Statistics and Its Interface Volume 12 (2019) 355–363

Optimal treatment assignment of multiple treatments with analysis of variance decomposition

Zhilan Lou*, Jun Shao, and Menggang Yu

Personalized medicine to identify individualized treatment assignment rules has received increasing interest. When there are more than two treatments, the outcome weighted learning framework builds an optimal assignment rule via the skill of reproducing kernel Hilbert space. One main challenge is that the interpretation of covariates is difficult since the solution is a black-box classifier. Consequently, we establish a structured optimal treatment assignment rule with the functional analysis of variance decomposition. The method promotes the sparsity of the final solution by using structured kernel function and an l_1 penalty term. Meanwhile, we propose an easy-handling iterative procedure to overcome the calculation problem. Convergence of the risk function for resulting estimator is shown in the paper. The finite sample performance of the proposed method is demonstrated by simulation studies and a real data analysis.

Keywords and phrases: Personalized medicine, Treatment assignment rule, Analysis of variance decomposition, Structured multi-category support vector machine.

1. INTRODUCTION

Personalized medicine means providing the right patient with the right drug at the right dose at the right time using individual patient characteristics, including patient demographics, genomic information, treatment and outcome history and so on. The significant heterogeneity across patients in response to treatments is the reason why we should consider personalized medicine. A drug that works for a majority of individuals may not work for a subset of patients with certain characteristics. For example, molecularly targeted cancer drugs are only effective for patients with tumors expressing targets [3], and significant heterogeneity exists in responses among patients with different levels of psychiatric symptoms [14]. At recent, this topic is becoming an increasingly popular research topic among clinical and intervention scientists [15, 8] who try to find an individualized treatment assignment rule to optimize patient responses.

The classical approach to search for an optimal treatment assignment rule involves assuming a parametric or semiparametric model. However, the model assumptions may not be valid due to the complex disease mechanism and individual heterogeneity [6, 20, 12, 18], while these methods emphasize prediction accuracy of response model rather than directly constructing an optimal treatment assignment rule. An alternative approach for the binary treatments tries to construct an treatment assignment rule by maximizing the expected clinical outcome within a weighted classification framework [25, 24, 23, 5], called outcome weighted learning. For the case of multiple treatments, [13] proposed an extended outcome weighted learning framework to construct the treatment assignment rule under equal or unequal loss, together with the Fisher consistency and some asymptotic properties. This approach spares modeling of the covariate main effects and covariate-treatment interactions.

The motivation for this article is the interpretability of covariates. The flexibility of support vector machines [21, 16] in outcome weighted learning is to transform the covariates into the high-dimensional feature space, and the hyperplane in the high-dimensional space can distinguish the different classes well. Recent applications often involve a large number of covariates. However, this kind of transformation is hard to clearly indicated in the applications. The solution of the treatment assignment rule is usually expressed as a linear combination of representers which resulting in a black-box classifier, while identifying important predictors is often crucial in practical applications.

In this article, we propose a method to estimate the optimal treatment rule with clear interpretability of covariates for the case of multiple treatments. In Section 2, we first turn the optimal treatment assignment rule that maximizing the expected clinical outcome into minimizing a risk related with a convex vector hinge loss weighted by clinical outcomes. Furthermore, motivated by the COSSO method [11] that produces sparse solutions, we use structured kernel function [9] and an additional l_1 penalty term to enhance the interpretability of covariates and promote the sparsity of the final solution. Similar to the LASSO method proposed by [19], the l_1 penalty term achieves the effect of variable selection by shrinking the weight of less relevant variables to zero. We then follow the idea in [10] that applies the technique of Reproducing Kernel Hilbert Space (RKHS) to turn the problem into quadratic programming problems for easy computation. To overcome the calculation shortcoming, we propose an easyhandling iterative procedure which guaran-

*Corresponding author.

tees convergence under given tuning parameters. In Section 3, our proposed method is compared with some other recent methods through simulation studies. The proposed method is also applied to a breast cancer behavioral study with four treatment arms in Section 4. We conclude with a discussion of our proposed method and future work in Section 5.

