *estimates*, whereas the estimation of a causal effect would

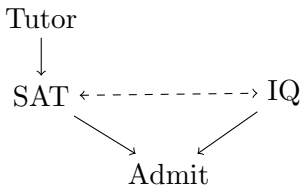


Figure 13: Example Directed Acyclic Graph (DAG)

attempt to answer if the *data* are fair.

Figure 13 provides an example DAG representing a possible set of variables and relationships in a simplified college admissions process. The variables are Tutor, SAT Score, IQ, and Admit. These indicate whether the student received SAT tutoring, their SAT and IQ scores, and whether they were admitted to a given university. In the graph, the variables are called nodes and are connected via directed edges. This direction captures a causal relationship: the value of Tutor causally relates to the value of SAT score. Dashed edged indicate a latent common cause: an unseen variable U that causally effects both nodes. We use DAGs to concisely represent conditional independence assumptions. In the language of DAGs, two nodes are called "d-separated" by a set of nodes B if all of the paths (series of edges, regardless of direction) connecting the two nodes are "blocked" by a collection of nodes. The only criteria for blocked paths we will use is the following: a path is blocked if it contains a chain $i \to b \to j$ such that b is in B. If two nodes are d-separated given B, then the nodes are conditionally independent given B. For example, Tutor and Admit are d-separated (conditionally independent) given SAT and IQ. For further information on DAGs, see Pearl (2009).

While our models take the form of DAGs similar to causal models used in economics or the social sciences, they do not require the same type of causal interpretation. This stems from a different object of interest: typically one cares about a parameter or direct effect of the model whereas in FADM we care about the estimates produced by the model. Estimating a causal effect requires considering counterfactuals. For example, the average treatment effect of a drug is the average difference in individuals' response when they are given the

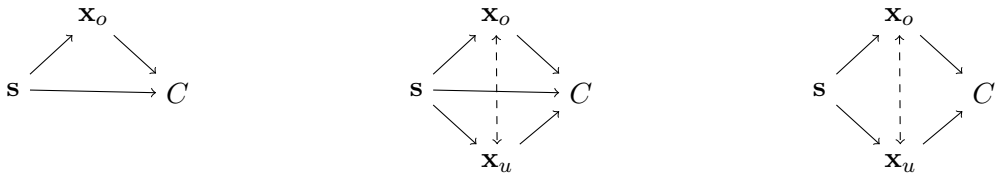
treatment versus control. Obviously patients are either given or not given the treatment. As such, he average treatment effect requires considering the counterfactual of their response under the alternative treatment.

Counterfactuals are easily computed in FADM by merely changing the observation on which the estimate is produced. The modular change is trivial to accomplish because it only requires changing the data set. We need not consider the performance of an individual with that set of covariates (or even if it exists). The hypothetical counterfactual exists in either instance and can be used to define fairness. Therefore, we do not need recourse to the interventionist or causal components of standard causal models and can deal only with their predictive components. In short, we only use DAGs to represents the conditional independence assumptions made between variables. Later we will need recourse to causal language, but it will not be for the estimation of treatment effects etc.

As a first step in operationalizing fairness in statistics, we provide models in which the assumption of fairness will be tractable. Consider an idealized population model that includes all possible covariates. For the i'th individual, C_i is credit risk, $\mathbf{s}_i$ contains the sensitive attributes (race, gender, age, etc), $\mathbf{x}_{o,i}$ contains the observed, legitimate covariates, and $\mathbf{x}_{u,i}$ contains the unobserved, legitimate covariates. Covariates $\mathbf{s}_i$, $\mathbf{x}_{o,i}$, and $\mathbf{x}_{u,i}$ are all bold to indicate they are column vectors. Unobserved covariates could be potentially observable such as drug use, or unknowable such as future income during the term of the loan. The data are assumed to have a joint distribution $\mathbb{P}(C_i, \mathbf{s}_i, \mathbf{x}_{o,i}, \mathbf{x}_{u,i})$, from which n observations are drawn. Society's fairness assumption is that $\mathbf{s}_i$ is not relevant to credit risk given full information:

$$\mathbb{P}(C_i = 1|\mathbf{s}_i, \mathbf{x}_{o,i}, \mathbf{x}_{u,i}) = \mathbb{P}(C_i = 1|\mathbf{x}_{oi}, \mathbf{x}_{u,i}).$$

It is important to posit the existence of both observed and unobserved legitimate covariates to capture the often observed relationship between sensitive covariates and the response.



(a) Observed Data (b) Unrestricted Data (c) Fair Data

Figure 14: Observationally Equivalent Data Generating Models

Specifically, observed data often show

$$\mathbb{P}(C_i = 1|\mathbf{s}_i, \mathbf{x}_{o,i}) \neq \mathbb{P}(C_i = 1|\mathbf{x}_{oi}.$$

This lack of independence violates the intuitive notion that sensitive features are uninformative.

Since assumptions like these will need to be presented many times, they will be succinctly captured using DAGs such as Figure 14. Observed data are often only representable by a fully connected graph which contains no conditional independence properties (Figure 14a). This observed distribution can be generated from multiple full-information models. The first possible representation of the full data is an unrestricted model (Figure 14b). In this case, sensitive covariates are not conditionally independent of the response given full information. Such a model states that there are different risk properties between protected groups even after considering *full* information. Society's motivation for fairness aware data mining is captured in Figure 14c: C is d-separated from s given x_u and x_o. Stated differently, credit risk is conditionally independent of the sensitive covariates given full information. Therefore, the apparent importance of sensitive information in the observed data is only due to unobserved covariates.

To motivate our construction of fair estimates we use a thought experiment similar to John Rawls' veil of ignorance: consider a hypothetical scenario in which two individuals apply for a loan. Their legitimate covariates are all identical but their sensitive attributes are different. We then decide which individual we would rather be for the purposes of

acquiring the loan. This is similar to the veil of ignorance if we consider being one of the two applicants, just unsure of which one we are. In this way, the sensitive information is "hidden" by the veil of ignorance. There are two different ways to consider fairness in this thought experiment. The first is "fairness as indifference" and the second is "fairness as no-direct-effect" of sensitive information.

Such a scenario has been experimentally tested using job applications in Chicago and Boston Bertrand and Mullainathan (2003). Researchers applied to many jobs using resumes which only differed in terms of the name used. One resume had a stereotypically black name while the other had a stereotypically white name. There was a significantly higher response rate to the resumes with the white name. In this case, it is clear that we are not indifferent between which resume we would rather use. This does not appear fair under the colloquial use of the term and violates the doctrine of disparate treatment. Therefore, indifference is a necessary component of fairness; however, this is not sufficient.

The historical examples of redlining demonstrate that one can be indifferent to sensitive characteristics but still receive unfair treatment. For example, suppose our applicants were both from "hazardous" areas in Philadelphia. Both black and white borrowers would be denied credit. We are indifferent to sensitive information because both groups are being discriminated against. Discrimination occurs because the information contained in location is being used incorrectly. Fairness is challenging because it requires estimating the response under an assumption that does not hold in the data; however, the unrestricted model in Figure 14b is equivalent to the fair model in Figure 14c if the direct effect of $\mathbf{s}$ on C is zero. This provides insight into the manner in which the fair estimate of C will be constructed: fairness requires constraining the sensitive covariate to have "no-direct-effect" on the estimates.

For clarity, this will be described using linear regression models of credit risk. Our goal is to describe fairness in general models: $\mathbb{P}(C_i|\mathbf{s}_i, \mathbf{x}_{o,i}) = f(\mathbf{s}_i, \mathbf{x}_{o,i})$; however, understanding the problem in a linear model provides not only tractable solutions but also insight into how the

goal can be accomplished in general. The insight is gained through properly understanding standard effect decompositions. Conceptually identical, but non-standard, decompositions also yield novel connections to both legal and philosophical standards for fairness.

Compare the classical full and restricted regression models. The full regression model includes both the sensitive and legitimate covariates as explanatory variables, while the restricted or marginal regression model only includes legitimate covariates as explanatory variables. Coefficients estimated in these models are given subscripts f and r, respectively. In both cases ϵ_m has mean 0, $m = f, r$.[3] While the notation is similar to that of multiple and simple regression models, respectively, covariates are potentially vector valued. Since this distinction is clear, we still use the terminology partial and marginal coefficients for the full and restricted models, respectively.

