Statistical Methodology 8 (2011) 42–55

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Variable selection for qualitative interactions

L. Gunter *, J. Zhu, S.A. Murphy

Department of Statistics, University of Michigan Ann Arbor, MI 48109, USA

ARTICLE INFO

Article history:
Received 28 November 2008
Received in revised form
5 May 2009
Accepted 9 May 2009

Keywords:
Decision making
Variable selection
Depression
Machine learning

ABSTRACT

In this article, we discuss variable selection for decision making with a focus on decisions regarding when to provide treatment and which treatment to provide. Current variable selection techniques were developed for use in a supervised learning setting where the goal is prediction of the response. These techniques often downplay the importance of interaction variables that have small predictive ability but that are critical when the ultimate goal is decision making rather than prediction. We propose two new techniques designed specifically to find variables that aid in decision making. Simulation results are given, along with an application of the methods on data from a randomized controlled trial for the treatment of depression.

1. Introduction

This article describes two new methods to select variables useful for decision making. Applications that deal with decision making occur in many different fields such as computer science, engineering, economics and medicine. In medicine, deciding when a patient needs treatment and which treatment is best are critical decisions. Clinical trials, particularly those involving heterogeneous patients, collect a large amount of potentially useful information that can aid in making these decisions. However, in clinical practice much of this information is expensive, time-consuming, and/or burdensome to collect. Thus, variable selection is needed to help inform the clinicians which variables are most important.

Variable selection techniques have been developed to enhance prediction, but their use in decision making has not been well tested. In our research we have found these techniques often miss or downplay the importance of certain interaction variables that are key to making decisions. The variable selection techniques we propose focus on finding these important interactions.

* Corresponding author.
E-mail addresses: lgunter@umich.edu (L. Gunter), jizhu@umich.edu (J. Zhu), samurphy@umich.edu (S.A. Murphy).

1572-3127/$ – see front matter
doi:10.1016/j.stamet.2009.05.003

This work is motivated in part by the the Nefazodone CBASP trial data. The Nefazodone CBASP trial [1] was a randomized controlled trial conducted to compare the efficacy of three alternate treatments for patients with chronic depression. The study randomized 681 patients with non-psychotic chronic major depressive disorder (MDD) to either Nefazodone, cognitive behavioral-analysis system of psychotherapy (CBASP) or the combination of the two treatments. Analysis of the trial data showed the combination treatment to be superior to the two singleton treatments overall. We wanted to know whether this relationship held true for all subsets of patients, and if not, to discover which patient characteristics help to determine the optimal depression treatment for an individual patient.

The remainder of this article is organized as follows: Sections 2 and 3 give background material on optimal decision making and discuss what makes a variable important for decision making. Section 4 provides two ranking techniques designed to find variables useful for decision making followed by an algorithm for using these techniques. Section 5 presents some simulation experiments, and Section 6 illustrates the methods using data from the Nefazodone CBASP study. Concluding remarks are given in Section 7.

2. Optimal decision making

We consider variable selection in the simplest decision-making setting in which one must decide between two actions. The idea is to use observations about a subject $X = (X_1, X_2, \ldots, X_p)$, to choose a treatment action A. Following the action, a response occurs. The response, R, gives us some indication of the desirability of the chosen action. The goal is to choose actions that maximize the response. A policy, π, is a stochastic or deterministic decision rule mapping the space of observations, X, to the space of the action, A. In other words, π defines the probability of choosing action $A = a$ given the observations $X = x$. So the goal can be restated as finding a policy π^* that maximizes the response.

A simple example of a decision-making problem is a clinical trial to test two alternative drug treatments. The observation vector, X, would consist of baseline variables, such as the patient's background, medical history and current symptoms. The action would be the treatment assigned to the patient, and the response could be the patient's condition or symptoms after receiving treatment. The goal is to determine which treatment is optimal for any given future patient, using the data obtained in the trial.

Alternate policies can be compared via the expected mean response, called the Value of a policy [2]. Let the distribution of X be a fixed distribution f, and let the distribution of R given (X, A) be a fixed distribution g. Then when actions are chosen according to a policy π, the trajectory (X, A, R) has distribution

$$f(x)\pi(a|x)g(r|x, a), \tag{1}$$

If $E_\pi[\,]$ denotes the expectation over the above distribution, then the Value of π is

$$V_\pi = E_\pi[R].$$

The optimal policy, π^*, is then defined as

$$\pi^* = \arg\max_\pi V_\pi = \arg\max_\pi E_\pi[R],$$

or equivalently

$$\pi^*(x) = \arg\max_a E[R|X = x, A = a].$$

If we knew the multivariate distribution of (X, A, R), the best treatment for future use could be found by calculating $E[R|X = x, A = a]$ for every possible (x, a) combination and then selecting the action leading to the highest conditional expectation of R for each x. In practice, however, we do not know this distribution. So we must use data to estimate the optimal future treatment. We do this by first estimating $E[R|X = x, A = a]$ for each (x, a) using a predictive model and learner, such as a multiple linear regression. We then use the estimated regression function to 'estimate' the best future treatment for each x. For example, if we used the data to estimate $E[R|X = x, A = a]$ by

$$\hat{E}[R|X = x, A = a] = \hat{\beta}_0 + x\hat{\beta}_1 + a\hat{\beta}_2 + xa\hat{\beta}_3,$$

for $a \in \{0, 1\}$, our estimated optimal future treatment actions would be

$$\hat{\pi}^*(x) = I(\hat{\beta}_2 + x\hat{\beta}_3 > 0).$$

3. Variable selection

There are multiple reasons why variable selection might be necessary in a decision making application. One reason is that finding the optimal policy becomes more difficult as the number of spurious variables included in the model increases. Thus, careful variable selection could lead to better policies. Also, due to limited resources, it may only be possible to collect a small number of variables when enacting a policy in a real world setting. Researchers are often unsure which variables would be most important to collect. Variable selection techniques could help identify these variables. In addition, policies with fewer variables are often easier to understand, so variable selection can improve interpretability.