2. METHODOLOGY

2.1 Outcome weighted learning for estimating optimal treatment rule

Assume data are collected from a randomized trial with k different treatments indexed by $A \in \{1, \ldots, k\}$. Let X be a p-dimensional covariate vector associated with a clinical outcome. Let $Y^{(j)}$ be the clinical outcome when treatment $A = j$, $j = 1, ..., k$. Since each patient receives one and only one treatment, the observed clinical outcome is $Y = \sum_{j=1}^{k} I_{\{A=j\}} Y^{(j)}$, where I is the indicator function. We observe $(Y_i, X_i, A_i), i = 1, ..., n$ in a given randomized trial, which are independent and identically distributed as (Y, X, A). From the definition, we can see that the treatment assignment A is related with Y, but A is independent of X and $Y^{(1)}, ..., Y^{(k)}$. And we assume that a large value of the clinical outcome is preferred.

Based on the observed data, our statistical goal is to construct an treatment assignment rule $D(X) \in \{1, \ldots, k\}$. Thus a future patient with collected covariate information X will receive the personalized treatment $D(X)$ that leads to a larger clinical outcome $Y^{(D)}$. Since we assume the larger $Y^{(D)}$ the better, D should be constructed to maximize the expected outcome $E(Y^{(D)})$, where E is the expectation with respect to the distribution of (Y, X, A). Using the independence between A and $(X, Y^{(1)}, ..., Y^{(k)})$, we can find that

$$\begin{aligned}
E(Y^{(D)}) &= E\left\{\sum_{j=1}^{k} I_{\{D(X)=j\}} Y^{(j)}\right\} \\
&= E\left\{\sum_{j=1}^{k} I_{\{D(X)=j\}} Y^{(j)} \middle| A = j\right\} \\
&= \sum_{j=1}^{k} E\left\{\frac{I_{\{D(X)=j\}} Y}{\pi(j)} \middle| A = j\right\} P(A = j) \\
&= E\left[E\left\{\frac{I_{\{D(X)=A\}} Y}{\pi(T)} \middle| A\right\}\right] \\
&= E\left\{\frac{I_{\{T=D(X)\}} Y}{\pi(A)}\right\} \\
&= E\left\{\frac{Y}{\pi(A)}\right\} - E\left\{\frac{I_{\{A\neq D(X)\}} Y}{\pi(A)}\right\} \\
&= E\left\{\frac{Y}{\pi(A)}\right\} - \sum_{j=1}^{k} E\{I_{\{D(X)\neq j\}} E(Y|A = j, X)\},
\end{aligned}$$

where $\pi(A)$ is a function of A with $\pi(j) = P(A = j)$. The optimal treatment assignment rule $D^*(X)$ is defined as the rule that maximizes $E(Y^{(D)})$, i.e.,

$$\begin{aligned}
D^* &= \arg\min_D E\left\{\frac{I_{\{A\neq D(X)\}} Y}{\pi(A)}\right\} \\
&= \arg\min_D \sum_{j=1}^{k} E\{I_{\{D(X)\neq j\}} E(Y|A = j, X)\}.
\end{aligned} \tag{1}$$

We can see that $D^*(X) = \arg\max_{j\leq k} E(Y|A = j, X)$. That is the reason why traditional regression based approaches to estimate the optimal rule try to construct a good estimator of $E(Y|A = j, X)$, $j = 1, ..., k$. To avoid the estimation of conditional expectations as we discussed in the introduction, we focus on finding the optimal assignment rule by directly solving the minimization problem (1) via outcome weighted learning method.

