HHS Public Access

Author manuscript

Stat. Author manuscript; available in PMC 2016 March 06.

Published in final edited form as:

Stat. 2015 ; 4(1): 59–68. doi:10.1002/sta4.78.

On Sparse representation for Optimal Individualized Treatment Selection with Penalized Outcome Weighted Learning

Rui Song[a,*], **Michael Kosorok**[b], **Donglin Zeng**[b], **Yingqi Zhao**[c], **Eric Laber**[a], and **Ming Yuan**[c]

[a]Department of Statistics, North Carolina State University, Raleigh, NC 27695

[b]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599

[c]Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin, 53792

Abstract

As a new strategy for treatment which takes individual heterogeneity into consideration, personalized medicine is of growing interest. Discovering individualized treatment rules (ITRs) for patients who have heterogeneous responses to treatment is one of the important areas in developing personalized medicine. As more and more information per individual is being collected in clinical studies and not all of the information is relevant for treatment discovery, variable selection becomes increasingly important in discovering individualized treatment rules. In this article, we develop a variable selection method based on penalized outcome weighted learning through which an optimal treatment rule is considered as a classification problem where each subject is weighted proportional to his or her clinical outcome. We show that the resulting estimator of the treatment rule is consistent and establish variable selection consistency and the asymptotic distribution of the estimators. The performance of the proposed approach is demonstrated via simulation studies and an analysis of chronic depression data.

Keywords

Penalization; Personalized medicine; Support vector machine

1. Introduction

It is well known that patients possess strong heterogeneity in response to treatments in modern clinical studies, especially in chronic diseases. Consequently, a drug may have significant treatment effect for some patients with certain characteristics, but not work for other patients. Treating each individual based on his or her genomic or prognostic information rather than via “one size fits all” approach can thus potentially improve drastically the effect of treatment for public health. This significantly motivates interest in discovering personalized treatment strategies in statistical research, where an optimal personalized treatment strategy is a set of treatment rules tailored to individuals, to

*rsong@ncsu.edu.

maximize long-term clinical outcomes for individual patients, see Murphy (2003); Robins (2004); Moodie et al. (2007).

Various methods have been proposed in the statistical literature to estimate the optimal personalized treatment strategy. Q-learning (Watkins, 1989; Watkins & Dayan, 1992; Murphy, 2005; Zhao et al., 2011; Chakraborty et al., 2010; Song et al., 2014) and A-learning (Murphy, 2003; Robins, 2004) are two backward induction methods for deriving optimal multistage treatment regimes. Other related approaches include likelihood-based methods (Thall et al., 2000, 2002, 2007) and semiparametric methods (Lunceford et al., 2002; Wahed & Tsiatis, 2004, 2006; Moodie et al., 2009; Zhang et al., 2012). Zhao et al. (2012) proposed outcome weighted learning, which can be viewed as a weighted classification problem using the covariate information weighted by the individual response to achieve the goal of maximizing the overall outcome. Since more and more individual-level information are collected in clinical studies, especially in multi-stage clinical trials, these approaches are often challenged by the curse of dimensionality in the presence of high-dimensional covariates. Variable selection hence becomes important in discovering individualized treatment rules.

In spite of a large effort being devoted to variable selection methods, the topic of variable selection for personalized and dynamic treatment regimens has received little attention. Some variable selection techniques used in this area include the Lasso (Loth et al., 2007), decision trees (Ernst et al., 2005) and Bayesian variable selection (Chen et al., 2009). Biernot & Moodie (2010) conducted numerical studies to compare several variable selection methods. In some clinical trials where medical decision making is needed, qualitative interaction tests (Gail & Simon, 1985; Piantadosi & Gail, 1993; Yan, 2004) have been used to test a small number of expert determined pre-specified interactions, such as Allhat et al. (2002); Reynolds et al. (2006). Many of the tests were designed for testing only qualitative interactions between categorical variables and the treatment. Moreover, when the number of covariates is large, these tests are too conservative when controlling the error rate due to multiple testing.