Currently, variable selection for decision making in many fields is predominantly guided by expert opinion. Expert opinion can be a good starting place when there is sufficient domain knowledge and expertise. Some predictive variable selection techniques, such as Lasso [3], have been suggested [4]. In clinical trials, a combination of predictive variable selection techniques and statistical testing of a small number of interaction variables suggested by expert opinion are most commonly used [5–7]. Little research has been carried out to evaluate these techniques in decision making, or to suggest how they might be improved.

When selecting variables for decision making, a distinction should be made between variables that are included merely to facilitate estimation as opposed to variables involved in the decision rules. *Predictive* variables are variables used to reduce the variability and increase the accuracy of the estimator. Variables that help prescribe the optimal action for a given patient are *prescriptive* variables [8]. For optimal estimation results, it is best to select both types of variables. However, only *prescriptive* variables need to be collected when implementing the policy.

For a variable to be *prescriptive*, it must have a qualitative interaction with the action [9]. A variable X_j is said to qualitatively interact with the action, A, if there exists at least two distinct, non-empty sets, $S_1, S_2 \subset space(X_j)$ for which

$$\arg\max_a E[R|X_j = x_{j1}, A = a] \neq \arg\max_a E[R|X_j = x_{j2}, A = a],$$

for all $x_{j1} \in S_1$, and $x_{j2} \in S_2$. These variables are useful for decision making because they help decipher which action is optimal for each individual patient.

To illustrate this idea, see the plots in Fig. 1. These plots depict different possible relationships between the conditional mean of R, A and a particular X_j, when averaging over all other X_i, $i \neq j$. Fig. 1(a), shows a variable, X_1, which does not interact with the action. Fig. 1(b) shows a variable, X_2, that interacts with the action, A, but does not qualitatively interact with the action. In both plots, the optimal action is $A = 1$. Knowledge of X_1 or X_2 is useful for predicting the response for a given action, but should not affect which action should be chosen. Fig. 1(c), shows a variable, X_3, which qualitatively interacts with the action. We can see that the optimal action in this plot is $A = 0$ when $X_3 \leq 0.5$, and $A = 1$ when $X_3 > 0.5$. Knowledge of X_3 impacts the best choice of the action and likewise the response, thus it is important for decision making.

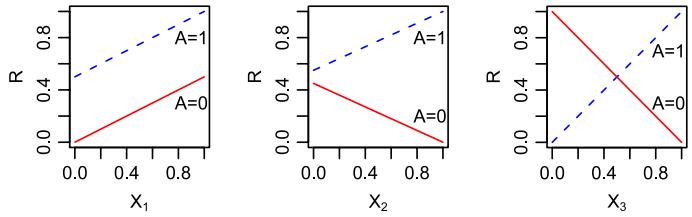
The degree to which a prescriptive variable is useful depends on two factors:

(1) *Interaction*: the magnitude of the interaction between the variable and the action. For an action with two possible values, $A \in \{0, 1\}$, this is the degree to which the following quantity varies as x varies

$$E[R|X = x, A = 1] - E[R|X = x, A = 0]. \tag{2}$$

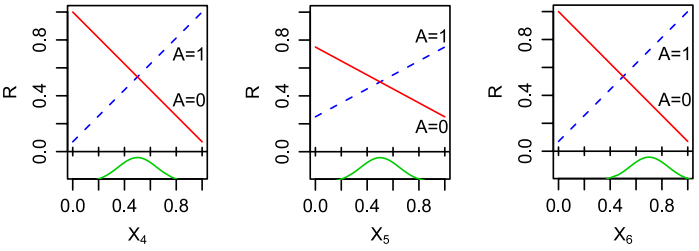
(2) *Proportion*: the proportion of patients whose optimal choice of action changes given a knowledge of the variable. If $a^* = \arg\max_a E[R|A = a]$, this is the proportion of patients for which the following holds:

$$\arg\max_a E[R|X = x, A = a] \neq a^*. \tag{3}$$



(a) No interaction. (b) Non-qualitative interaction. (c) Qualitative interaction.

Fig. 1. Plots demonstrating qualitative and non-qualitative interactions.



(a) Large interaction large proportion. (b) Small interaction, large proportion. (c) Large interaction, small proportion.

Fig. 2. Plots demonstrating usefulness factors of qualitative interactions.

Consider the plots in Fig. 2. Fig. 2(a) shows the relationship between the conditional mean of R, A, and a variable X_4, with an underlying plot giving the distribution of X_4. Fig. 2(b), (c) are similar to Fig. 2(a), but for variables X_5 and X_6. Notice that X_4 and X_5 have the same distribution. However, the interaction between X_4 and A is much stronger than the interaction between X_5 and A. Therefore, the effect of choosing the optimal action is much greater given X_4 than it is given X_5. Now notice that X_4 and X_6 have the same relationship with the conditional mean of R and A but are distributed differently. The distribution of X_4 is centered at the point of intersection, so half of the subjects would do better choosing $A = 0$ over $A = 1$. Whereas, the proportion of patients benefiting from choosing $A = 0$ is much smaller with X_6. Thus X_4 would be more useful in decision making than X_5 or X_6.

Since both of these factors also affect the predictive ability of a qualitative interaction, it may not be readily apparent why current variable selection techniques designed for prediction are not well equipped to find these prescriptive variables. One reason current variable selection methods aimed at prediction may have problems detecting prescriptive variables could be due to the way prescriptive variables occur in nature. In real world applications, individual variables, rather than interactions between variables, tend to explain most of the variation in the outcome and thus are most important for good prediction. This individual effect a variable has on the response is often referred to as the 'main effect' of the variable and most variable selection techniques are good at finding main effects. Furthermore, while non-qualitative treatment–covariate interactions do occur quite frequently in real world applications, it is commonly assumed that qualitative interactions are rare in nature [10,11].

There is an abundance of literature discussing qualitative interactions (e.g. [12,13,9,14,15,11,16]). Much of the statistical literature suggests that the search for qualitative interactions should be severely limited and qualitative interactions that are found should be initially mistrusted [9,14,11,16]. This point of view is fueled by a myriad of papers publishing claims of finding a qualitative interaction during exploratory data analysis of a controlled trial followed by subsequent studies in which the interaction did not replicate [17,11].