Note that the optimal rule $D^*(X)$ does not change if we replace Y by $Y - c(X)$ for any function $c(X)$. Consequently, we can assume that $E(Y|A = j, X) \geq 0$ for all $j = 1, ..., k$. Since each treatment assignment rule $D(X) \in \{1, \ldots, k\}$ can be represented by $\arg\max_{j\leq k} f_j(X)$ for functions $f_1, ..., f_k$ on $\mathcal{X}$ with the sum-to-zero constraint $\sum_{j=1}^{k} f_j(X) = 0$, we can use the observed data $(Y_i, X_i, A_i), i = 1, ..., n$ to estimate the risk function in (1) by

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{\pi(A_i)} I_{\{A_i\neq D(X_i)\}} = \\
&\qquad \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{\pi(A_i)} I_{\{A_i\neq \arg\max_{j\leq k} f_j(X_i)\}}
\end{aligned} \tag{2}$$

where $\sum_{j=1}^{k} f_j(X_i) = 0$. However, it is difficult to solve the minimization problem over $f = (f_1, ..., f_k)$ due to the discontinuity and nonconvexity. To alleviate these difficulties, we use the vector hinge loss [13] and apply the technique of Reproducing Kernel Hilbert Space to estimate the treatment assignment rule.

Consider $f = (f_1, \ldots, f_k)$ with $f_j(x) = h_j(x) + b_j$, $x \in \mathcal{X}$, where b_j's are constants and h_j's are in the RKHS H_K associated with a positive definite kernel function K on $\mathcal{X} \times \mathcal{X}$, the closure of linear span of the set of functions $\{K(y, \cdot) : y \in \mathcal{X}\}$. Then the minimization problem in (2) turns into

$$\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{\pi(A_i)} \sum_{j=1}^{k} \left\{f_j(X_i) + \frac{1}{k-1}\right\}_+ + \frac{\lambda}{2}\sum_{j=1}^{k} \|h_j\|^2_{H_K} \tag{3}$$

over f with $f_j(x) = h_j(x) + b_j$, $x \in \mathcal{X}$, $h_j \in H_K$, and $\sum_{j=1}^{k} f_j(x) = 0$, where $\|\cdot\|^2_{H_K}$ is the squared norm in H_K generated by the inner product $\langle K_x, K_y\rangle_K = K(x, y)$, $K_x = K(x, \cdot)$, and λ is a tuning parameter.

2.2 Optimal treatment assignment rule via the functional analysis of variance decomposition

Unfortunately, the solution of (3) is a black-box function. We adopt the functional analysis of variance decomposition

to enhance the interpretability of covariates. First we review the functional analysis of variance decomposition as a structured representation of a multivariate function for describing a relationship f between covariates $x = (x_1, ..., x_p)$ and the response y, where $x \in \mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_p$ with $X_\alpha \in \mathcal{X}_\alpha$. The analysis of variance decomposition of f is

$$f(X) = b + \sum_{\alpha=1}^{p} f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(x_\alpha, x_\beta) + \cdots, \tag{4}$$

where b is a constant, and the functional components f_S for $S \subseteq \{1, ..., p\}$ satisfy side conditions for identifiability. The component f_α can be viewed as the main effect of x_α, $f_{\alpha\beta}$ as the two-factor interaction of x_α and x_β, and so on. For simplicity and estimation accuracy of f, the analysis of variance decomposition is truncated after lower-order interaction terms in practice.

The smooth function space that facilitates the analysis of variance decomposition in (4) is briefly described as below. Assume that the function f is in H, a reproducing kernel Hilbert space of functions defined on $\mathcal{X}$. Details about reproducing kernel Hilbert spaces and the general properties can be found in [2]. Then the space H is constructed as a tensor product of functional subspace H_α, a reproducing kernel Hilbert space of functions on H_α for $\alpha = 1, \ldots, p$, which can be further decomposed as $\{1\} \oplus \bar{H}_\alpha$, where $\bar{H}_\alpha$ is the subspace of H_α orthogonal to $\{1\}$. The space H is given by

$$\begin{aligned} H &= \otimes_{\alpha=1}^{p}(\{1\} \oplus \bar{H}_\alpha) \\ &= \{1\} \oplus \sum_{\alpha=1}^{p} \bar{H}_\alpha \oplus \sum_{\alpha<\beta} (\bar{H}_\alpha \otimes \bar{H}_\beta) \oplus \ldots \end{aligned} \tag{5}$$