When the dimension of the covariate space is high and there are many irrelevant variables for decision making, imposing sparsity in the parameters via an automatic variable selection procedure can significantly enhance the decision making accuracy. Penalized methods have also been studied to identify variables important for making treatment decisions. Qian & Murphy (2011) developed a two-stage procedure in the framework of Q-learning, where they used L_1 penalized least squares to estimate optimal treatment regimes. Lu et al. (2013) proposed a penalized quadratic loss in the framework of A-learning. Gunter et al. (2011) proposed variable selection methods for qualitative interactions, where two variable-ranking quantities were presented. Existing penalization methods are primarily designed for linear regression models, which may not be appropriate for variable selection in treatment decision making when the true model is not correctly specified. We are therefore motivated to develop variable selection methods that do not depend on the parametric modeling of the value function.

In this paper, we propose to adopt the outcome weighted learning framework to simultaneously estimate the optimal decision rule and incorporate sparsity in a way that retains the computational advantages and theoretical validity. Our proposed method is called penalized outcome weighted learning (POWL). We demonstrate later in the paper that POWL can achieve comparable or better classification accuracy while, at the same time, choosing relevant features. The main idea is to introduce the SCAD penalty proposed in Fan & Li (2001) to outcome weighted learning. The resulting optimization problem is a linear program and can be efficiently solved. Our approach can handle very high-dimensional state spaces. Furthermore, we study Fisher consistency of the linear POWL. We also investigate asymptotic properties of the variable coefficients in the SVM solution for linear classification. We focus on the case where the decision function is a linear function of the covariates. Various variable selection methods were studied in recent years to encourage sparsity in support vector machines (Cortes & Vapnik, 1995). References include Zhu et al. (2003); Wu et al. (2008); Park et al. (2012) etc. Our work can be viewed as a penalized weighted SVM.

The rest of the paper is organized as follows. The proposed POWL is introduced in Section 2. In Section 3, we study the theoretical properties of POWL and numerical studies are presented in Section 4.We apply the proposed method to a chronic depression data set in Section 5, and conclude with a brief discussion in Section 6.

2. Method

2.1. Outcome-Weighted Learning (OWL)

Consider data from a randomized trial. By notation, we use A to denote treatment assignment, which takes values −1 and 1. The probability of treatment assignment A is given as $P(A = 1) = \pi$ and $P(A = -1) = 1 - \pi$, where $0 < \pi < 1$. We use $\mathbf{X} = (X_1,\ldots, X_p)'$ to denote p-dimensional biomarker and prognostic information associated with the patient and Let $\tilde{\mathbf{X}} = (1, \mathbf{X}')'$. Let R be the clinical outcome of interest (assuming large values are desirable) which we also call the "reward." Let $f(\mathbf{x}, r)$ and $g(\mathbf{x}, r)$ denote the conditional density of $(\mathbf{X}, R)$ given $A = 1$ and −1 respectively. In this randomized trial setting, A is randomly assigned to patients and the observed data from n i.i.d patients has the form $(A_i, \mathbf{X}_i, R_i)$, $i = 1,\ldots, n$. Our goal is to identify a deterministic decision rule, $d(\mathbf{x})$, which takes a given value $\mathbf{x}$ of $\mathbf{X}$ and returns a treatment choice from a space $\mathscr{A}$. We denote the distribution of (X, A, R) by P and expectation with respect to this distribution by E. Let P^d denote the distribution of $(\mathbf{X},A, R)$ when $A = d(\mathbf{X})$, the treatment is determined by the rule d. The expectation with respect to this distribution is denoted by E^d, where the individualized treatment rule (ITR) $d(\mathbf{x})$ is used to assign treatments. Define the value function as $V(d) = E^d(R)$. Thus, an optimal individualized treatment rule, d_0, is a rule that has the maximal value, i.e., d_0 is the maximizer of $V(d)$ over decision rules d. Specifically, $V(d_0) = E\{\max_{A\in\mathscr{A}} Q(\mathbf{X}, d_0(\mathbf{X}))\}$ where $Q(\mathbf{X}, A) \equiv E[R|A = d(\mathbf{x}), \mathbf{X} = \mathbf{x}]$. The expected reward under ITR $\mathscr{D}$ is given as

$$E^{\mathscr{D}}(R)=\int RdP^{\mathscr{D}}=\int R\frac{dP^{\mathscr{D}}}{dP}dP=E\left[\frac{I(A=\mathscr{D}(\mathbf{X}))}{A\pi+(1-A)/2}R\right]. \quad (1)$$