	Full Regression	Restricted Regression
Model	$C_i = \gamma_f + \mathbf{s}_i'\alpha_f + \mathbf{x}_{o,i}'\beta_f + \epsilon_f$	$C_i = \gamma_r + \mathbf{x}_{o,i}'\beta_r + \epsilon_r$
Estimated Coefficients	$\hat{\gamma}_f, \hat{\alpha}_f,$ and $\hat{\beta}_f$	$\hat{\gamma}_r$ and $\hat{\beta}_r$

A standard decomposition demonstrates that the marginal coefficient $\hat{\beta}_r$ can be represented as a function of the partial coefficients $\hat{\beta}_f$ and $\hat{\alpha}_f$ Stine and Foster (2013). This separates the marginal coefficient into direct and indirect effects:

$$\underbrace{\hat{\beta}_r}_{\text{marginal}} = \underbrace{\hat{\beta}_f}_{\text{direct}} + \underbrace{\hat{\nu}\hat{\alpha}_f}_{\text{indirect}} . \tag{5.1}$$

where $\hat{\nu}$ is estimated from

$$\mathbf{s}_i = \gamma + \mathbf{x}_{o,i}'\nu + \epsilon$$

As alluded to previously, an ideal case is when $\hat{\alpha}_f = 0$. Here the naive solution of ignoring $\mathbf{s}$ during model fitting does not change anything. In general, the same effect can be accomplished by fitting the full regression model to generate coefficient estimates, but making

[3]The only property used below is orthogonality inherent to regression procedures. No inference is conducted at this stage, so no further assumptions are required.

predictions that only use $\hat{\omega}_f$ and $\hat{\beta}_f$. This removes the *influence* of $\mathbf{s}$. While only legitimate covariates are used, coefficients must be estimated in the full model, else the relationship between sensitive and legitimate covariates will bias the estimated effect of $\mathbf{x}_o$. Such bias allows lingering discriminatory effects to bias the estimation. The estimates are fair under FEO since we are not addressing the possibility of discrimination in $\mathbf{x}_o$. The reverse regression literature in economics uses these estimates as a preprocessing step Goldberger (1984). That literature did not justify this as a fair estimate. We do so here and extend the estimates to more philosophically robust settings.

Definition 10 (Fair Estimate: Formal Equality of Opportunity)**.** $\hat{C}_i = \hat{\gamma}_f + \mathbf{x}'_{o,i}\hat{\beta}_f$

The standard decomposition in equation (5.1) can be presented in a non-standard way to yield additional insight. Collect the observations into matrices $\mathbf{C}$, $\mathbf{S}$, and $\mathbf{X}$ and consider writing the *estimated* response from the full regression. By decomposing this expression we can identify components which are of philosophical and legal interest. Separate the sensitive covariates into the component which is orthogonal to the legitimate covariates and that which is correlated with them. We will refer to these as the "unique" and "shared" components, respectively. It is important to note that the coefficient is computed only from the unique component. This decomposition can be done by considering the projection matrix on the column space of $\mathbf{X}_o$. For a general matrix $\mathbf{M}$, the projection or hat matrix is $\mathbf{H}_\mathbf{M} = \mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'$.

$$\begin{aligned}\hat{\mathbf{C}} &= \hat{\gamma}_f + \qquad\qquad \mathbf{S}\hat{\alpha}_f \qquad\qquad + \mathbf{X}_o\hat{\beta}_f \\ \hat{\mathbf{C}} &= \hat{\gamma}_f + \underbrace{\mathbf{H}_{\mathbf{X}_o}\mathbf{S}}_{di}\hat{\alpha}_f + \overbrace{\underbrace{(\mathbf{I} - \mathbf{H}_{\mathbf{X}_o})\mathbf{S}}_{dt}\hat{\alpha}_f} + \mathbf{X}_o\hat{\beta}_f \end{aligned} \tag{5.2}$$

The resulting terms are identified in equation (5.2) as *di* and *dt*, to indicate their legal significance. The term *di* captures the disparate treatment effect: it is the component of the estimate which is due to the unique variability of $\mathbf{S}$. Given the fair model in Figure 14c, we know the apparent importance of $\mathbf{S}$ (signified by the magnitude of $\hat{\alpha}_f$) is due to

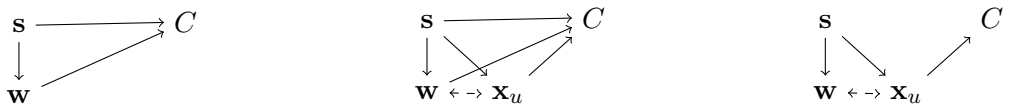
excluded covariates; however, it is identified by $\mathbf{S}$ in the observed data. While this may be a "sufficiently accurate generalization," this is illegal statistical discrimination and is the common way the term "statistical discrimination" is used in social science.

The term dt captures the disparate impact effect. We refer to it as the *informative* redlining effect in order to contrast it with an effect identified later. Intuitively, it is the misuse of a legitimately informative variable. It is the result of the ability to estimate $\mathbf{S}$ with other covariates. It is an adjustment to the influence of $\mathbf{X}_o$ that accounts for different performance between groups of $\mathbf{S}$. It is important that the adjustment is caused by variability in $\mathbf{S}$ instead of $\mathbf{X}_o$, as seen in equation (5.1). Identifying a disparate impact effect may be challenging because it is in the space spanned by the legitimate covariate, $\mathbf{X}_o$. The current legal solution merely removes the sensitive features from the analysis which allows for redlining via the term dt.

5.1.2. New Variable Type

The FADM literature currently only separates covariates into sensitive and legitimate groups. We relax this dichotomy and introduce a third class of "proxy" variables given by $\mathbf{w}$. Proxy variables, also known as information carriers, do not directly influence C given full information. While this is a similar property as $\mathbf{s}$, they are not considered to protected characteristics. A common example of an information carrier is location. Living in a particular location does not *make* someone of higher merit for most applications, but it may be indicative of things that do so. For example, suppose that information on education is missing in the data set. Location can be used as a proxy for education.

For simplicity, our current discussion will not include observed legitimate covariates. These are added in Section 5.1.5. The data are assumed to have joint distribution $\mathbb{P}(C_i, \mathbf{s}_i, \mathbf{x}_{u,i}, \mathbf{w}_i)$, from which n observations are drawn. As before, consider a linear model of credit score given the sensitive features and proxy variables. The coefficients given the subscript p, and



(a) Observed Data (b) Unrestricted Data (c) Fair Data

Figure 15: DAGs Using Proxy Variables

ϵ_p has mean 0.

$$\text{Proxy Model:} \quad C_i = \gamma_p + \mathbf{s}_i'\alpha_p + \mathbf{w}_i'\delta_p + \epsilon_p$$

$$\text{Estimated Coefficients:} \quad \hat{\gamma}_p, \hat{\alpha}_p, \text{ and } \hat{\delta}_p$$

DAGs similar to those in the previous subsection visually present the definition of proxy variables: $\mathbf{w}$ is conditionally independent of C given $\mathbf{x}_u$ (Figure 15c). When $\mathbf{x}_u$ is missing, $\mathbf{w}$ is often not conditionally independent from C. This is given in Figure 15a and can also be understood through decompositions similar to equation (5.1). As before, this observed data structure can be generated from multiple full data models which include the unobserved, legitimate covariates. An unrestricted data model, Figure 15b, posits no conditional independence between variables. This violates both the fairness assumption for sensitive features as well as the definition of proxy variables. The fair data model, Figure 15c, respects both of these constraints. This captures the intuition of a proxy variable; if location is a proxy for income, but income is already in the data set, then location will be uninformative.

One concern in using proxy variables in FADM is that most neighborhoods are racially homogeneous. The famous Schelling segregation models demonstrate that this happens even without the presence of strong discrimination Schelling (1971). Fair estimation in this setting should allow location to be a proxy for a legitimate variable but must *not* use location as a proxy for race or other sensitive covariates. Given the difficulty of exactly separating these two components, conservative estimates can be used to ensure that location is not use for redlining.