Then the corresponding simplification of $f \in H$ is yield by truncating for higher-order interactions. Relabel the remaining truncated subspaces as F_v, for $v = 1, \ldots, d$, and let the resulting reproducing kernel Hilbert space be $F = \{1\} \oplus \bar{F}$, where $\bar{F} = \oplus_{v=1}^{d} F_v$. If $f \in F$, then f is represented as a sum of functional components. Using F, the general regularization approach turns into finding $\widehat{f} \in F$ so as to minimize

$$\frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda \sum_{v} \theta_v^{-1} \|P^v f\|^2,$$

where L is the loss function and $\|.\|$ is the norm defined on the reproducing kernel Hilbert space F, P^v is the orthogonal projection operator on to F_v, and $\theta_v \geq 0$. The minimizer is taken to satisfy $\|P^v f\|^2 = 0$ when $\theta_v = 0$. The penalty term $\sum_v \theta_v^{-1} \|P^v f\|^2$ with rescaling parameters θ_v entails the following reproducing kernel for $\bar{F}$:

$$K(s,t) = \sum_{v=1}^{d} \theta_v K_v(s,t) \tag{6}$$

for $s, t \in \mathcal{X}$, where K_v is the reproducing kernel for F_v. The tuning parameter θ_v amounts to rescale of the component spaces F_v, and both the set of θ_v values and λ affect the model complexity.

For structured representation of f, we consider the analysis of variance decomposition corresponding to functional subspaces in (5). Suppose that $f_j = b_j + h_j(x) \in \{1\} \oplus \bar{F}$, $j = 1, \ldots, k$. Then h_j can be expressed as $h_j = \sum_{v=1}^{d} h_{vj}$ with $h_{vj} \in F_v$. Similar to the LASSO method in linear models that produces sparse solutions, we impose an additional l_1 penalty on the sum of the parameters that can further force those covariates with negligible weights to be zero. The rescaling parameter θ_v for F_v allows a systematic way of selecting the most relevant components to Y. Thus, the regularization method tries to find the optimal rule so as to minimize

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{\pi(A_i)} \sum_{j=1}^{k} \left\{ f_j(X_i) + \frac{1}{k-1} \right\}_+ \\ &+ \frac{\lambda}{2} \sum_{j=1}^{k} \left(\sum_{v=1}^{d} \theta_v^{-1} \|h_{vj}\|^2_{H_{K_v}} \right) + \lambda_\theta \sum_{v=1}^{d} \theta_v \end{aligned} \tag{7}$$

s.t.,

$$\theta_v \geq 0, \quad v = 1, \ldots, d.$$

By the representer theorem, its solution admits a finite-dimensional representation. For fixed $\theta = (\theta_1, \ldots, \theta_d)^T$, substituting the rescaled reproducing kernel in (6) into the finite-dimensional representation, each coordinate of $\widehat{f}$ is given by

$$\widehat{f}_j(x) = b_j + \sum_{i=1}^{n} c_{ij} \sum_{v=1}^{d} \theta_v K_v(x_i, x), \tag{8}$$

with $\widehat{h}_{vj} = \theta_v \sum_{i=1}^{n} c_{ij} K_v(x_i, x)$ as the vth functional component.