It can be seen from (1) that maximizing $V(\mathscr{D})$ is equivalent to a weighted classification problem, where we classify subjects with $\mathbf{X}$ according to treatment A but weight each subject by $R/(A_\pi + (1 - A)/2)$. Based on this idea, outcome weighted learning was developed in Zhao et al. (2012). Because direct optimization of (1) is intractable due to the nonconvex and discontinuous 0−1 loss, Zhao et al. (2012) propose to use convex relaxation in combination with techniques applied in support vector machines (Cortes & Vapnik, 1995).

In the next section, we generalize this learning approach to allow simultaneous maximization of the value function and variable selection, which we call penalized outcome-weighted learning (POWL). We consider the situation where the optimal decision rule belongs to a linear function class of the feature covariates.

2.2. Penalized Outcome-weighted Learning (POWL)

Let $p_\lambda(\boldsymbol{\beta})$ denote some penalty function, where $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ is the parameter of interest and λ is the tuning parameter. The POWL minimizes the following objective function

$$Q_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{A_i\pi + (1 - A_i)/2}(1 - A_i h(\mathbf{X}_i, \boldsymbol{\beta}))_+ + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \quad (2)$$

where $h(\cdot)$ is the decision function, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_-')'$ and $\boldsymbol{\beta}_- = (\beta_1, \ldots, \beta_p)'$, $(z)_+ = \max(z, 0)$. Choices of the penalty functions include the lasso (Tibshirani, 1996) and the SCAD (Fan and Li, 2001). The limit of $Q_n(\boldsymbol{\beta})$ is defined as

$$Q(\boldsymbol{\beta}) = E\left[\frac{R}{A\pi + (1 - A)/2}(1 - Ah(\mathbf{X}; \boldsymbol{\beta}))_+\right]. \quad (3)$$

We note that it is the same as the population objective function of the OWL proposed in Zhao et al. (2012).

Suppose that the decision function $h(\mathbf{X}; \boldsymbol{\beta})$ minimizing (3) is a linear function of $\mathbf{X}$ with true parameter value $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, that is, $h(\mathbf{X}, \boldsymbol{\beta}) = \mathbf{X}'\tilde{\boldsymbol{\beta}}^0$ and $\boldsymbol{\beta}^0 = \text{argmin}_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$. Then the corresponding individualized optimal rule will assign a subject with prognostic value $\mathbf{X}$ into treatment 1 if $\mathbf{X}'\tilde{\boldsymbol{\beta}}^0 > 0$ and −1 otherwise. Let $\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta})$. The POWL will classify a subject $(\mathbf{X}_i, A_i, R_i)$ by the sign of $h(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$ correspondingly. Define $\mathscr{M} = \{j : \beta_j^0 \neq 0\}$. Let $s = |\mathscr{M}|$, the cardinality of the non-sparse set $\mathscr{M}$. We aim to estimate $\boldsymbol{\beta}^0$ and recover the set $\mathscr{M}$. To facilitate the theoretical derivation and the numerical calculation, we consider the SCAD penalty proposed in Fan & Li (2001) defined as

$$p'_{\lambda_n}(\theta) = \lambda_n \left\{1(\theta \leq \lambda_n) + \frac{(a\lambda_n - \theta)_+}{\lambda_n(a - 1)} 1(\theta > \lambda_n)\right\},$$

where $a > 2$ and $\lambda_n > 0$ are the tuning parameters.

We propose to use the one-step estimator in Zou & Li (2008) to approximate the penalty term. Minimizing (2) can be rewritten as

$$\max_{\beta,\|\beta_-\|=1}\sum_{i=1}^{n}\frac{R_i}{A_i\pi+(1-A_i)/2}\xi_i+\sum_{i=0}^{p}p'_{\lambda_n}(\beta_i)|\beta_i|,$$

$$\text{subject to}|A_i(\beta_0+\beta'_-\mathbf{X}_i)|\geq(1-\xi_i),\xi_i\geq 0,$$

where ξ_i is a slacking variable for subject *i* to allow a small portion of wrong classification. Therefore we use linear programming to solve the above problem.