As before, decompose the estimates from the proxy model, where $\mathbf{W}$ is the matrix of proxy variables with $\mathbf{w}_i'$ as rows. All explanatory variables are separated into shared and unique components. In both decompositions, there are components identified as disparate treatment and disparate impact. Further information can be gained by considering the decompositions of $\mathbf{W}$ and $\mathbf{X}$ into their constituent parts. There is a unique component, unrelated to $\mathbf{S}$, as well as components labeled $sd+$ and $sd-$.

$$\hat{C}_f = \hat{\gamma}_f + \underbrace{\mathbf{H}_{\mathbf{X}_o}\mathbf{S}}_{di}\hat{\alpha}_f + \underbrace{(\mathbf{I}-\mathbf{H}_{\mathbf{X}_o})\mathbf{S}}_{dt}\hat{\alpha}_f + \underbrace{\mathbf{H}_{\mathbf{S}}\mathbf{X}_o}_{sd^+}\hat{\beta}_f + \underbrace{(\mathbf{I}-\mathbf{H}_{\mathbf{S}})\mathbf{X}_o}_{u}\hat{\beta}_f \tag{5.3}$$

$$\hat{C}_p = \hat{\gamma}_p + \underbrace{\mathbf{H}_{\mathbf{W}}\mathbf{S}}_{di}\hat{\alpha}_p + \underbrace{(\mathbf{I}-\mathbf{H}_{\mathbf{W}})\mathbf{S}}_{dt}\hat{\alpha}_p + \underbrace{\mathbf{H}_{\mathbf{S}}\mathbf{W}}_{sd^-}\hat{\delta}_p + \underbrace{(\mathbf{I}-\mathbf{H}_{\mathbf{S}})\mathbf{W}}_{u}\hat{\delta}_p \tag{5.4}$$

In equation (5.3), previous discussions of redlining do not distinguish between the terms di and $sd-$ Kamishima et al. (2012); Kamiran et al. (2013) because they are both due to the correlation between $\mathbf{X}_o$ and $\mathbf{S}$. It is clear that they are different, as $\mathbf{H}_{\mathbf{X}_o}\mathbf{S}$ is in the space spanned by $\mathbf{X}_o$ and $\mathbf{H}_{\mathbf{S}}\mathbf{X}_o$, is in the space spanned by $\mathbf{S}$. Furthermore, the coefficients attached to these terms are estimated from different sources. Intuition may suggest we remove all components in the space spanned by $\mathbf{S}$, but this is often incorrect. The term $sd+$ can be included in many models because it accounts for the group means of $\mathbf{X}$. Excluding $sd+$ implies that the *level* of $\mathbf{X}$ is not important but that an individual's *deviation* from their group mean is. This makes group membership a hindrance or advantage and is inappropriate for a legitimate covariate. For example, if an individual's group mean is low, a moderately high value will result in a large change in estimates. Similarly, if an individual's group mean is high, a moderately low value will have a large negative impact. These deleterious effects are due to group membership and not due to individual characteristics. Therefore, $sd+$ should be included if $\mathbf{x}$ is legitimate.

In the decomposition of the proxy model, equation (5.4), $sd-$ addresses the concern that the group means are potentially unfair or could be used to discriminate. For example, if $\mathbf{w}$ is location, $sd-$ accounts for the race distributions in a neighborhood. Given that

proxy variables $\mathbf{w}$ are not considered directly informative, it is unclear what this race distribution can legitimately contribute. If there is racial bias in the demographics of neighborhoods, using such information would perpetuate this discrimination. Ensuring that this is not perpetuated requires removing $sd-$ from the estimates of $\mathbf{C}$. This identifies a new type of redlining effect that we call *uninformative* redlining; it is the sum of di and $sd-$. Uninformative redlining can be identified visually using the graphs in Figure 15. Fairness requires consideration of the information contained in the arrow $\mathbf{s} \rightarrow C$ as well as information conveyed in the path $\mathbf{s} \rightarrow \mathbf{w} \rightarrow C$. This is because $\mathbf{s} \rightarrow \mathbf{w}$ is potentially discriminatory. Therefore, fair estimates with proxy variables only use the unique variability in $\mathbf{W}$. An important consequence of this estimate is that average estimates are the same for different groups of $\mathbf{s}$.

Definition 11 (Fair Estimate: Proxy Variables)**.** $\hat{C}_i = \hat{\gamma}_p + (\mathbf{I} - \mathbf{H_S})\mathbf{W}\hat{\delta}_p$

5.1.3. Substantive Equality of Opportunity

One objection to this model is the assumption that all $\mathbf{x}$ covariates are legitimate. Thus, while default can be explained in terms of $\mathbf{x}$ without recourse to $\mathbf{s}$, that is only because the covariates $\mathbf{x}$ are the result of discrimination. This critique stems from concerns over substantive equality of opportunity: different $\mathbf{s}$ groups may not have the same possibility of being of high merit as measured by $\mathbf{x}$. If this is driven by societal factors such as a class hierarchy or a cycle of poverty, these covariates may be suspect, and their use could perpetuate the disenfranchisement of historically marginalized groups. Such seemingly legitimate variables which are prejudicially associated with sensitive covariates need to be considered differently as they are simultaneously potentially informative and discriminatory. This class of "potentially illegitimate covariates" can be treated as proxy variables $\mathbf{w}$, given their conservative treatment: information from $\mathbf{w}$ can be used to estimate merit, but only in such a way that does not distinguish between groups in $\mathbf{s}$.

A full discussion of this topic requires positing different models for the relationship between sensitive and legitimate covariates. This lies at the heart of not only legal cases and social

science literature but also the public debates about fair treatment. The three models are independence, benign association, and prejudicial association. Often, independence between $\mathbf{s}$ and $\mathbf{w}$ is inaccurate but can be checked. As such, we focus on distinguishing the two types of association. As we will see in the simple example of Section 5.1.4, these models must be considered if one is to justify the types of estimates constructed in the FADM literature. This discussion is completely lacking, leaving all previous estimates unjustified.

In the benign association model, the relationship between $\mathbf{s}$ and $\mathbf{x}$ is not prejudicial. Here, differences in conditional distributions, $\mathbb{P}(\mathbf{x}|\mathbf{x})$, are the result of justified, individual choices. For example, suppose differences between groups are the result of different motivation via familial socialization. Broadly speaking, if some communities or cultures impart a higher value of education to their children than other cultures, the conditional distributions for educational attainment may be significantly different. These differences, however, appear legitimate and raise the question if the parents' rights to raise their children take priority over strict adherence to substantive equality of opportunity Arneson (2015); Brighouse and Swift (2009).

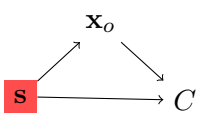
Alternatively, suppose the relationship between $\mathbf{s}$ and $\mathbf{x}$ is the *result* of either social restrictions or social benefits. For example, one group could be historically denied admission to university due to their group membership. This can produce the same observable differences between covariate distributions, in that the favored group has higher educational attainment than the disfavored group. In the context of fairness-aware data mining, these two cases need to be treated differently; however, differentiating them is beyond the scope of this paper. Our interest is not in specifying which variables have benign versus prejudicial association, but in constructing fair estimates once such a determination has been made.

Determining which case accurately describes a given society is crucially important, but is in the domain of causal inference and social science. Some cultures may differ in terms of which variables are prejudicially associated with sensitive information. To continue the education example, compare and contrast the Finnish education system with the United

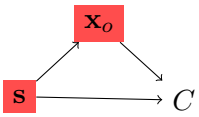
States education system. The Finnish system is predicated on equality of opportunity instead of educational quality. Regardless of the community in which students are raised, there is a reasonable expectation that they are provided the same access to education. In the United States, however, there are large differences in school quality, particularly between wealthy and poor areas. This may require education to be treated differently if one desires fair estimates in Finnish data or United States data.