Let K_v be a $n \times n$ matrix with (l, m)th entry $K_v(x_l, x_m)$ and set $b = (b_1, \ldots, b_k)^T$, $c = (c_1, \ldots, c_k)$ with $c_j = (c_{1j}, \ldots, c_{nj})^T$. By the reproducing property and (8), $\sum_{v=1}^{d} \theta_v^{-1} \|\widehat{h}_{vj}\|^2_{H_{K_v}} = c_j^T (\sum_{v=1}^{d} \theta_v K_v) c_j$. Given θ, let $K_\theta = \sum_{d=1}^{v} \theta_v K_v$. Then the risk function in (7) can be rewritten as a finite-dimensional problem of finding θ and (b, c) that minimizes

$$\begin{aligned} L(\theta, b, c) =& \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{\pi(A_i)} \sum_{j=1}^{k} \left\{ f_j(X_i) + \frac{1}{k-1} \right\}_+ \\ &+ \frac{\lambda}{2} \sum_{j=1}^{k} c_j^T K_\theta c_j + \lambda_\theta \sum_{v=1}^{d} \theta_v \end{aligned} \tag{9}$$

s.t.,

$$\begin{aligned} &\theta_v \geq 0, \qquad v = 1, ..., d, \\ &\sum_{j=1}^{k} (b_j \mathbf{1} + K_\theta c_j) = \mathbf{0}, \end{aligned} \tag{10}$$

where $\mathbf{1}$ is a vector whose components are all equal to one and $\mathbf{0}$ is a vector whose components are all equal to zero, for two vectors a and b, $a = b$ means that all components of $a - b$ are all zero. This finite-dimensional problem involves θ and (b, c) jointly. We use the iterative scheme in terms of two well-defined convex optimization problems referred to below as the c-step and θ-step for finding $\widehat{f}$. First, initialize $\theta^{(0)} = (1, \ldots, 1)^T$ and $(b^{(0)}, c^{(0)}) = \arg\min L(\theta^{(0)}, b, c)$. At the mth stage ($m = 1, 2, \ldots$), carry out the following two steps:

- θ-step: find $\theta^{(m)}$ to minimize $L(\theta, b^{(m-1)}, c^{(m-1)})$ with $(b^{(m-1)}, c^{(m-1)})$ fixed;
- c-step: find $(b^{(m)}, c^{(m)})$ to minimize $L(\theta^{(m)}, b, c)$ with $\theta^{(m)}$ fixed.

Note that the optimization problem in c-step reduces to the support vector machine with the reproducing kernel K_θ which is clearly explained in [13]. For the optimization problem in θ-step, let $\xi_j = (\xi_{1j}, \ldots, \xi_{nj})^T$ be the non-negative slack variables, $j = 1, \ldots, k$. Set $\xi = (\xi_1, \ldots, \xi_k)$ and let W be the $n \times n$ diagonal matrix whose ith diagonal entry is $Y_i/\pi(A_i)$. For fixed b and c, θ-step turns into an easy handling linear programming problem to find θ that minimizes

$$(11)\quad L(\theta, \xi) = \frac{1}{n}\sum_{j=1}^{k} W\xi_j + \sum_{v=1}^{d}\theta_v \left\{ \frac{\lambda}{2}\sum_{j=1}^{k} c_j^T K_v c_j + \lambda_\theta \right\},$$

s.t.,

$$(12)\quad \begin{aligned} & b_j\mathbf{1} + \sum_{v=1}^{d}\theta_v K_v c_j + (k-1)^{-1}\mathbf{1} \le \xi_j, \qquad j = 1, ..., k, \\ & \xi_j \ge \mathbf{0}, \qquad j = 1, ..., k, \\ & \theta_v \ge 0, \qquad v = 1, ..., d. \end{aligned}$$

Let $\widehat{f}^{(m)}$ denote the minimizer at the mth step.We can see that $\widehat{f}^{(0)}$ is solution of the support vector machine with $\theta^{(0)}$. We now show the asymptotic property of the risk corresponding to $\widehat{f}^{(m)}$ as generated by the alternating algorithm.