Skepticism concerning the validity of qualitative interactions is partially due to the way many clinical trials are conducted. Entry criteria are restrictive for many clinical trials. This results in data with minimal variation in the X variables, representing only a small subset of the treatable population.

In this case it is often reasonable to assume there are no genuine qualitative interactions within the range of the data. However, this does not imply that genuine qualitative interactions do not exist over the range of X for the entire treatable population. For this reason, the methods we present are most useful when applied to data representative of the entire treatable population (or at least a substantial proportion of the population).

Skepticism also exists due to the way analyses of clinical trials are reported. Since journals traditionally publish only significant results, it is tempting to comb through the data in post hoc analysis looking for anything that is significant and interesting. Many times, this "data fishing" includes looking for significant qualitative interactions. When significance levels are not corrected for the number of tests performed, researchers can often find at least one significant qualitative interaction spuriously. This problem would not occur as much if researchers were more forthcoming to journals about the number of tests they performed and the significance levels they used [18]. However, it is important to note that genuine qualitative interactions with small effect sizes will be undetectable in some data sets, especially those of small sample size. Despite this skepticism, medical researchers continue to look for qualitative interactions. They look for them because it is an underlying goal of clinical research to find the best treatment for each individual patient.

Our goal is to develop approaches that assist in finding qualitative interactions, but are less susceptible to finding spurious results. To address the concern that the proposed methods are equivalent to testing large numbers of interactions with uncorrected significance levels, we thoroughly test them on simulated data generated without qualitative interactions. Beyond this, we feel it is also important to emphasize that the goal of these methods is not to find the 'correct' underlying model. Rather, the driving force for this variable selection is to facilitate and improve decision making by reducing the number of variables that need to be considered when constructing a policy.

4. Qualitative interaction ranking

As we discussed in the previous section, variable selection in decision making should focus on variables that qualitatively interact with the action. In this section we will present two variable-ranking techniques that rank the variables in X based on their potential for a qualitative interaction with the action variable. We conclude the section with a proposed complete algorithm for variable selection in a decision making application.

The first variable-ranking method is based upon the two usefulness factors for a qualitative variable discussed in the previous section (see quantities (2) and (3)). Assume we have a data set of n subjects, with p baseline observations taken on each subject, making up the $n \times p$ observation matrix X. Also assume that in the data the action, $A = \{0, 1\}$ is randomized. The response is denoted by R. Consider the evaluation of the jth variable, X_j (the jth column of X). Then, given an estimator of $E\left[R|X_j = x_j, A = a\right]$, say $\hat{E}\left[R|X_j = x_j, A = a\right]$, define the following quantities for $j = 1, \ldots, p$:

$$D_j = \left(\max_{1 \le i \le n} \left(\hat{E}\left[R|X_j = x_{ij}, A = a^*\right] - \hat{E}\left[R|X_j = x_{ij}, A \neq a^*\right] \right) - \min_{1 \le i \le n} \left(\hat{E}\left[R|X_j = x_{ij}, A = a^*\right] - \hat{E}\left[R|X_j = x_{ij}, A \neq a^*\right] \right) \right) \tag{4}$$

and

$$P_j = \frac{1}{n} \sum_{i=1}^{n} 1\left\{ \arg\max_a \hat{E}\left[R|X_j = x_{ij}, A = a\right] \neq a^* \right\}, \tag{5}$$

where $1\{\cdot\}$ is 1 if '$\cdot$' is true and 0 otherwise and $a^* = \arg\max_a \hat{E}\left[R|A = a\right]$ is the overall optimal action.

Here D_j is a measure of the magnitude of the interaction. The P_j is a measure of the proportion of subjects affected by a change in the optimal choice of action due to the inclusion of an interaction

involving X_j. These two quantities can be combined to make a score, U_j, for ranking the variables:

$$U_j = \left(\frac{D_j - \min_{1\le k\le p} D_k}{\max_{1\le k\le p} D_k - \min_{1\le k\le p} D_k} \right) \left(\frac{P_j - \min_{1\le k\le p} P_k}{\max_{1\le k\le p} P_k - \min_{1\le k\le p} P_k} \right). \tag{6}$$

The first term in parentheses provides the relative (as compared to the other variables in X) magnitude of X_j's interaction with the action; the second term in parentheses provides the relative proportion of affected subjects in X_j used to select the action. The U_j is a product because we want to select X_j only if both D_j and P_j are relatively large. The first variable ranking procedure will rank variables in terms of their U_j.

The second ranking procedure looks directly at the expected increase in the estimated optimal Value due to the knowledge of the variable X_j. Recall, the Value is the expected response. The ranking procedure estimates the quantity described by [19] as the value of information. Define the score S_j as

$$S_j = \sum_{i=1}^{n} \left[\max_a \hat{E}\left[R|X_j = x_{ij}, A = a\right] - \hat{E}\left[R|X_j = x_{ij}, A = a^*\right] \right]. \tag{7}$$

Both of these scores, U and S, can be used to rank the variables. They have been defined generically to allow different models for $E[R|X, A]$. In the numerical section that follows, we use a linear model to estimate the conditional expectation and obtain $\hat{E}$.

Although not explicitly shown in the notation, predictive variables may also be used in the estimation of the conditional expectation. When testing for the interaction between X_j and A, researchers often prefer to maintain a hierarchical ordering [20] and thus the main effect of the variable X_j and the main effect of the action should be included. This helps to avoid finding spurious interactions that may appear because the main effect is important but is not included in the estimation. It is also wise to include other important main effects of the variables in X on R to help reduce variability in the estimation.

4.1. Variable selection algorithm

The following is an overview of an algorithm for variable selection.