We can now consider the philosophical differences between formal equality of opportunity and substantive equality of opportunity that are relevant to FADM. The motivation for incorporating substantive equality of opportunity is twofold: first, it addresses the concern that all people may not have a legitimate opportunity to be of high merit. In this way, substantive equality of opportunity is intimately connected with social mobility. It reflects a belief in a prejudicial, caste system along $\mathbf{s}$ with respect to a covariate $\mathbf{x}$. From this perspective, addressing the concern requires removing the group differences when making predictions using $\mathbf{x}$. A second motivation is to balance the allocation of benefits. The aim here is to attempt to move towards a more balanced society by increasing social mobility via cycles of poverty and wealth. The distinction between FEO and SEO revolves around whether there is benign or prejudicial association between variables, and changes the way a covariate must treated in our analysis.

Again, conduct a thought experiment using the veil of ignorance in the resume example. Instead of merely changing names on otherwise identical resumes, consider the following construction of a resume. Fix all information besides race and education; on one resume, put the stereotypical black name as well as a level of education sampled from the conditional distribution of education given race. Do the same for the resume with the stereotypical white name. This is a way to operationalize the veil of ignorance in statistical modeling: we must compare the two resumes constructed while hiding both race and education. Figure 16 visually demonstrates this by shading information that is hidden by the veil of ignorance. Asking for indifference between which resume to accept in this scenario treats education



(a) Formal EO



(b) Substantive EO

Figure 16: Veil of Ignorance Hides More Information

differently between the two groups. If education distributions differ, this places a stronger constraint to achieve fair treatment. This is only accurate in the setting where the covariate is partially the result of discrimination.

Fairness requires average group differences to be removed because averages differ due to discrimination. The veil of ignorance needs to hide all of the information on which average differences are the result of discrimination. Mathematically this is easy to accomplish in our framework. Notice that the story of average difference being unfair is the exact same as the story we provided for proxy variables. Therefore satisfying substantive equality of opportunity with respect to a suspect covariate merely requires treating that covariate as a proxy, instead of a legitimate, covariate. This allows the covariate to be meaningfully used without changing average differences in estimates between protected groups. Other papers have advocated a regression approach where all variables are considered proxy variables Calders et al. (2013). Without a proper understanding of the implications of this viewpoint, however, the results are highly unsatisfactory. This will be discussed in detail via example in Section 5.1.4.[4]

In light of this discussion, fair estimates are given as

Definition 12 (Fair Estimate: Substantive Equality of Opportunity)**.** $\hat{C}_i = \hat{\gamma}_p + (\mathbf{I} - \mathbf{H_S})\mathbf{W}\hat{\delta}_p$

[4]It is interesting to note that hiding *all* information behind the veil of ignorance is closely related to Rawlsian fair equality of opportunity and even results in his maxi-min perspective on social justice. These ideas are pursued elsewhere as they extend beyond FADM applications.

5.1.4. Simple Example

This section provides a simplified example to demonstrate different fair estimates. This has been overlooked in the literature, which favors providing a mathematical statement of discrimination for the set of predictions and demonstrating that the measure has been satisfied. It is important to understand what the estimates themselves look like. Claims about fairness are often made my individuals: the applicant in our loan example wants to be treated fairly. As such, it is ironic that such an analysis has never been conducted. We demonstrate how ignoring this is such a large oversight of previous works. Without a proper generative story "fair" estimates can appear decidedly *unfair*, possible drastically so.

Consider a simple example with only two covariates: income, x and sensitive group, s.[5] Suppose the data is collected on individuals who took out a loan of a given size. In this case, higher income is indicative of better repayment. As an additional simplification, suppose that income is split into only two categories: high and low. Lastly, to see the relevant issues, s and x need to be associated. The two sensitive groups will be written as $s+$ and $s-$ merely to indicate which group is, in general, of higher income. As such, the majority of the low income group is $s-$ and the majority of the high income group is $s+$. The response is the indicator of default D_i. The data and estimates are provided in Table 10, in which there exist direct effects for both s and x. This is consistent with the observed data DAGs in previous sections. As discussed at the end of the paper, the framework presented in this paper is equally applicable to binary data and generalized linear models. This case is simple enough that linear regression produces accurate conditional probability estimates. Therefore, different fair estimates can be directly analyzed for fairness.

Five possible estimates within our framework are compared in Table 10: the full OLS model, the restricted regression which excludes s, the formal equality of opportunity model in which income is considered a legitimate variable, the substantive equality of opportunity model in which income is considered a proxy or discriminatory variable, and the marginal model

[5]Variables are no longer in bold because they are not vectors.

Table 10: Simplified Loan Repayment data.

Income ($\mathbb{P}(x)$)	Low (.6)		High (.4)			
Group ($\mathbb{P}(s\|x)$)	$s-$ (.75)	$s+$	$s-$ (.25)	$s+$	Total	
Default Yes	225	60	20	30	335	
Default No	225	90	80	270	665	

	$\hat{\mathbb{P}}(D_i = 1\|x_i, s_i)$				DS	RMSE
Full Model	.5	.4	.2	.1	-.25	13.84
Exclude **s**	.475	.475	.125	.125	-.17	13.91
Formal EO	.455	.455	.155	.155	-.15	13.93
Sub. EO	.39	.535	.09	.235	0.00	14.37
Marginal	.35	.35	.35	.35	0.00	14.93

which estimates the marginal probability of default without any covariates. Estimates are presented along with the "discrimination score" and the RMSE from estimating the true default indicator. The discrimination score is merely the difference in the estimate probability of default between the two groups: $\hat{\mathbb{P}}(D = 1|s+) - \hat{\mathbb{P}}(D = 1|s-)$ Calders and Verwer (2010). This, however, is a misnomer since it does not separate explainable from discriminatory variation. It provides a useful perspective given its widespread use in the literature.

Since income is the only covariate that can measure similarity, the colloquial notion of fairness dictates that estimates should be constant for individuals with the same income. This is easily accomplished by the legal prescription of excluding s. If the information is unobserved, it cannot lead to differences in estimation. The formal equality of opportunity model satisfies this as well. As seen in the standard decomposition in equation (5.1) the only difference between the two estimates is the coefficient on x. Said differently, the term di in equation (5.2) lies in the space spanned by x. Therefore its removal only changes estimates for income groups. For reasons given in Section 5.1.1, the FEO estimates are considered fair. Excluding s permits redlining because it increases the estimated disparity between low- and high-income groups. This disproportionately effects those in $s-$ as they constitute the majority of the low income group. The FEO estimates result in some average differences between groups, but this is acceptable if the association between x and s is benign as in our

education example. This accurate measures the proportional differences desired by Banks (2001) for fair treatment.

The SEO estimates are highly counter-intuitive: although $s+$ performs better in our data set even after accounting for income, these "fair" estimates predict the opposite. Understanding this requires accepting the world view implicit in the SEO estimates: average income differences are the result of discrimination. Members of $\mathbf{s}-$ in the high income group have a much higher income than average for $\mathbf{s}-$. Similarly, members of $\mathbf{s}+$ who are in the high income group have a larger income than average for $\mathbf{s}+$, but not by as much since $\mathbf{s}+$ has a higher average income than $\mathbf{s}-$. The magnitude of these differences is given importance, not the income level. Intuitively, it implies that high-income, $s-$ individuals overcame a type of hindrance and may be of higher quality than corresponding $s+$ individuals that received a benefit. Therefore, the benefit the high-income, $s+$ group received due to being of high income is smaller than the benefit that the high-income, $s-$ group receives. The opposite is seen in the low income category. Low income, $\mathbf{s}+$ individuals have income much lower than average $\mathbf{s}+$ income, while low income, $\mathbf{s}-$ individuals have income lower than, but closer to average $\mathbf{s}-$ income. Therefore, low income $\mathbf{s}+$ individuals receive a larger detriment to being low income than $\mathbf{s}-$ individuals. In the SEO world view, low income $\mathbf{s}+$ individuals are low income *in spite* of having received a benefit.

These effects balance the differences in income distributions, resulting in both groups having the same average estimated default. This is seen in the discrimination score of 0. Without this framework of x being discriminatory, the SEO estimates discriminate against $s+$. All estimation methods previously considered in the literature produce estimates relevantly similar to SEO in this regard. This was never acknowledged because a direct comparison of the change in individual estimates was never provided. Furthermore, if not all $s+$ individuals are given a benefit or not all $s-$ individuals are given a detriment, then these models are merely approximations of the fair correction. An ideal protected or sensitive covariate s is exactly that which accounts for differences in the opportunity of being high merit. This is

more in line with Rawl's conception of equality of opportunity as pertaining chiefly to the socioeconomic status into which people are born Rawls (2001). While important, this line of inquiry is beyond the scope of this paper.