Theorem 1. Given λ and λ_θ, the algorithm yields a sequence of $\widehat{f}^{(m)}$ with feasible $(\theta^{(m)}, b^{(m)}, c^{(m)})$ and nonincreasing $L(\theta^{(m)}, b^{(m)}, c^{(m)})$; that is, $L(\theta^{(m+1)}, b^{(m+1)}, c^{(m+1)}) \le L(\theta^{(m)}, b^{(m)}, c^{(m)})$.

Corollary 1. Given λ and λ_θ, the sequence of $L(\theta^{(m)}, b^{(m)}, c^{(m)})$ generated by the algorithm converges as $m \to \infty$.

Actually, the original minimization problem (1) amounts to compare the magnitude of $E(Y|A = j, X)$ with different treatments, i.e., treating different treatments equally. Nowadays we always need to consider weighted versions of the conditional expectations in some medical problems. For example, if treatment j is more expensive, toxic, or laborious than treatment l, then we may only prefer $A = j$ when $E(Y|A = j, X)$ is larger than $E(Y|A = l, X)$ to a certain factor. We can extend our approach to weighted version via the $0-q$ loss. More details can be found in [13]. Although the outcome weighted learning method in our approach shares that in [13], the recognized contribution of our work is the functional variance decomposition to produce sparse solutions, which improves the interpretability of prediction.

3. SIMULATION RESULTS

Some simulation studies were conducted to evaluate the finite sample performance of the proposed method and compare the proposed method with the following two methods. The first method, called one versus others, applies the method for two treatments in [25] to compare treatment j versus all others, $j = 1, ..., k$, and then picks the best in these k comparisons as the optimal solution. The second method is the outcome weighted learning framework for multiple treatments proposed by [13] that shares good theoretical properties for the case of $k \ge 2$. To evaluate the performance of these three methods, we generated a independent validation data set following exactly the same procedure as the training data set except that the sample size is 1,000. The performances are assessed by two criteria: the misclassification error rate of the estimated optimal rule compared with the true optimal rule and the magnitude of the excess risk $R(\widehat{f}) - R^*$ of rule $\widehat{f}$. We consider the following eight scenarios. The training dataset was generated as follows.

Scenario 1:
$X = (X^{(1)}, X^{(2)})$ and $X^{(1)}$ and $X^{(2)}$ were independently generated from uniform distribution $U(0, 1)$; the treatment A was generated from $\{1, 2, 3\}$ independently of covariates with equal probability $1/3$. The true optimal treatment A^* is 1 if $X^{(1)} \le 1/3$, 3 if $X^{(1)} \ge 2/3$, and 2; otherwise, the outcome variable $Y^{(A)} = 2I_{\{A^*=A\}} + X^{(2)}$.

Scenario 2:
$X = (X^{(1)}, ..., X^{(4)})$ and $X^{(\nu)}$'s were generated independently from $U(0, 1)$; the actual treatment and the optimal treatment were generated the same as in the first scenario; the outcome $Y^{(A)} = 2I_{\{A^*=A\}} + (X^{(1)})^2 + \exp\{-X^{(3)} - X^{(4)} + \epsilon\}$, where ϵ is a noise term generated independently with X from $U(0, 1)$.

Scenario 3:
It is the same as scenario 1 except that $Y = I_{\{A^*=A\}} + X^{(2)}$.

Scenario 4:
It is the same as the second scenario except that $X^{(3)}$ is binary from a Bernoulli distribution with success probability 0.5 and the outcome $Y = I_{\{A^*=A\}} + 1 + (0.2X^{(1)} + 0.25X^{(2)})^2 - X^{(3)}$.