(1) **Select Important Predictors:** Select important predictive variables of R among $(X, A * X)$ using a Lasso with the penalty parameter chosen by Bayesian Information Criterion (BIC).
(2) **Rank Interactions Individually:** Rank the variables in X using either U or S. Use the main effect variables selected in step 1 to help decrease the variability in the estimator $\hat{E}$. Select the top H variables in rank, where $H =$ the number of variables having non-zero U or S scores.
(3) **Create Nested Subsets of Chosen Predictive and Prescriptive Variables:**
  (a) Collect the following K variables:
    (i) the predictive variables chosen in step 1,
    (ii) the main effects of the top H ranked variables in step 2, and
    (iii) the interactions between A and the top H ranked variables in step 2.
  (b) Run a weighted Lasso using a weighting scheme that satisfies the following properties:
    (i) all main effect variables and all interaction variables chosen in step 1 only are given a weight $w = 1$;
    (ii) all interaction variables chosen in step 2 are given a weight $0 < w \le 1$ which is a non-increasing function of the U or S score.
  (c) Create K nested subsets based on the order of entry of the K variables in the weighted Lasso in the previous step.
(4) **Select Subset Using Adjusted Gain in Value Criterion:**
  (a) For each subset $k = 1, \ldots, K$, estimate the maximal Value, e.g.
    (i) use the subset to estimate $\hat{E}$;
    (ii) estimate the optimal policy, $\hat{\pi}_k^*(x) = \arg\max_a \hat{E}[R|X = x, A = a]$;

(iii) estimate the Value of $\hat{\pi}_k^*$ by:

$$\hat{V}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{E}[R|X = x_i, A = \hat{\pi}_k^*(x_i)].$$

(b) Select the subset, k^*, that has the highest Adjusted Gain in Value (AGV) criterion:

$$AGV_k = \frac{\hat{V}_k - \hat{V}_0}{\hat{V}_{m^*} - \hat{V}_0}\left(\frac{m^*}{k}\right),$$

where $m^* = \arg\max_k \hat{V}_k$ and $\hat{V}_0$ is the estimated Value of the policy $\hat{\pi}_0^* = \arg\max_a \hat{E}\,[R|A = a]$.

In step 1 we use Lasso to find the variables among $(X, A*X)$ that are important predictors of R. We chose Bayesian Information Criterion to select the penalty parameter [21] because of its conservative nature, to ensure only strong predictors enter the model. Predictive variables are important for reducing variability in the estimations. However, predictive variables are only part of the puzzle, so we add to step 1 a few more steps to help our algorithm select both prescriptive and predictive variables. In step 2 we look for qualitative interactions individually using an approach which rates each variable in X based on its potential for a qualitative interaction with the action. We look at each of the interaction variables individually to avoid problems with collinearity. In steps 3 and 4 we seek to further refine the set of variables collected in steps 1 and 2. In step 3 we seek a quick way to navigate through the space of all possible combinations of the variables collected in steps 1 and 2. Thus we chose to create nested subsets from the variables based on order of selection in a weighted Lasso. This ordering by the weighted Lasso gives us a joint ranking of all the variables selected in steps 1 and 2. We use the weighting scheme in the weighted Lasso to balance the importance of both predictive and prescriptive variables in the decision making process. Since Lasso favors variables that are predictive we offset this by down-weighting the prescriptive variables. In step 4 we select between the different subsets using the AGV criterion, a criterion that trades off between the complexity and the observed Value of each of the models.

The AGV criterion selects the subset of variables with the maximum proportion of increase in Value per variable. It is similar in idea to the adjusted R^2 value. The model with $m^* = \arg\max_k \hat{V}_k$ variables is akin to a saturated model, because the addition of more variables does not improve the Value of the model. Thus the denominator is the observed maximum gain in value, among the different variable subsets, divided by m^*, an estimate of the degrees of freedom used to achieve that gain in Value. The numerator then measures the gain in Value of the intermediate model, the model with k variables, divided by k, the estimated degrees of freedom needed to achieve that gain in Value.

An alternate way to look at the AGV criterion is that the quotient $(\hat{V}_k - \hat{V}_0)/(\hat{V}_{m^*} - \hat{V}_0)$ compares the gain in Value for the current subset of variables against the maximum gain in Value over all the subsets of variables. Ideally this term stays fairly stationary whenever a main effect variable is added to the model and increases when a qualitative interaction is added to the model. Thus this quotient is expected to be approximately monotone increasing with k. The quotient, m^*/k, acts as a penalty on the inclusion of variables that do not substantially increase the Value. We include main effect variables in the counts m^* and k because each main effect variable that is included decreases the degrees of freedom. Also, the inclusion of main effects in the counts quickly deflates the quotient as k increases, leading to a less severe penalty on larger models. This is helpful since there is often many more useful predictive variables than prescriptive variables.

In the next section we test this algorithm on simulated data. We reference the algorithm as Method U or Method S depending on the scoring function U or S that was used in step 2. For the weighting scheme in step 3(b) we tried multiple different schemes (inverse, exponential, etc.). In practice, the weighting scheme that worked best is listed below:

(1) all predictive variables are given a weight $w = 1$;
(2) all prescriptive variables are given a weight $w = 1 - \frac{U}{\max(U)+\epsilon}$ or $w = 1 - \frac{S}{\max(S)+\epsilon}$, respectively.

The ϵ term in the weight is needed to ensure $w \neq 0$. So ϵ can be thought of as a stabilizing factor, but it can also be thought of as the balancing factor between prescriptive and predictive variables (i.e. large ϵ favors predictive variables, small ϵ favors prescriptive variables). In experimentation we found $\epsilon = H/n$ to be a good value.

5. Simulations

To test the performance of the new techniques, we ran them on realistically designed simulation data and compared the results to using Lasso [3]. Lasso was used to select from the set of main effects of X, and the interactions between A and each variable in X. The main effect of A was not subject to selection, that is the coefficient of A was unconstrained by the L_1 penalty function. We tested two different methods for choosing the penalty parameter. The first method we used was the Bayesian Information Criterion (BIC) as defined in Zou and [21]. We reference this method as BIC Lasso. Zou et al. [21] recommend this method when using Lasso primarily for variable selection. The second method we used for choosing the penalty parameter was 5-fold cross-validation on the prediction error of the Lasso model [3]. This is a standard method for choosing the penalty parameter and we reference this method as CV Lasso. Note that our method uses Lasso as well, but only to select predictive variables in step 1 and to order variables in step 3(a).