3. Theoretical Results

In this section, we study the theoretical properties of the proposed POWL with SCAD penalty. First we consider the Fisher consistency of the linear POWL. It was shown in Zhao et al. (2012) that the decision function based on the surrogate loss $Q(\beta)$ possess Fisher consistency. Since the population version of our objective function remains the same as OWL, our procedure also possess Fisher consistency, that is, the population version of our estimation procedure is the same as the target of the estimation, and the population minimizer of the surrogate objective function yields a classification rule with the same sign as that of the Bayes rule.

We define the score as:

$$S(\boldsymbol{\beta})=-E\left[\frac{R}{A\pi+(1-A)/2}1(1-Ah(\mathbf{X},\boldsymbol{\beta})\geq 0)A\tilde{\mathbf{X}}\right].$$

We also define

$$H(\boldsymbol{\beta})=E\left[\frac{R}{A\pi+(1-A)/2}\delta(1-Ah(\mathbf{X},\boldsymbol{\beta}))A\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\right],\text{ and}$$

$$G(\boldsymbol{\beta})=E\left[\frac{R}{A\pi+(1-A)/2}1(1-Ah(\mathbf{X},\boldsymbol{\beta})\geq 0)A\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\right],$$

where the δ function is the Dirac delta function, that is, on the real number line it is zero everywhere except at zero, with the integral function as the indicator function $1(t\geq 0)$.

Next we establish consistency and convergence rates of the estimator $\hat{\beta}$ under the following conditions. We use lower case to represent the realizations of the random variables. Denote $f(\mathbf{x})$ and $g(\mathbf{x})$ as the conditional densities of $\mathbf{x}$ given $A = 1$ and -1 respectively.

- **A1** The densities $f(\mathbf{x}, r)$ and $g(\mathbf{x}, r)$ are continuous with finite second moments.
- **A2** There exists $B(\mathbf{x}_0, \delta_0)$, a ball centered at $\mathbf{x}_0$ with radius $\delta_0 > 0$ such that $f(\mathbf{x}) > C_1$, $g(\mathbf{x}) > C_1$ for any $\mathbf{x} \in B(\mathbf{x}_0, \delta_0)$.

A3 For some $1 \le j \le p$, $E(X_jR|A = 1) \neq E(X_jR|A = -1)$.

A4 Let $M^+ = \{\mathbf{x} \in \mathcal{X} | \mathbf{x}'\boldsymbol{\beta}^0 = 1\}$ and $M^- = \{\mathbf{x} \in \mathcal{X} | \mathbf{x}'\boldsymbol{\beta}^0 = -1\}$. For an orthogonal transformation φ_j that maps $\boldsymbol{\beta}_+^0 / \|\boldsymbol{\beta}_+^0\|$ to the unit vector e_j whose $j-$th element is one and the other elements are zero, for some $1 \le j \le d$, there exist rectangles

$$D^+ = \{\mathbf{x} \in M^+ : l_i \le (\phi_j \mathbf{x})_i \le v_i, \text{for} i \neq j\}, \text{and}$$

$$D^- = \{\mathbf{x} \in M^- : l_i \le (\phi_j \mathbf{x})_i \le v_i, \text{for} i \neq j\}.$$

Condition A1 and A4 are needed for the Hessian matrix $H(\boldsymbol{\beta})$ to be continuous and positive definite near $\boldsymbol{\beta}^0$. Condition A2 and A3 ensure the weighted classification problem is nonseparable and $\boldsymbol{\beta}^0$ is nonzero.