Table 1. Two-dimensional three-treatments example in Scenario 1, $n = 200$. Vector hinge loss minimizing values of λ_θ at the θ-step and λ at the c-step

Iteration	$\log_2(\widehat{\lambda}_\theta)$ (CV: hinge)	$\log_2(\widehat{\lambda}_\theta)$ (CV: hinge)	$\widehat{\theta}_1$	$\widehat{\theta}_2$
0		-16 (1.8329)	1	1
1	-2 (1.8212)	-16 (1.8172)	1	0
2	-2 (1.8516)	-16 (1.8190)	1	0
3	-2 (1.8649)	-16 (1.8210)	1	0

Scenario 5:
It is the same as scenario 1 except that the true optimal treatment A^* is 1 if $0.5(X^{(2)}-0.5)^2 - X^{(1)} + 0.65 < 0$, 3 if $0.5(X^{(2)}-0.5)^2 + X^{(1)} - 0.4 > 0$, and 2 otherwise.

Scenario 6:
It is the same as scenario 1 except that the true optimal treatment is the same as scenario 1 with probability 0.9 and randomly assigned to the other treatments with probability 0.1.

Scenario 7:
The covariate and outcome were generated the same as in scenario 1; the treatment A was generated from $\{1, 2, 3, 4\}$ independently of covariates with equal probability 1/4; the true optimal treatment A^* is 1 if $X^{(1)} \le 1/4$, 2 if $1/4 < X^{(1)} \le 2/4$, 3 or 4 similarly.

Scenario 8:
It is the same as scenario 7 except that the true optimal treatment A^* is 1 if $0.5(X^{(2)}-0.5)^2 - X^{(1)} + 0.7 < 0$, 3 if $0.3 < 0.5(X^{(2)}-0.5)^2 + X^{(1)} \le 0.55$, 4 if $0.5(X^{(2)}-0.5)^2 + X^{(1)} \le 0.3$, and 2 otherwise.

There are three treatments in scenarios 1-6, but with various optimal treatment structures and different outcome structures. In particular, Scenario 1 considers a simple linear boundary with one covariate for the optimal treatment. Scenario 2 involves a more complex main effect structure compared to Scenario 1. Thus we can examine the impact of the main effect when the optimal treatment structure is the same as that in Scenario 1. Scenario 3 examines the effect of reduced treatment interaction. Scenario 4 is the same as Scenario 2 but a binary covariate is used instead of all continuous covariate. Scenario 5 has a nonlinear boundary in the optimal rule with two covariates. We apply a non-zero Bayes error, 0.1, in Scenario 6. The training data sample sizes are 100, 150, and 200 in all these six scenarios. The remaining scenarios 7-8 involves four treatments. In particular, Scenario 7 has a linear boundary similar to Scenario 1 and Scenario 8 has a nonlinear boundary similar to Scenario 5. For scenarios 7-8, we consider sample sizes 200, 300, and 400.

Although any positive definite function K can be chosen as a reproducing kernel in the proposed method, we consider only flexible and structured kernels that facilitate the analysis of variance decomposition. Since reproducing kernels are closed under tensor summation and multiplication, we only define a univariate kernel function. For example, the spline kernel on the unit interval $[0, 1]$, $K(s,t) = k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s-t|)$ for s and $t \in [0,1]$, where $k_1(t) = t - \frac{1}{2}$, $k_2(t) = (k_1^2(t) - \frac{1}{12})/2$ and $k_4(t) = (k_1^4(t) - k_1^2(t)/2 + \frac{7}{240})/24$; more details can be found in [22]. For any covariate, we can do transformation so that it lies in $[0, 1]$ via

$$\tilde{x}_\alpha = \frac{x_\alpha - \min(x_\alpha)}{\max(x_\alpha) - \min(x_\alpha)},$$

where $\min(x_\alpha)$ and $\max(x_\alpha)$ are the minimum and the maximum values of the covariate in the training data set.

The tuning parameters λ and λ_θ are chosen by cross validation so as to minimize the prediction error determined by a loss function. We applied a five fold cross-validation procedure to tune the parameters. We adopted a one-step update procedure that alternates tuning of λ at the c-step and of λ_θ at the θ-step which is summarised as follows. Let $\widehat{E}$ denote a generic estimate of prediction error as a function of λ and λ_θ. The procedure consists of the following steps.