To generate realistic simulation data, we randomly selected rows, with replacement from X, the observation matrix from the Nefazodone CBASP trial data. We generated new actions, A, and new responses, R, that covered a wide variety of models. We report results for the following generative models:

(1) Main effects of X only, no treatment effect and no interactions with treatment;
(2) Main effects of X, moderate treatment effect and no interactions with treatment;
(3) Main effects of X, moderate treatment effect, multiple medium to small non-qualitative interactions with treatment, no qualitative interaction with treatment;
(4) Main effects of X, small treatment effect, small qualitative interaction with a binary variable, no non-qualitative interactions;
(5) Main effects of X, small treatment effect, small qualitative interaction with a continuous variable, no non-qualitative interactions;
(6) Main effects of X, small treatment effect, multiple moderate to small non-qualitative interactions with treatment, small to moderate qualitative interaction with a binary variable and treatment;
(7) Main effects of X, small treatment effect, multiple small non-qualitative interactions with treatment, small qualitative interaction with a continuous variable and treatment.

For each generative model, we used main effect coefficients for the variables X, estimated in an analysis of the real data set. In generative models 3–7 we randomly selected variables from the Nefazodone CBASP data for each treatment covariate interaction and used these same variables for each repetition. The treatment, qualitative interaction and non-qualitative interaction coefficients were set using a variant of Cohen's D effect size measure [10] shown below:

$$D = \frac{\beta\sqrt{\mathrm{Var}(R)}}{\sqrt{\mathrm{Var}(X_j)}}. \tag{8}$$

We altered this formula by replacing the marginal variance, $\mathrm{Var}(R)$, with the conditional variance of the response $\mathrm{Var}(R|X, A)$. However, we maintained the definitions of 'small' and 'moderate' effect sizes suggested by [10] as $D = 0.2$ and $D = 0.5$ respectively. Thus the effects are slightly smaller than the traditional definition.

For each generative model, we ran CV Lasso, BIC Lasso, Method U and Method S to see which interaction variables were selected by each method. We repeated this 1000 times and recorded the percentage of time each variable was selected for each method and the sign of the coefficient of each interaction selected.

For each repetition, we also calculated the following statistic for each method

$$T = \frac{V_{\hat{\pi}^*} - V_\pi}{V_{\pi^*} - V_\pi},$$

where V_{π^*} is the Value of the true optimal policy, π^*, V_π is the Value of an 'agnostic' policy π which gives equal probability to each action and $V_{\hat{\pi}^*}$ is the Value of the estimated optimal policy given the selected variables. We estimated the policy $\hat{\pi}^*$ by first fitting a linear model of the selected variables on the response using the training set and then optimizing the fitted model with respect to the action.

Table 1
Simulation results: BL stands for BIC Lasso, *U* for method U and *S* for method S. The first two columns summarize the difference in percentage statistics *T* between BIC Lasso and the two new methods; values denoted with a * are significantly different from zero using a two-sided *t*-test with $\alpha = .05$. Note: model 1 has no treatment effect or interactions with treatment, thus all policies return the same Value. The next three columns give the average number of spurious interactions selected by the three methods over the 1000 repetitions. The last three columns give the selection percentage of the qualitative interaction (when one existed) for each method.

Generative model	Ave		Ave # of spur. interact.			Selection percentage		
	$T_U - T_{BL}$	$T_S - T_{BL}$	*BL*	*U*	*S*	*BL*	*U*	*S*
1	NA	NA	0.04	1.9	1.3	–	NA	–
2	−0.027*	−0.025*	0.03	0.6	0.5	–	NA	–
3	0.000	0.000	0.4	0.6	0.5	–	NA	–
4	0.212*	0.322*	0.1	1.7	1.0	6	24	27
5	0.280*	0.226*	0.1	1.2	1.2	6	35	27
6	0.219*	0.387*	0.1	1.0	0.3	25	53	74
7	0.128*	0.103*	0.1	0.9	0.8	12	60	49

Table 2
Simulation results: CL stands for Cross-validated Lasso, *U* for method U and *S* for method S. The first two columns summarize the difference in percentage statistics *T* between CV Lasso and the two new methods; values denoted with a * are significantly different from zero using a two-sided *t*-test with $\alpha = .05$. Note: model 1 has no treatment effect or interactions with treatment, thus all policies return the same Value. The next three columns give the average number of spurious interactions selected by the three methods over the 1000 repetitions. The last three columns give the selection percentage of the qualitative interaction (when one existed) for each method.

Generative model	Ave		Ave # of spur. interact.			Ave selection percentage		
	$T_U - T_{CL}$	$T_S - T_{CL}$	*CL*	*U*	*S*	*CL*	*U*	*S*
1	NA	NA	4.9	1.9	1.3	–	NA	–
2	0.021*	0.022*	4.8	0.6	0.5	–	NA	–
3	0.009*	0.009*	8.4	0.6	0.5	–	NA	–
4	0.031*	0.140*	5.2	1.7	1.0	40	24	27
5	0.113*	0.059*	4.6	1.2	1.2	37	35	27
6	0.095*	0.263*	5.0	1.0	0.3	69	53	74
7	0.097*	0.072*	5.6	0.9	0.8	57	60	49

The statistic *T* gives the percentage of gain in Value when using the estimated optimal policy as opposed to an 'agnostic' policy relative to the percentage of gain in Value when using the true optimal policy as opposed to an agnostic policy. We compared the new methods with both of the Lasso competitors by looking at the difference in their *T* statistics. The results are listed in Tables 1 and 2. Differences denoted with a * are significantly different from zero using a two sided *t*-test with $\alpha = .05$. Note that since generative model 1 has no treatment effect and no interactions with treatment, all policies will have the same Value resulting in an undefined *T* statistic. The tables also list the average number of spurious interactions selected by each method and the selection percentage of the qualitative interaction (if one existed) over the 1000 repetitions.

Looking over Table 1 we see that BIC Lasso tends to include a slightly smaller number of spurious interactions, as expected, due to its conservative nature [21]. Its conservative nature is also a bonus in terms of the average Value in the rare situation when no interactions exist in the generative model (generative model 2). However, the use of this method results in a dramatic loss in the average Value when a qualitative interaction does exist because the qualitative interaction is often left out.

Table 2 shows that while CV Lasso is good at selecting the qualitative interaction, it tends to include several more spurious interactions than the new methods. This leads to a significant loss in the average Value due to policies with bad decisions based on the spurious variables selected when using this method.