Theorem 1

Under Conditions A1–A4, if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, then with probability tending to one we have

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0\| = O_p(n^{-1/2}).$$

To discuss the oracle properties of the POWL with SCAD penalty, we need the following notation. Let $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_1^{0\prime}, 0')'$, where $\boldsymbol{\beta}_1^0 \in \mathbb{R}^{S+1}$ represents the true value for $\boldsymbol{\beta}_1$, the nonzero elements of $\boldsymbol{\beta}$. Let

$$H(\boldsymbol{\beta}) = \begin{pmatrix} H_{11}(\boldsymbol{\beta}) & H_{12}(\boldsymbol{\beta}) \\ H_{21}(\boldsymbol{\beta}) & H_{22}(\boldsymbol{\beta}) \end{pmatrix}, \text{and} G(\boldsymbol{\beta}) = \begin{pmatrix} G_{11}(\boldsymbol{\beta}) & G_{12}(\boldsymbol{\beta}) \\ G_{21}(\boldsymbol{\beta}) & G_{22}(\boldsymbol{\beta}) \end{pmatrix},$$

where $H_{11}(\boldsymbol{\beta})$ and $G_{11}(\boldsymbol{\beta})$ are the top $(1 + s)$ square sub-matrix of $H(\boldsymbol{\beta})$ and $G(\boldsymbol{\beta})$, respectively. The sparsity and the oracle property of the estimators are established in the following theorem:

Theorem 2

Denote $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^{'}, \hat{\boldsymbol{\beta}}_2^{'})'$, where $\hat{\boldsymbol{\beta}_1} \in \mathbb{R}^{s+1}$ estimates $\boldsymbol{\beta}_1$ and $\hat{\boldsymbol{\beta}_2} \in \mathbb{R}^{p-s}$ estimates $\boldsymbol{\beta}_2$. Under Conditions A1–A4, if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, then $\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}_2} = 0) = 1$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0) \to_d N(0, H_{11}^{-1}(\boldsymbol{\beta}^0) G_{11}(\boldsymbol{\beta}^0) H_{11}^{-1}(\boldsymbol{\beta}^0))$.

4. Simulation Studies

Zhao et al. (2012) compared the numerical performance of OWL and the L_1 method in Qian & Murphy (2011), where OWL demonstrated better performance in value function estimation. Therefore in this section we focus on comparing the numerical performance of OWL and POWL in variable selection and value function estimation. We generated 50

dimensional vectors of prognostic variables $X_1,\ldots, X_{50}$. They are independently uniformly distributed random variables on [−1, 1]. The treatment A takes values −1 and 1 and is independent of the covariates. The response variable R is normally distributed with mean $Q_0 = 1 + 2X_1 + X_2 + 0.5X_3 + T(X, A)$ and variance 1. The term $T(X, A)$ represents the interaction between treatment and prognostic variables. The first s variables are relevant while the remaining $p - s$ variables are noise variables. We consider the following two scenarios with different degrees of sparsity level:

1. Scenario 1: $T(X, A) = (X_1 + X_2)A$;
2. Scenario 2: $T(X, A) = (X_1 + X_2 + X_3 - X_4 + X_5 - X_6)A$.

Therefore the optimal rule for Scenario 1 is $I(X_1 + X_2 > 0)$ and that for Scenario 2 is $I(X_1 + X_2 + X_3 - X_4 + X_5 - X_6 > 0)$. Based on large scale Monte Carlo simulations, the optimal values for two scenarios are 1.66 and 2.13 respectively. We generated n training samples with n ranging over 30, 100, 200, 400 and 800. For each scenario we estimate the optimal ITR by applying OWL and POWL and repeated the procedure on 200 independent data sets. The BIC criteria is used to select the optimal values of the tuning parameters. The grid for the tuning parameter λ_n is $\{2^{-5}, 2^{-4},\ldots, 2^5\}$.

The performances of the methods are evaluated by the following four criteria. The first is to evaluate the value function using the estimated optimal individual treatment rule when applying to an independent validation data set with sample size 10000. We evaluate the estimated value function for any ITR with the method in Murphy et al. (2001). The second is to evaluate the misclassification rates of the estimated optimal ITR from the true optimal ITR using the validation data. The third is to evaluate the size of the selected model. The fourth is to evaluate the number of true positives in the selected model. Because the size of the selected model and the number of true positives are closely related with the choice of tuning parameter, we also record the minimal model sizes that are needed to cover the true model to gauge the difficulty of the problem. The simulation results are summarized in Tables 1 and 2.