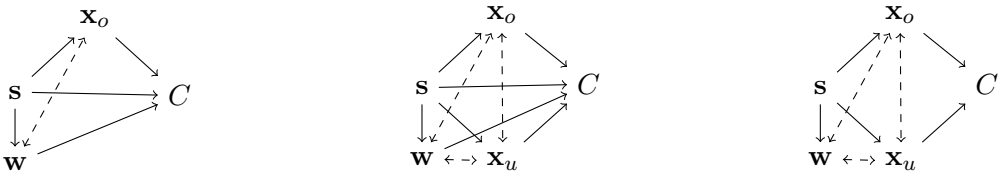
The SEO estimates show another important property: their RMSE is lower than that of the marginal estimate of default while still minimizing the discrimination score. As such, if a bank is required to minimize differences between groups in the interest of fairness, it would rather use the SEO estimates than the marginal estimate. The difference between the two estimates is that SEO still acknowledges that income in an informative predictor and contains an income effect. Furthermore, formal equality of opportunity is not satisfied in the marginal case because all merit information is ignored. See Arneson (2015) for a more detailed discussion.

5.1.5. Total Model

As a final level of complexity, consider models with sensitive, legitimate, and proxy variables. The data are assumed to have a joint distribution $\mathbb{P}(C_i, \mathbf{s}_i, \mathbf{x}_{o,i}, \mathbf{x}_{u,i}, \mathbf{w}_i)$, from which n observations are drawn. The relevant DAGs are given in Figure 17. A before, the observed data and unrestricted full data contain no conditional independence relationships. The fair data model captures the fairness constraint on $\mathbf{s}$ as well as the definition of the proxy variables $\mathbf{w}$. Again consider a simple linear model for clarity. Coefficients given the subscript t and ϵ_t has mean 0.

$$\begin{aligned} \text{Total Model:} \quad & C_i = \gamma_t + \mathbf{s}_i'\alpha_t + \mathbf{x}_{o,i}'\beta_t + \mathbf{w}_i'\delta_t + \epsilon_t \\ \text{Estimated Coefficients:} \quad & \hat{\gamma}_t, \hat{\alpha}_t, \hat{\beta}_t, \text{ and } \hat{\delta}_t \end{aligned} \tag{5.5}$$

The now familiar decomposition into unique and shared components is more complex because "shared" components exist across multiple dimensions. Equation (5.6) separates each term into its unique component and the component which is correlated with the other



Observed Data | Unrestricted Data | Fair Data

Figure 17: DAGs of Total Model

variables.

$$\begin{aligned}\hat{C} = \hat{\gamma}_t &+ \underbrace{\mathbf{H}_{[\mathbf{X}_o,\mathbf{W}]}\mathbf{S}}_{di}\hat{\alpha}_t + \underbrace{(\mathbf{I}-\mathbf{H}_{[\mathbf{X}_o,\mathbf{W}]})\mathbf{S}}_{dt}\hat{\alpha}_t \\ &+ \underbrace{\mathbf{H}_{[\mathbf{S},\mathbf{W}]}\mathbf{X}_o}_{sd^+}\hat{\beta}_t + \underbrace{(\mathbf{I}-\mathbf{H}_{[\mathbf{S},\mathbf{W}]})\mathbf{X}_o}_{u}\hat{\beta}_t \\ &+ \underbrace{\mathbf{H}_{[\mathbf{X}_o,\mathbf{S}]}\mathbf{W}}_{sd^-,sd^+}\hat{\delta}_t + \underbrace{(\mathbf{I}-\mathbf{H}_{[\mathbf{X}_o,\mathbf{S}]})\mathbf{W}}_{u}\hat{\delta}_t\end{aligned}$$

The sensitive features are again separated into disparate impact and disparate treatment components. Similarly, the legitimate variables are separated into permissible statistical discrimination and a unique component. The proxy variables, however, display an interesting property. There is the usual unique component but the correlated component is more complex. The term labels both $sd+$ and $sd-$ indicates that this combines both legal and illegal forms of statistical discrimination. The notation, $\mathbf{H}_{[\mathbf{X}_o,\mathbf{S}]}\mathbf{W}$, indicates that the shared component is the best linear estimate of $\mathbf{W}$ given both $\mathbf{S}$ and $\mathbf{X}$. This is the original problem of creating a fair estimate in the presence of sensitive and legitimate covariates. The only difference is that $\mathbf{W}$ is the response instead of $\mathbf{C}$.

Note that we assume the categories $\mathbf{s}$, $\mathbf{x}$, and $\mathbf{w}$ are given. This accounts for the distinction between formal and substantive equality of opportunity, allowing the definition below to be applicable in both scenarios.

Definition 13 (Fair Estimate: Total Model)**.** *General fair estimates are created with the following multi-step procedure:*

1. *Estimate the total model (5.5) to produce* $\hat{\gamma}_t$, $\hat{\alpha}_t$, $\hat{\beta}_t$, *and* $\hat{\delta}_t$.
2. *Create a fair estimate of* $\mathbf{W}$ *by*
   a. *Estimate* $\mathbf{w}_i = \gamma + \mathbf{s}_i'\alpha + \mathbf{x}_i'\beta + \epsilon$ *to produce* $\hat{\gamma}$, $\hat{\alpha}$ *and* $\hat{\beta}$.
   b. *Set* $\hat{\mathbf{W}} = \hat{\gamma} + \mathbf{x}_i'\hat{\beta}$.
3. *Set* $\hat{C} = \hat{\gamma}_t + \mathbf{X}_o\hat{\beta}_t + \hat{\mathbf{W}}\hat{\delta}_t + (\mathbf{I} - \mathbf{H}_{[\mathbf{X}_o,\mathbf{S}]})\mathbf{W}\hat{\delta}_t$.

5.1.6. Novel Applications

Now that fair estimates have been provided, a simple comparison can be made to identify the cost of disparate impact and governmental policies. The cost of fairness in these cases can be quantified by considering different actors as making the decision. Suppose that a privately owned, profit-maximizing bank is providing credit. As such, the best estimates are those which most accurately predict the riskiness of a loan while operating satisfying Title VII of the Civil Rights Act. Contrast this with a government-owned company that is interested in maximizing social utility or welfare instead of profit. This bank can constrain profit to achieve fairness goals. Conceptually, the cost of the government policy is the difference in the expected profit between the estimation methods these two banks would use.

The formal equality of opportunity models provide the minimally constrained fair estimates. A private bank may argue in favor of this method even if some covariates are prejudicial, because covariate generation is outside of the scope of the bank's operation. For example, discrimination in education is not within the power of the bank to change. That discussion, however, is well beyond the scope of this paper. Therefore, the private bank's estimates will be from the FEO model. The government may want to use a partially or fully substantive equality of opportunity model due to discrimination in the generation of some explanatory variables. The state-owned bank's estimates are from the corresponding SEO model. Using these estimates of risk, a bank would provide different loans at different rates.

One could consider that loans are provided at rates determined by the state-owned estimates

but whose cost to the bank is computed using the privately-owned estimates. The bank expects this to result in lower profit given its estimates of risk. The difference in expected profit between this loan scheme and the desired, profit-maximizing scheme is an estimate of the cost of the government program. Said differently, it is the price that the government should pay the bank for accomplishing the government's goals. Examples of the different estimates can be seen in our simplified data example in Table 10.

The cost of disparate impact can be estimated similarly. Instead of the reference being the ideal private bank and the state-owned bank, the comparison is between the rates the bank actually provided and what the ideal private bank *would* have provided given the same data. Therefore, instead of comparing the SEO and FEO models, one would compare the bank's actual lending history to what would be predicted under the FEO risk estimates. For example, suppose the bank followed the current minimal legal prescription of excluding the sensitive information. The difference between the actual and ideal estimates are again seen in our simplified data example in Table 10. The additional revenue the bank received due to their discriminatory lending practices can be estimated as the difference in the two estimation methods. Elaboration of these applications with real data is a subject of current research.