Overall, we found that the two new methods perform well. While the competing Lasso methods each have their appeal in terms of selection, when considering the average Value returns and the purpose of the variable selection, the new methods appear more advantageous.

6. Application: Nefazodone CBASP trial

To apply this method to a real data set we suggest augmenting this algorithm in two ways. First we use bootstrap sampling [22] of the original data to give a measure of reliability on the results. That is, take 1000 bootstrap samples of the data, run the algorithm and record the interaction variables that are selected along with the sign of the interaction coefficient for each bootstrap sample. This will give a percentage of time each interaction variable is selected by the method. Define the adjusted selection percentage to be the absolute value of the number of times an interaction is selected with a positive coefficient minus the number of times an interaction is selected with a negative coefficient. This adjustment eliminates variables that, across the bootstrap samples, do not consistently interact in one direction with the action.

Second, we construct a threshold to determine which interaction variables to include in the final model. The threshold estimates the selection percentages we would expect to see if the data contained no interactions. To compute the threshold, we first remove the interaction effects within the data by randomly reassigning the observed values for the interaction variables to different subjects. In other words, permute the X values of the $X*A$ interactions in the $(X, A, X*A)$ model matrix. After obtaining 100 permuted data sets, on each permuted data set we run the same analysis of taking 1000 bootstrap samples, running the algorithm and recording the selection percentage of each interaction variable over the 1000 bootstrap samples. We then calculate the maximum adjusted selection percentage over the p interaction variables for each permuted data set. The threshold is then set to be the $(1-\alpha)$th percentile over the 100 maximum selection percentages. We found in simulations that the threshold effectively controlled the family-wise error rate to be approximately α giving us in any given experiment $(1-\alpha)$% confidence that a variable with a selection percentage above this threshold interacts with the action.

This augmentation by bootstrap resampling and thresholding helps to stabilize the results and it is possible to apply it to other variable selection algorithms, not just the new methods suggested in this paper. In simulations we found the bootstrap resampling and thresholding also effectively controlled the family-wise error for BIC Lasso. The bootstrap resampling can also be done with CV Lasso. However, this threshold was far too conservative to control the family-wise error rate for the CV Lasso.

To demonstrate these new methods along with this augmentation, we applied them to a real data set dealing with depression. As introduced previously, the Nefazodone CBASP trial [1] was conducted to compare the efficacy of three alternate treatments for patients with chronic depression. We applied the methods to pinpoint if any of the patient characteristics might help to determine the optimal depression treatment for each patient.

The study randomized 681 patients with non-psychotic chronic major depressive disorder (MDD) to either Nefazodone, cognitive behavioral-analysis system of psychotherapy (CBASP) or the combination of the two treatments. For detailed study design and primary analysis see [1]. We considered $p = 61$ baseline covariates for our observation matrix X; these variables are listed in Table 3. The outcome, R, was the 24-item Hamilton Rating Scale for Depression score [23], observed post treatment. For simplicity, we only allowed the action to vary between two treatments at a time. Since the primary analysis of the data showed the combination treatment to be superior to either individual treatment alone, we ran the variable selection techniques twice: the first time with the action varying between the combination treatment and Nefazodone alone, and the second time with the action varying between the combination treatment and CBASP alone.

The results of our first analysis, comparing the combination treatment to Nefazodone alone, are shown in Table 3 and Fig. 3. The adjusted selection percentages for each variable are listed in Table 3 along with 80% and 90% thresholds at the bottom (i.e. alpha equal to 0.2 and 0.1). Fig. 3 shows plots of these adjusted selection percentages and thresholds. The x-axis in each plot corresponds to the variable numbers listed in Table 3. The horizontal dashed lines are 80% thresholds and the horizontal solid lines are 90% thresholds.

Only the adjusted selection percentages from the bootstrap resampling of CV Lasso are plotted in the first plot. Absent a working threshold, it is not very clear which variables should be selected for further analysis. The next three plots are for BIC Lasso, method U and method S. All three of these methods had one variable with an adjusted selection percentage exceeding the 80% threshold. For

Table 3
Results from variable selection techniques on the Nefazodone CBASP trial data comparing the combination treatment against Nefazodone alone.

Variable	Adjusted selection percentages			
	CV Lasso	BIC Lasso	Method U	Method S
1 Gender	43.1	8.3	4.0	3.4
2 Racial category	6.2	1.5	0.4	0.9
3–4 Marital status	12.3, 12.4	0.3, 1.6	0.1, 0.3	0, 0.6
5 Body mass index	2.2	1.2	1.0	0.8
6 Age in years at screening	20.8	2.6	1.8	1.1
7 Family/friend support system	2.3	0.7	0.3	0.1
8 Treated current depression	5.4	0.6	0.1	0.2
9 Psychotherapy current depression	28.5	2.9	1.3	1.2
10 Medication current depression	31.8	4.1	0.4	0.8
11 Treated past depression	35.1	18.7	3.5	3.8
12 Psychotherapy past depression	53.7	33.1	5.2	6.0
13 Medication past depression	21.7	13.5	1.8	2.0
14 Age of MDD Onset	12.8	3.5	2.2	2.0
15–17 Depressive episodes count	23.6, 37.7, 14.6	4.3, 2.4, 2.7	0.4, 1.6, 1.5	1.0, 0.7, 0.9
18 Length current episode	38.0	3.6	1.5	0.4
19–20 MDD current episode type	29.3, 35.5	2.5, 3.7	5.4, 5.5	5.2, 6.3
21–22 MDD current severity	20.5, 9.3	1.1, 0.6	2.9, 1.8	3.6, 1.3
23 Dysthymia onset	27.6	1.3	0.1	0.1
24 Length current dysthymia	15.7	1.9	0.2	0.1
25–26 Generalized anxiety	34.8, 13.8	10.6, 0.9	2.8, 1.5	4.2, 2.7
27 Anxiety disorder NOS	49.3	18.9	1.1	0.5
28–29 Panic disorder	35.2, 38.8	5.4, 18.3	1.3, 3.7	0.4, 3.5
30–31 Social phobia	5.1, 41.3	1.1, 7.0	1.4, 2.3	0.4, 1.8
32–33 Specific phobia	6.0, 36.1	0.1, 10.6	0.8, 7.3	0.2, 11.6
34 Obsessive compulsive	59.8	**47.3**	12.6	12.6
35 Body dysmorphic current	23.9	2.1	2.2	3.6
36 Anorexia or bulimia nervosa	12.9	0.0	0.7	0.3
37–38 Alcohol abuse/dependence	48.3, 62.0	19.0, 35.4	25.4, **45.7**	24.1, **44.9**
39 Drug abuse	1.9	3.1	1.1	1.1
40–41 Post traumatic stress	2.7, 16.4	2.1, 2.9	0.7, 0	0.1, 0.1
42 Other psychological problems	21.9	5.7	2.6	2.7
43 Global assessment of function	9.5	2.6	2.0	1.2
44–45 Main study diagnosis	3.8, 36.9	0.3, 6.0	0.7, 2.3	0.2, 2.8
46 Severity of illness	9.8	0.7	0.1	0.2
47 Total HAMA score	22.9	8.0	9.7	5.4
48 HAMA sleep disturbance	10.8	0.7	0.1	0.2
49 HAMA psychic anxiety score	6.4	0.4	1.0	0.4
50 HAMA somatic anxiety score	57.1	26.9	30.3	23.0
51 Total HAMD-24 score	3.0	0.6	0.3	0.0
52 Total HAMD-17 score	20.3	0.6	0	0
53 HAMD cognitive disturbance	2.0	0	0.7	0.1
54 HAMD retardation score	2.7	0.2	0.1	0.2
55 HAMD anxiety/somatic	2.1	0.3	0.3	0.1
56 IDSSR total score	8.8	0.3	0.2	0
57 IDSSR anxious depression Type	6.7	0.3	0	0
58 IDSSR general/mood cognition	23.4	6.3	4.4	2.7
59 IDSSR anxiety/arousal score	4.3	0.1	0.1	0.1
60–61 IDSSR sleep scores	22.8, 9.1	2.5, 0.7	1.4, 0.5	0.9, 0.2
Thresholds: 80%, 90%	NA	**39.8, 50.3**	**42.4, 45.6**	**41.2, 46.5**