We report the sample mean of both value functions and misclassification rates with the associated sample variance in parenthesis. The sample median of the minimum model size, the size of the selected model and the number of true positives are also recorded with the associated robust standard deviation in parenthesis. As the sample size increases, the performance of OWL and POWL increase in the sense that the value function estimates get closer and closer to the optimal value function and the misclassification rates continue decreasing. POWL provides bigger value function estimates and smaller misclassification rates in all simulation settings. In terms of the variable selection results, POWL can recover the true model better and better as the sample size grows. The number of true positives increases as the sample size increases.

5. Data Analysis

We apply the proposed method to the data from the Nefazodone-CBASP trial (Keller et al., 2000). This trial was conducted to compare the efficacy of several alternative treatments for patients with chronic depression. In the study, 681 patients with nonpsychotic chronic major

depressive disorder (MDD) were randomized to either Nefazodone, cognitive behavioral-analysis system of psychotherapy (CBASP), or to a combination of both treatments. The primary outcome was the score on the 24-item Hamilton Rating Scale for Depression (HRSD). Low HRSD scores are desirable.

We used a subset of the Nefazodone-CBASP data consisting of 647 patients, where 211, 216 and 220 patients were assigned to the combined treatment, Nafazodone and CBASP group, respectively. Among the three treatments, pairwise comparisons showed that the combination treatment significantly lowered the HRSD scores compared to either of the single treatments. There was no overall difference between the two single treatments.

We used the proposed POWL to estimate the optimal individualized treatment rule with 50 pretreatment variables. Pairwise comparisons were conducted among the three treatments. We calculated the value functions from a five-fold cross-validation analysis, where the estimated rules were obtained based on four folds of the data and the estimated value functions were obtained use the remaining fifth fold of the data. The value functions calculated this way should better represent expected value functions for future subjects, as compared to calculating value functions based on the training data. The averages of the cross-validation value functions from the POWL and the OWL approach are presented in Table 3.

As can be seen in Table 3, the POWL approach produced slightly larger value functions, corresponding to smaller HRSD values compared with the OWL approach. When comparing combination treatment with nefazodone only, both OWL and POWL recommended the combination treatment to all patients in the validation data in each round of the cross validation procedure. When the two single treatments are studied, there are only negligible differences in the estimated value functions for the two methods and the selection results also indicate an insignificant difference between them. Meanwhile, POWL needs less variables for decision making. In summery, POWL not only yields individualized treatment rules with the best clinical outcomes, but also enjoys more parsimonious decision rules.

6. Discussion

In this article, we proposed a new variable selection method for optimal treatment decision making, which does not need modeling of the value function. Our method can be applied to data with continuous outcomes and binary treatment options from a clinical trial or an observational study. Simulation studies demonstrate that our method can identify the most important variables under various settings and can provide a good estimated optimal treatment regime which has a small error rate and a high average outcome value.

Dynamic treatment regimes, which are sequential decision rules for individual patients that can adapt over time, are more useful than single-stage decision rules in some applications, see the discussion in Murphy (2003). There are great recent interests on developing statistical inference for dynamic treatment regimes (Robins, 2004; Moodie et al., 2007; Zhao et al., 2011; Song et al., 2014). It will be interesting and useful to generalize the proposal approach for variable selection in developing dynamic treatment regimes.

References

Allhat O, et al. Major outcomes in moderately hypercholesterolemic, hypertensive patients randomized to pravastatin vs usual care: the antihypertensive and lipid-lowering treatment to prevent heart attack trial (allhatllt). JAMA: the Journal of the American Medical Association. 2002; 288(23): 2998.

Biernot P, Moodie EE. A comparison of variable selection approaches for dynamic treatment regimes. The international journal of biostatistics. 2010; 6(1)

Chakraborty B, Murphy S, Strecher V. Inference for non-regular parameters in optimal dynamic treatment regimes. Statistical methods in medical research. 2010; 19(3):317–343. [PubMed: 19608604]

Chen W, Ghosh D, Raghunathan TE, Sargent DJ. Bayesian variable selection with joint modeling of categorical and survival outcomes: an application to individualizing chemotherapy treatment in advanced colorectal cancer. Biometrics. 2009; 65(4):1030–1040. [PubMed: 19210736]

Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995:273–297.