5.2. Correcting Estimates

Suppose that we have estimates given by an unknown model with unknown inputs. The model may use sensitive information to be intentionally or unintentionally discriminatory. This is a challenging but necessary case to consider. Most models used by private companies are proprietary. Society requires methods that can check if a model is fair merely by observing its output. Therefore, we must be completely agnostic as to the construction of these black-box estimates given as $C^{\dagger}$. It is perhaps surprising that fairness can be easily accomplish by considering such estimates as another explanatory variable similar to

a stacked regression:

$$C_i = \gamma_s + \mathbf{s}_i'\alpha_s + \mathbf{x}_{o,i}'\beta_s + \mathbf{w}_i'\delta_s + C_i^{\dagger}\lambda_s + \epsilon_s$$

Intuitively these black-box estimates are potentially predictive, but there is no guarantee that they are fair. This identically matches the description of proxy variables considered in Section 5.1.3. If we treat $C_i^{\dagger}$ as a proxy variable, its information can be used but not in a way that makes distinctions between protected groups. This allows us to easily correct black-box estimates.

A numeric example will solidify this idea and demonstrate the efficacy of our methods. Our data contains rankings of wine quality and is taken from the UCI Machine Learning Repository Lichman (2013). There are ratings for both red and white wines and a reasonable request is that ranking be "fair" between the two groups. This data captures all of the relevant issues that arise in FADM and was also considered in Calders et al. (2013). It captures the similarity between fairness-aware data mining and controlling for batch effects if data is aggregated from multiple sources.

Our response is the rating of wine quality measured between 0 and 10, the sensitive feature is wine type (red or white), and there are 10 explanatory variables such as acidity, sugar, and pH. Of the approximately 6,500 ratings, about 25% are for red wines. The distributions are very similar, as seen in Figure 18.

Calders and Verwer (2010); Calders et al. (2013) claim that the measure of "unfairness" in this case is given by the average rating difference between the two groups. This paper has argued that understand a bias in ratings requires explainable differences to be taken into account. This is evidence for treating the multiple regression coefficient as an estimate of unfairness. We do not advocate doing so for reasons made clear shortly, but the difference in measures is worth discussion. The two measures can disagree substantially, not only in magnitude but even in sign. The average difference in wine rating is .24, with white

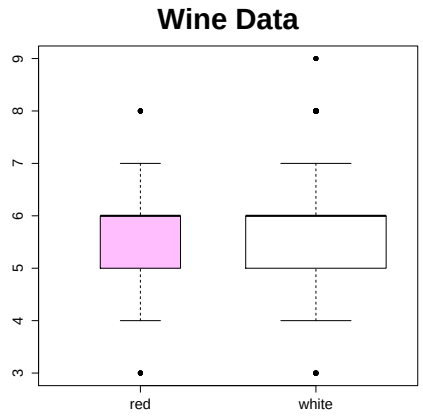


Figure 18: Histogram of wine data rankings.

wine being preferred; however, a multiple regression analysis indicates that white wine is, on average, rated .14 *lower* than red wines, ceteris paribus. The difference in sign is an example of Simpson's Paradox and is more generally related to the problem of suppression Johnson et al. (2015c).

Estimating the degree of discrimination in this case is a far more challenging problem than providing fair estimates. It is identical to asking for an estimate of the average treatment effect, where the treatment is wine type. It requires a set of causal assumptions to be made about the comparability of the two groups. In the language of causal inference, there may be insufficient covariate overlap between the two groups given the required differences between the two types of wine. This prevents accurate estimation of a fairness measure. The complexities inherent in estimating discrimination are insufficiently acknowledged in the FADM literature. Interested readers are directed toward Pearl (2009); Rosenbaum (2010) for introductions to various perspectives on causal inference.

Our validation framework follows that of Calders et al. (2013). Models will be trained on intentionally biased data and tested on the real data. Note that this is not exactly the desired measure; ideally one should test on appropriately "fair" data. This presupposes an answer to the problem at hand; however, by including bias in the training data the test

data is at least more fair. Data is biased by randomly selecting 70% of the white wines and increasing their rating by 1. This results in group mean differences (white-red) of .94 on the biased data and .24 on the raw data. Importantly the entirety of this bias is picked up by the multiple regression coefficient on wine type (WHITE). This gives more evidence to the massive difference in perspectives afforded by the multiple regression viewpoint. The multiple regression coefficients on the biased data is .56WHITE whereas the multiple regression coefficient on the raw data is -.14WHITE. All other coefficient estimates are unchanged. Therefore our fair estimates are exactly the same if produced using the biased or raw data. This is precisely the type of result necessary for fair algorithms; the multiple regression model is exactly reversing the bias added to the data. Admittedly, this is trivial since the bias is randomly attributed to 70% of white wines. That being said, this is heretofore unrecognized in the literature and not acknowledged in Calders et al. (2013).

The hope in this validation scenario is that the fairness constraint will improve performance on the test sample even when training on biased data. We compare our models with those of Calders et al. (2013) and indicate them by "Calders". They attempt to capture explainable variability through the use of propensity score stratification. This method estimates the probability that an observation is from a red or white wine. Observation are then binned into 5 groups based on this propensity score. In each group, a full substantive equality of opportunity model is fit where all explanatory variables are considered discriminatory. Therefore, within each strata, there is no average difference between groups. There is a large problem with this perspective since there is often enough information to predict the sensitive attribute almost perfectly. This results in largely homogeneous strata and the method fails. Calders et al. (2013) focus on a smaller problem with only two covariates in which this issue does not arise. Furthermore merely using 5 bins does not provide sufficient covariate overlap for the guarantees surrounding propensity score matching to hold. Other methods such as those of Kamiran et al. (2013) only handle a single sensitive and single legitimate covariate. Our estimates satisfy all of their fairness requirements in a more general and flexible framework.

Table 11 shows the results of using linear methods on this data. As a baseline measure, we use ordinary least squares (OLS), which is merely the estimates from the full model. This contains the disparate treatment and disparate impact effect. The formal equality of opportunity model treats all non-sensitive explanatory variables as legitimate, while the substantive equality of opportunity model treats all non-sensitive covariates as proxy variables. There is a spectrum of SEO models that consider different sets as fair or discriminatory. We chose to consider the full SEO model so as to permit easy comparison to the Calders estimates. Three performance measures are given: RMSE measured on the biased data (in-sample), RMSE on the test sample (raw data), and the mean difference in estimates between groups (DS).

Table 11: Regression Performance

	OLS	Formal EO	Sub. EO	Calders
RMSE-biased	0.84	0.85	0.93	0.86
RMSE-raw	0.95	0.92	0.91	0.92
DS	0.94	0.60	-0.00	0.79

The general direction of performance results are as expected: fairness constraints worsen performance on the biased data but improve performance on the test data. Even using linear regression models, we provide more accurate estimates of the test data while minimizing the mean difference between groups. This significantly improves upon the results of the Calders estimates. This is in spite of the fact that the Calders estimates require at least 6 times as many parameters. One set of parameters is used in the propensity score model to perform stratification, and additional parameters are estimated in each of the 5 strata.

From a different perspective, the results in Table 11 may not be particularly impressive in that wine ratings are not well estimated. This is to be expected, in part because wine quality is far more complex than these 10 explanatory variables, but also because rating is most likely not a linear function of the explanatory variables. Therefore, consider a more complex, black-box estimate of rating that can account for these nonlinearities. A random forest Breiman (2001) will be our canonical black-box as it an off-the-shelf method which

performs well in a variety of scenarios but has largely unknown complexity. It is challenging to consider how **s** should be used constructively in a random-forest algorithm while ensuring fairness. The estimates can be easily corrected, though, as outlined above.

Table 12 contains the results of the random forest models. The formal equality of opportunity model treats the random forest estimates as potentially discriminatory but the others as legitimate, while the substantive quality of opportunity model considers all variables as potentially discriminatory. These are again compared to the base-line random forest estimates and the Calders estimates. The general trend is the same as before: fairness constraints improve performance on the test sample but worsen performance on the biased sample. Using a random forest significantly reduced RMSE while not worsening the fairness measures. The corrected random forests estimates significantly outperform the Calders estimates along all measures. It does so even while exactly satisfying their desired constraint of zero mean difference between groups.