BIC Lasso, this variable was variable 34, *Obsessive Compulsive Disorder*, whereas, for both of the new methods the variable was variable 38, past history of *Alcohol Dependence*. Further analysis of the two variables confirmed that the interaction with *Obsessive Compulsive Disorder* and the action was non-qualitative in the data, whereas the interaction between past *Alcohol Dependence* and the action had good potential for being qualitative. More study should be done to determine the usefulness of past history of *Alcohol Dependence* for selecting treatments. Also, in this study around 20% of subjects in

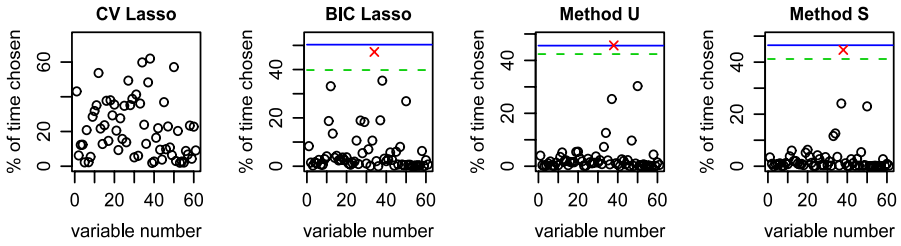


Fig. 3. Plots of interaction variables selected from Nefazodone CBASP trial data comparing the combination treatment to Nefazodone alone. In each plot x-axis is the variable number given in Table 3, and y-axis is adjusted percent of time the variables were selected by the method. Dashed horizontal line is the 80% threshold and solid horizontal line is the 90% threshold. In the second plot the $\times$ identifies the *Obsessive Compulsive Disorder* variable, whereas the $\times$ in the third and fourth plots denotes *Alcohol Dependence*.

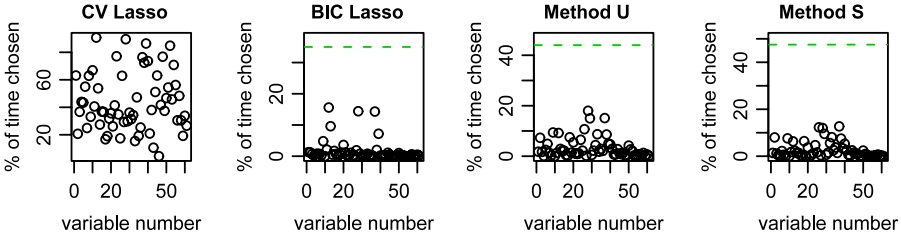


Fig. 4. Plots of interaction variables selected from Nefazodone CBASP trial data comparing the combination treatment to CBASP alone. In each plot x-axis is the variable number given in Table 3, and y-axis is adjusted percent of time the variables were selected by the method. The dashed horizontal lines are 80% thresholds.

each group left the study early. Here R is the last observed Hamilton Rating, which is a worst case scenario under the assumption that depressed subjects who drop out of the study do not improve. Although we included all available good predictors of R in the model of $E[R|X, A]$, it may be the case that unobserved determinants of dropout provide an alternate explanation for the apparent qualitative interaction with past *Alcohol Dependence*.

The results of our second analysis comparing the combination treatment to CBASP alone are shown in Fig. 4. The figure shows plots of the adjusted selection percentages for each method along with 80% thresholds. The x-axis in each plot corresponds to the variable numbers listed in Table 3. The horizontal dashed lines are 80% thresholds. As shown in the plot, no variables were selected by either of the new methods or BIC Lasso. For brevity we forgo listing the individual selection percentages. This analysis suggests there are no true qualitative interactions between the baseline covariates and the two treatment options. Many researchers believe this is the most likely scenario in medical decision-making applications. We conclude that the combination treatment is better than CBASP alone for all patient subsets tested.

7. Discussion

In this article, we discussed when a variable is important in decision making and why variable selection techniques designed for prediction may not perform well in a decision-making setting. We presented two new techniques explicitly designed to select variables for decision making. These techniques focus on interaction variables that are good candidates for playing a role in the actual decision rules.