Ernst D, Geurts P, Wehenkel L. Tree-based batch mode reinforcement learning. Journal of Machine Learning Research. 2005:503–556.

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

Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics. 1985:361–372. [PubMed: 4027319]

Gunter L, Zhu J, Murphy S. Variable selection for qualitative interactions. Statistical methodology. 2011; 8(1):42–55. [PubMed: 21179592]

Keller MB, McCullough JP Jr, Klein DN, Arnow B, Dunner D, Gelenberg A, Markowitz J, Nemeroff C, Russell J, Thase M, Trivedi M, Zajecka J. A comparison of nefazodone, the cognitive behavioral analysis system of psychotherapy, and their combination for the treatment of chronic depression. New England Journal of Medicine. 2000; 342:1462–1470. [PubMed: 10816183]

Loth, M.; Davy, M.; Preux, P. Approximate Dynamic Programming and Reinforcement Learning, 2007. ADPRL 2007. IEEE International Symposium on. IEEE; 2007. Sparse temporal difference learning using lasso; p. 352-359.

Lu W, Zhang HH, Zeng D. Variable selection for optimal treatment decision. Statistical methods in medical research. 2013; 22(5):493–504. [PubMed: 22116341]

Lunceford J, Davidian M, Tsiatis A. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. Biometrics. 2002; 58(1):48–57. [PubMed: 11890326]

Moodie EE, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. Biometrics. 2007; 63(2):447–455. [PubMed: 17688497]

Moodie EEM, Platt RW, Kramer MS. Estimating Response-Maximized Decision Rules With Applications to Breastfeeding. Journal of the American Statistical Association. 2009; 104:155–165.

Murphy SA. Optimal dynamic treatment regimes. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003; 65(2):331–355.

Murphy SA. An experimental design for the development of adaptive treatment strategies. Statistics in medicine. 2005; 24(10):1455–1481. [PubMed: 15586395]

Park C, Kim K, Myung R, Koo J. Oracle properties of scad-penalized support vector machine. Journal of Statistical Planning and Inference. 2012; 142(8):2257–2270.

Piantadosi S, Gail M. A comparison of the power of two tests for qualitative interactions. Statistics in Medicine. 1993; 12(13):1239–1248. [PubMed: 8210823]

Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Annals of statistics. 2011; 39(2):1180. [PubMed: 21666835]

Reynolds CF, Dew MA, Pollock BG, Mulsant BH, Frank E, Miller MD, Houck PR, Mazumdar S, Butters MA, Stack JA, et al. Maintenance treatment of major depression in old age. New England Journal of Medicine. 2006; 354(11):1130–1138. [PubMed: 16540613]

Robins, JM. Proceedings of the second seattle Symposium in Biostatistics. Springer; 2004. Optimal structural nested models for optimal sequential decisions; p. 189-326.

Song R, Wang W, Zeng D, Kosorok MR. Penalized q-learning for dynamic treatment regimes. Statistica Sinica (To appear). 2014

Thall P, Millikan R, Sung H. Evaluating multiple treatment courses in clinical trials. Statistics In Medicine. 2000; 19(8):1011–1028. [PubMed: 10790677]

Thall P, Sung H, Estey E. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. Journal of the American Statistical Association. 2002; 97(457):29–39.

Thall PF, Wooten LH, Logothetis CJ, Millikan RE, Tannir NM. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. Statistics In Medicine. 2007; 26(26):4687–4702. [PubMed: 17427204]

Wahed A, Tsiatis A. Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. Biometrika. 2006; 93(1):163–177.

Wahed AS, Tsiatis AA. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomised designs in clinical trials. Biometrics. 2004; 60(5):124–133. [PubMed: 15032782]

Watkins, CJ. Ph.D. thesis. England: University of Cambridge; 1989. Learning from delayed rewards.

Watkins CJ, Dayan P. Q-learning. Machine Learning. 1992; 8(3–4):279–292.