Table 12: Corrected Random Forest Performance

	RF	Formal EO RF	Sub. EO RF	Calders
RMSE-biased	0.73	0.74	0.83	0.86
RMSE-raw	0.87	0.82	0.81	0.92
DS	0.93	0.62	-0.00	0.79

CHAPTER 6 : DISCUSSION

This dissertation has discussed two frameworks for addressing inference during model selection and fairness-aware data mining. The different viewpoint afforded by these frameworks provides simple, algorithms that perform very well. We close by discussing insights provided by the algorithms and future projects in these domains.

6.1. Valid Stepwise Regression

Separation of the ranking and testing effects is important in its own right as it demonstrates that the current research on post-selection inference ignores the testing effect. Furthermore, the magnitude of the testing effect is often much larger than that of the ranking effect. Revisiting sequential methods are designed to control these selection effects, producing simple procedures that outperform the more complicated ones from the conditional inference framework.

The sequential selection framework provides many practical benefits. The core improvement is the ability to have dynamic algorithms that do not require a fixed set of hypotheses to be specified in advance. This allows for revisiting and directed search procedures. This flexibility also highlights the needs for new notions of false rejections. For example, identifying the interaction $\mathbf{X}_1\mathbf{X}_2$ in our framework requires both marginal terms to be included initially, even though neither is in the true model. The marginal terms are often considered false rejections even though, in the model in which they are tested, they capture significant signal. Progress on revising measures of false rejections to account for these considerations has been made by (G'Sell et al., 2013).

Future projects in this domain are as follows:

- One straightforward extension to Chapter 3 is to explore the use of Revisiting Alpha-Investing in generalized linear models. As previously discussed, type-I error control is maintained in this domain, but similar proofs for the performance of the procedure and

the implications of submodularity are unknown. This is partly due to submodularity being closely tied to linear models theory, but this can be extended to generalized linear models via weighted least-squares. Even in linear models, however, assessing the approximate submodularity of interaction spaces is an interesting and open problem.

- Further computational improvements can be achieved when approximating forward stepwise by using "lazy evaluation," which is an alternative algorithm to approximately sort the p-values. Lazy evaluation begins with a sorted list of marginal p-values from which the most significant feature is chosen. In order to select the correct second feature, forward stepwise recomputes all stepwise p-values conditional on the previously selected feature. Lazy evaluation merely recomputes the second smallest p-value. If, after recomputing, it is still smaller than the third smallest p-value, the corresponding hypothesis test is rejected. Else, the third smallest p-value is recomputed and compared to the fourth-smallest. The process continues until the corrected p-value is smaller than all p-values from the previous step and the corrected p-values in the current step. Under submodularity, this process is exactly the same as forward stepwise. The performance and error control of lazy evaluation under approximate submodularity are open questions.
- Instead of recomputing p-values after each feature is added to a model, one could only recompute p-values after each testing pass of a revisiting procedure is complete. This may result in a different set of features selected; however, rejected features fall into two sets which are intuitively appealing. The first set contains those features which would have been rejected in the revisiting procedures of Chapters 2 and 3. These features contain unique signal not captured by the others. The second set of features are "redundant" in that multiple features convey the same information. Loosely speaking, forward stepwise only selects one from each set. While this does reduce the proportion of seemingly false rejections, when correlation is high as it is in real data, it can be difficult to separate true features from the false features with which they are highly correlated. Including groups of features in this way is similar to a discrete version of the group lasso (Yuan and Lin, 2006).

- The Benjamini-Hochberg step-up procedures are "forward looking" in that they do not provide an accurate stopping time like alpha-investing rules. The power of alpha-investing can be increased if it is allowed to "peek" at the results of subsequent tests. The challenge in doing so is that the martingale properties of mFDR need to be preserved. By leveraging tests of intersection hypotheses and the closure principle we conjecture that some future information can be used. For example, instead of merely testing an individual hypothesis H_1, consider a test of the joint hypothesis $H_{[4]} = H_1 \cap \ldots \cap H_4$. If H_1 fails to be rejected but $H_{[4]}$ is rejected, it indicates that there is a false hypotheses in subsequent tests. Intuitively, if alpha-investing could take out a "loan" of error probability, it may be able to "repay" the loan when rejecting the future tests. Such behavior may preserve the martingale structure over sets of tests.

6.2. Submodularity

Submodularity plays an important role in statistics because it characterizes the difficulty of the *search problem* of feature selection. Assumptions used to prove the success of the Lasso, such as the restricted eigenvalue and restricted isometry properties, bound minimum sparse eigenvalues and hence are stronger assumptions than submodularity. Similarly, SIS requires true model features to have a bounded discrepancy between their joint coefficient in the true model and their marginal coefficient from a simple regression. Bounding this discrepancy is stronger than approximate submodularity as all true features cannot become vastly *more* significant in the presence of others. Similarly, worst case data examples can be crafted by intentionally breaking submodularity. This can be seen in Berk et al. (2013) and Miller (2002). Due to the importance of submodularity in discrete optimization, it provides a more theoretically robust assumption than those more commonly considered in statistics. Furthermore, it characterizes a different dimension of difficulty than the signal to noise ratio. As such, it is an important statistic to report in simulated data analyses.

Future projects in this domain are as follows:

- If data is submodular then the forward stepwise table will have non-decreasing p-values; however, it is not known the extent to which the converse holds. It clearly does not need to be the case that sorted p-values implies that there are no conditional suppressor variables in the entire data set. It can easily be the case that suppression occurs in the "insignificant" features which are not selected by forward stepwise for many steps; however, many feature selection assumptions only need to hold on the restricted set of relevant variables. Therefore, we conjecture that sorted stepwise p-values have an important connection to the submodularity of the highly significant variables.
- The graphs and discussion of Chapter 4 indicate that regardless of the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$, there exists a response Y such that the data is submodular. We conjecture that this is a general property for any correlation matrix with fewer than n features. This would force researchers to focus more on the nature of their true response function when using simulated data.

6.3. Fairness-Aware Data Mining

Our models ensure black-box estimates are fair and can be used to quantify regulation and disparate impact. All of this is done through providing a clear statistical theory of fairness and explainable variability. While we discuss the classical method of multiple regression, its power and interpretability were heretofore not well-understood in fairness-aware data mining. There are two key components that are required to generalize this approach to other methods. First, we identified the direct effect of $\mathbf{x}_o$ by estimating the regression model using all available information. This is required for the coefficient estimates in both the formal and substantive equality of opportunity models. Second, variables were residualized to remove the direct effect of $\mathbf{s}$ from $\mathbf{w}$. This was necessary for the substantive equality of opportunity models and to correct black-box estimates. Both generalized linear models and generalized additive models can perform these tasks. Therefore, the theory provided here transfers to those domains. A more thorough investigation of these settings is currently being conducted.

Future projects in this domain are as follows:

- More nuanced relationships can be identified by increasing the complexity of the similarity metric between individuals as well as the functional form of the response. For example, decision trees can be built to identify group of similar individuals.
- There are many open questions for FADM even in the linear setting. For example, how should one conduct fair inference? If estimates produced by a proprietary model are close but not identical to our methods, can we conduct a statistical test for similarity? This is crucially important in legal cases if fault must be demonstrated and is closely related to inference under model-misspecification Buja et al. (2014).
- Another major challenge is to relax the assumption of known covariate groups. A method to interpolate the categories may remove some of the inherent difficulties in the classification.

BIBLIOGRAPHY

Texas Department of Housing and Community Affairs et al. v . Inclusive Communities Project, Inc., et al, 576 U.S. 7 (2015).

Raytheon Co. v. Hernandez, 540 U.S. 44, 52, 13 AD 1825 (2003)(quoting Hazan Paper Co. v. Biggens, 507 U.S. 604, 610 (1993)).

International Board of Teamsters v. United States, 431 U.S. 324 (1977).

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Arneson, R. (2015). Equality of Opportunity. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2015 edition.

Bach, F. (2011). Learning with Submodular Functions: A Convex Optimization Perspective. *CoRR*, abs/1111.6453.

Bachoc, F., Leeb, H., and Pötscher, B. M. (2014). Valid confidence intervals for post-model-selection predictors. *arXiv preprint arXiv:1412.4605*.