It should be noted that Lasso treats the indicator variables used to model a categorical variable as separate variables. It is well known that this can lead to over-selection of categorical variables with many categories. Consequently, the proposed method is subject to this problem. Therefore we recommend using Group Lasso [24,25], OSCAR [26] or elastic net type penalty [27] in step 3 of the algorithm when applying it to a data set with many multi-category variables.

The entire algorithm, including bootstrap sampling and thresholding, takes approximately 30 hours to run in Matlab on a 3 Ghz Intel Xeon X5355 processor for a data set of $p = 60$ baseline covariates and $n = 400$ subjects. The algorithm would require far less computation time if a more theoretically justified threshold could be determined, rather than using a permutation based threshold. This is an area for future work.

More research is needed to determine the oracle consistency properties of this algorithm and its performance on problems where $p > n$. Adjusting these methods to deal with dropout is also an open issue. Our long term goal is to extend these methods to settings with multiple decision time points.

Acknowledgments

We wish to thank Martin Keller and the investigators who conducted the trial 'A Comparison of Nefazodone, the Cognitive Behavioral-analysis System of Psychotherapy, and Their Combination for Treatment of Chronic Depression', for use of their data, and gratefully acknowledge Bristol-Myers Squibb for helping fund the study. We also acknowledge financial support from NIDA grants R21 DA019800, K02 DA15674, P50 DA10075, NIMH grant R01 MH080015 and NSF grants DMS 0505432 and DMS 0705532, and technical support from the Betty Jo Hay Chair in Mental Health, A. John Rush, MD at the University of Texas Southwestern Medical Center, Dallas.

References

[1] M.B. Keller, J.P. McCullough, D.N. Klein, B. Arnow, D.L. Dunner, A.J. Gelenberg, J.C. Marekowitz, C.B. Nemeroff, J.M. Russell, M.E. Thase, M.H. Trivedi, J. Zajecka, A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for treatment of chronic depression, New England Journal of Medicine 342 (2000) 331–336.
[2] R.S. Sutton, A.G. Barto, Reinforcement Lerning: An Introduction, MIT Press, Cambridge, MA, 1998.
[3] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B 58 (1996) 267–288.
[4] M. Loth, M. Davy, P. Preux, Sparse temporal difference learning using lasso, in: IEEE International Symposium on Approximate Dynamic Programming and Reinforcement, Springer, Hawaii, USA, 2006.
[5] The ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group, Major outcomes in moderately hypercholesterolemic, hypertensive patients randomized to pravastatin vs usual care: The antihypertensive and lipid-lowering treatment to prevent heart attack trial (allhat-llt), Journal of the American Medical Association 288 (2002) 2998–3007.
[6] J.H. Krystal, J.A. Cramer, W.F. Krol, G.F. Kirket, R. Rosenheck, Naltrexone in the treatment of alcohol dependence, New England Journal of Medicine 345 (2001) 1734–1739.
[7] C.F. Reynolds, M.A. Dew, B.G. Pollock, B.H. Mulsant, E. Frank, M.D. Miller, P.R. Houck, S. Mazumdar, M.A. Butters, J.A. Stack, M.A. Schlernitzauer, E.M. Whyte, A. Gildengers, J. Karp, E. Lenze, K. Szanto, S. Bensasi, D.J. Kupfer, Maintenance treatment of major depression in old age, New England Journal of Medicine 345 (2006) 1130–1138.
[8] S.D. Hollon, A.T. Beck, Cognitive and cognitive behavioral therapies, in: M.J. Lambert (Ed.), Garfield and Bergin's Handbook of Psychotherapy and Behavior Change: An Empirical Analysis, 5th edn., Wiley, New York, 2004, pp. 447–492.
[9] R. Peto, Statistical aspects of cancer trials, in: K.E. Halnan (Ed.), Treatment of Cancer, Chapman, London, UK, 1982, pp. 867–871.
[10] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd edn., Lawrence Earlbaum Associates, Hillsdale, NJ, 1988.
[11] S. Yusuf, J. Wittes, J. Probstfield, H.A. Tyrole, Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials, Journal of the American Medial Association 266 (1991) 93–98.
[12] D.P. Byar, D.K. Corle, Selecting optimal treatment in clinical trials using covariate information, Journal of Chronic Diseases 30 (1977) 445–459.
[13] M. Gail, R. Simon, Testing for qualitative interactions between treatment effects and patient subsets, Biometrics 41 (1985) 361–372.
[14] S. Lagakos, The challenge of subgroup analyses-reporting without distorting, New England Journal of Medicine 354 (2006) 1667–1669.
[15] J. Shuster, J. Van Eys, Interaction between prognostic factors and treatment, Controlled Clinical Trials 4 (1983) 209–214.
[16] S.J. Senn, Individual therapy: New dawn or false dawn, Drug Information Journal 35 (2001) 1479–1494.
[17] Beta-Blocker Pooling Project Research Group, The beta-blocker pooling project (bbpp): Subgroup findings from randomized trials in post infarction patients, The Eropean Heart Journal 9 (1988) 8–16.
[18] S.F. Assmann, S.J. Pocock, L.E. Enos, L.E. Kasten, Subgroup analysis and other (mis)uses of baseline data in clinical trials, The Lancet 355 (2000) 1064–1069.
[19] G. Parmigiani, Modeling in Medical Decision Making: A Baysian Approach, Wiley, West sussex, England, 2002.
[20] C.F. Wu, M. Hamada, Experiments: Planning, Analysis, and Parameter Design Optimization, Wiley, New York, 2000.
[21] H. Zou, T. Hastie, R. Tibshirani, On the "degrees of freedom" of the lasso, The Annals of Statistics 35 (2007) 2173–2192.
[22] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, New York, 1993.

[23] M. Hamilton, Development of a rating scale for primary depressive illness, British Journal of Social and Clinical Psychology 6 (1967) 278–296.
[24] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society, Series B 68 (2006) 49–67.
[25] P. Zhao, G. Rocha, B. Yu, Grouped and hierarchical model selection through composite absolute penalties. Technical Report 703, Department of Statistics University of California at Berkeley, 2006.
[26] H.D. Bondell, B.J. Reich, Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR, Biometrics 64 (2008) 115–123.
[27] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society, Series B 67 (2005) 301–320.