Wu S, Zou H, Yuan M. Structured variable selection in support vector machines. Electronic Journal of Statistics. 2008; 2:103–117.

Yan X. Test for qualitative interaction in equivalence trials when the number of centres is large. Statistics in Medicine. 2004; 23(5):711–722. [PubMed: 14981671]

Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012; 68(4):1010–1018. [PubMed: 22550953]

Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association. 2012; 107(499):1106–1118. [PubMed: 23630406]

Zhao Y, Zeng D, Socinski MA, Kosorok MR. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. Biometrics. 2011; 67(4):1422–1433. [PubMed: 21385164]

Zhu, J.; Rosset, S.; Hastie, T.; Tibshirani, R. Neural Information Processing Systems. MIT Press; 2003. 1-norm support vector machines; p. 16

Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. The Annals of Statistics. 2008; 36:1509–1533.

Table 1

Simulation results based on 200 independent datasets for Scenario 1. The sample mean of both value functions and misclassification rates (“MR”) with the associated sample variance in parenthesis are presented. The sample median of the minimum model size, the size of the selected model and the number of true positives are also recorded with the associated robust standard deviation in parenthesis. “MR” denotes misclassification rate. “MMMS” denotes minimum model size. “NV” denotes number of selected variables. “TP” denote number of true positives. The optimal value is 1.66.

Sample size	Methods	MR	Value	MMMS	NV	TP
30	OWL	0.3686(0.0312)	0.9008(0.1849)			
	POWL	0.3682(0.0349)	0.9015(0.2011)	27(37.52)	5(5.36)	0(1.34)
100	OWL	0.3352(0.0017)	1.1308 (0.0072)			
	POWL	0.3049(0.0130)	1.3765(0.0460)	20(44.22)	6(8.04)	1(1.34)
200	OWL	0.2912(0.0225)	1.0113(0.2650)			
	POWL	0.1722(0.0135)	1.2153(0.3874)	7(6.70)	3(2.68)	1(1.34)
400	OWL	0.2580(0.0074)	1.2875(0.1684)			
	POWL	0.1842(0.0065)	1.4152(0.1983)	2(0)	2(1.34)	2(1.34)
800	OWL	0.2045(0.0004)	1.5209(0.0011)			
	POWL	0.0649(0.0013)	1.6474(0.0008)	2(0)	4(2.68)	2(0)

Table 2

Simulation results based on 200 independent datasets for Scenario 2. The sample mean of both value functions and misclassification rates ("MR") with the associated sample variance in parenthesis are presented. The sample median of the minimum model size, the size of the selected model and the number of true positives are also recorded with the associated robust standard deviation in parenthesis. The notations are the same as these in Table 1. The optimal value is 2.13.

Sample size	Methods	MR	Value	MMMS	NV	TP
30	OWL	0.4050(0.0018)	1.3286(0.0218)			
	POWL	0.4280(0.0036)	1.2517(0.0448)	42(14.74)	6(2.68)	0(1.34)
100	OWL	0.3579(0.0020)	1.4801(0.0208)			
	POWL	0.2959(0.0035)	1.6761(0.0304)	34(20.10)	10(5.36)	4(2.68)
200	OWL	0.2749(0.0010)	1.7357(0.0092)			
	POWL	0.2548(0.0025)	1.7824(0.0121)	29(40.20)	8(4.02)	5(1.34)
400	OWL	0.2912(0.0225)	1.0113(0.2650)			
	POWL	0.1722(0.0135)	1.2153(0.3874)	10(0)	11(5.36)	6(0)
800	OWL	0.1345(0.0002)	2.0333(0.0011)			
	POWL	0.0523(0.0009)	2.1195(0.0014)	7(0)	6(2.68)	6(0)

Table 3

Mean depression scores from cross-validation procedure with OWL and POWL. "NVs" denotes the number of variables selected via POWL.

	OWL	POWL	
	HRSD	HRSD	NVs
Nefazodone vs CBASP	13.86	13.67	29
Combination vs Nefazodone	13.53	13.11	33
Combination vs CBASP	15.63	15.22	35