Badanidiyuru, A. and Vondrák, J. (2014). Fast algorithms for maximizing submodular functions. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM.

Banks, R. R. (2001). Race-based suspect selection and colorblind equal protection doctrine and discourse. *UCLA Law Review*, 48.

Barron, A. R., Cohen, A., Dahmen, W., and DeVore, R. A. (2008). Approximation and learning by greedy algorithms. *Annals of Statistics 2008, Vol. 36, No. 1, 64-94*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.*, 41(2):802–837.

Bertrand, M. and Mullainathan, S. (2003). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Technical report, National Bureau of Economic Research.

Bertsimas, D., King, A., and Mazumder, R. (2015). Best subset selection via a modern optimization lens. *arXiv preprint arXiv:1507.03133*.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.*, 5(1):232–253.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brighouse, H. and Swift, A. (2009). Legitimate parental partiality. *Philosophy & Public Affairs*, 37(1):43–80.

Brown, L. D. and Johnson, K. D. (2016). Commentary on Exact Post-selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association*. Accepted, to appear.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhan, K., and Zhao, L. (2014). Models as Approximations: How Random Predictors and Model Violations Invalidate Classical Inference in Regression. *ArXiv e-prints*.

Buja, A. and Brown, L. (2014). Discussion: "A significance test for the lasso". *Annals of Statistics 2014, Vol. 42, No. 2, 509-517*.

Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. (2013). Controlling Attribute Effect in Linear Regression. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 71–80.

Calders, T. and Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292.

Candes, E. and Tao, T. (2007). The Dantzig Selector: Statistical Estimation When p Is Much Larger than n. *The Annals of Statistics*, 35(6):pp. 2313–2351.

Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215.

Das, A. and Kempe, D. (2008). Algorithms for subset selection in linear regression. In *STOC*, pages 45–54.

Das, A. and Kempe, D. (2011). Submodular meets Spectral: Greedy Algorithms for Subset Selection, Sparse Approximation and Dictionary Selection. In Getoor, L. and Scheffer, T., editors, *ICML*, pages 1057–1064. Omnipress.

Draper, N. R., Guttman, I., and Kanemasu, H. (1971). The Distribution of Certain Regression Statistics. *Biometrika*, 58(2):pp. 295–298.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2):407–451.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fithian, W., Sun, D., and Taylor, J. (2015). Optimal Inference After Model Selection.

Foster, D., Stine, R., and Wyner, A. (2001). Universal Codes for Finite Sequences of Integers Drawn From Monotone. *IEEE Trans. on Info. Theory*, 48:1713–1720.

Foster, D. P. and George, E. I. (1994). The Risk Inflation Criterion for Multiple Regression. *The Annals of Statistics*, 22(4):pp. 1947–1975.

Foster, D. P. and Stine, R. A. (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444.

Fujishige, S. (2005). *Submodular Functions and Optimization*. Number Vol. 58 in Annals of Discrete Mathematics. Elsevier B.V.

George, E. I. and Foster, D. P. (2000). Calibration and Empirical Bayes Variable Selection. *Biometrika*, 87:731–747.

Goldberger, A. S. (1984). Reverse Regression and Salary Discrimination. *The Journal of Human Resources*, 19(3):pp. 293–318.

G'Sell, M. G., Hastie, T., and Tibshirani, R. (2013). False Variable Selection Rates in Regression.

G'Sell, M. G., Wager, S., Chouldechova, A., and Tibshirani, R. (2015). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Hastie, T. and Junyang, Q. (2014). Glmnet Vignette. Technical report, Stanford.

Holland, P. W. (2003). Causation and Race. *ETS Research Report Series*, 2003(1):i–21.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

House, W. (2014). Big Data - Seizing Opportunities, Preserving Values. Technical report.

Jaggi, M. (2013). Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. In Dasgupta, S. and Mcallester, D., editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 427–435. JMLR Workshop and Conference Proceedings.

Johnson, K. D., Brown, L. D., Foster, D. P., and Stine, R. A. (2016). Valid Stepwise Regression. In Preparation.

Johnson, K. D., Lin, D., Ungar, L. H., Foster, D. P., and Stine, R. A. (2015a). A Risk Ratio Comparison of l_0 and l_1 Penalized Regression. *ArXiv e-prints*. http://arxiv.org/abs/1510.06319.

Johnson, K. D., Stine, R. A., and Foster, D. P. (2015b). Revisiting Alpha-Investing: Conditionally Valid Stepwise Regression. *ArXiv e-prints*. http://arxiv.org/abs/1510.06322.

Johnson, K. D., Stine, R. A., and Foster, D. P. (2015c). Submodularity in Statistics: Comparing the Success of Model Selection Methods. *ArXiv e-prints*. http://arxiv.org/abs/1510.06301.

Kamiran, F., Zliobaite, I., and Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl. Inf. Syst.*, 35(3):613–644.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. In Flach, P. A., Bie, T. D., and Cristianini, N., editors, *ECML/PKDD (2)*, volume 7524 of *Lecture Notes in Computer Science*, pages 35–50. Springer.

Kumar, R., Moseley, B., Vassilvitskii, S., and Vattani, A. (2013). Fast Greedy Algorithms in Mapreduce and Streaming. In *Proceedings of the Twenty-fifth Annual ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA '13, pages 1–10, New York, NY, USA. ACM.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.*, 34(5):2554–2591.

Lichman, M. (2013). UCI Machine Learning Repository.

Lin, D., Foster, D. P., and Ungar, L. H. (2011). VIF Regression: A Fast Regression Algorithm for Large Data. *Journal of the American Statistical Association*, 106(493):232–247.

Loftus, J. R. (2015). Selective inference after cross-validation. *arXiv preprint arXiv:1511.08866*.

Mallows, C. L. (1973). Some Comments on CP. *Technometrics*, 15(4):pp. 661–675.

Miller, A. (2002). *Subset Selection in Regression*. Monographs on statistics and applied probability . Chapman and Hall/CRC 2002, 2nd edition.

Natarajan, B. K. (1995). Sparse Approximate Solutions to Linear Systems. *SIAM J. Comput.*, 24(2):227–234.

Nemhauser, G., Wolsey, L., and Fisher, M. (1978). An analysis of approximations for maximizing submodular set functions—II. In Balinski, M. and Hoffman, A., editors, *Polyhedral Combinatorics*, volume 8 of *Mathematical Programming Studies*, pages 73–87. Springer Berlin Heidelberg.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition.

Pedreschi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In Li, Y. B. L. and Sarawagi, S., editors, *KDD*, pages 560–568. ACM.

Pope, P. T. and Webster, J. T. (1972). The Use of an F-Statistic in Stepwise Regression Procedures. *Technometrics*, 14(2):pp. 327–340.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted Eigenvalue Properties for Correlated Gaussian Designs. *Journal of Machine Learning Research*, 11:2241–2259.

Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.

Ridgeway, G. (2016). Officer Risk Factors Associated with Police Shootings: A Matched Case–Control Study. *Statistics and Public Policy*, 3(1):1–6.

Risse, M. and Zeckhauser, R. (2004). Racial Profiling. *Philosophy & Public Affairs*, 32(2):131–170.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer Series in Statistics. Springer-Verlag New York.

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2):143–186.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Shao, J. (1997). An Asymptotic Theory for Linear Model Selection. *Statistica Sinica*, 7:221–264.

Stine, R. and Foster, D. (2013). *Statistics for Business: Decision Making and Analysis*. Pearson Education Limited.

Taylor, J., Lockhart, R., Tibshirani, R. J., and Tibshirani, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 7.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tzelgov, J. and Henik, A. (1991). Suppression Situations in Psycological Research: Definitions, Implications, and Applications. *Psychological Bulletin*, 109(3):524–536.

van de Geer, S. A. (2007). The Deterministic Lasso. American Statistical Association. In JSM proceedings.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Statist.*, 3:1360–1392.

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, F. (2006). *The Schur complement and its applications*, volume 4. Springer Science & Business Media.

Zhang, T. (2008). Adaptive Forward-Backward Greedy Algorithm for Sparse Learning with Linear Models. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *NIPS*, pages 1921–1928. Curran Associates, Inc.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101:1418–1429.