- Step 1. Initialize:
  - θ-step: initialize $\widehat{\theta}^{(0)}$;
  - c-step: find the initial multicategory support vector machine solution $(\widehat{b}^{(0)}, \widehat{c}^{(0)})$ that minimizes $L(\widehat{\theta}^{(0)}, b, c)$ in (9) at $\widehat{\lambda}^{(0)}$, which is a minimizer of $\widehat{E}(\lambda)$.
- Step 2. Update:
  - θ-step: find the rescaling parameters $\theta^{(1)}$ to minimize $L(\theta, b^{(0)}, c^{(0)})$ at $\widehat{\lambda}_\theta^{(1)}$, a minimizer of $\widehat{E}(\lambda_\theta)$;
  - c-step: find the one-step updated solution $(b^{(1)}, c^{(1)})$ to minimize $L(\theta^{(1)}, b, c)$ at $\widehat{\lambda}^{(1)}$, a new minimizer of $\widehat{E}(\lambda)$.

To show the reasonability of the one-step update procedure, we carried out a few more iterations in Scenario 1 and found that there is no noticeable change to the θ estimates, and the solutions from further iterations were virtually the same. Table 1 shows how the optimal pairs $(\widehat{\lambda}_\theta, \widehat{\lambda})$ changed and stabilized as we tuned λ at each c-step and λ_θ at each θ-step in the subsequent iterations. At the first θ-step, we have eliminated the irrelevant components and have correctly chosen the relevant components. This would result in

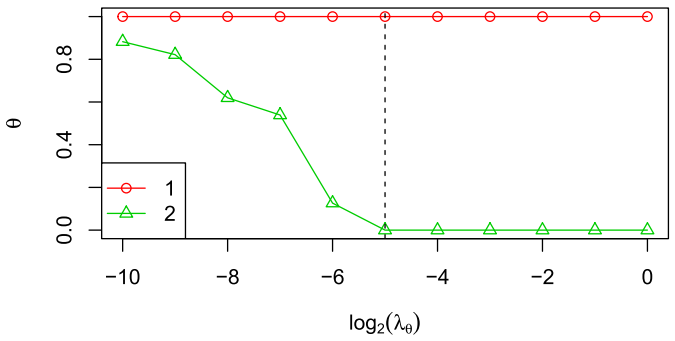

Figure 1. Two-dimensional three-treatments example in Scenario 1, $n = 200$. The trajectory for $\widehat{\theta}_1$ corresponding to X_1 is denoted by circle, that for $\widehat{\theta}_2$ corresponding to X_2 by triangle. The value of $(\widehat{\theta}_1, \widehat{\theta}_2) = (1, 0)$ at vector hinge loss largest minimizer $\widehat{\lambda}_\theta^{(1)} = 2^{-5}$ is indicated by dashed line.

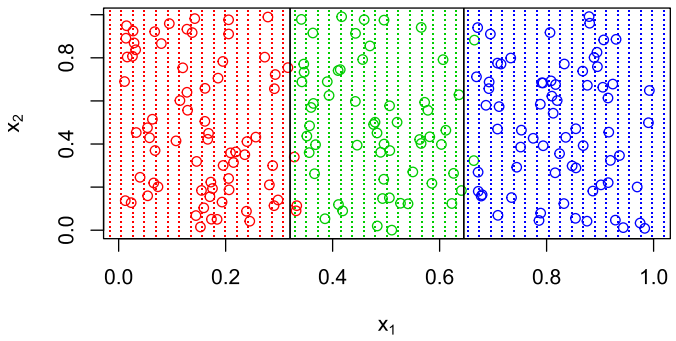

Figure 2. Two-dimensional three-treatments example of Scenario 1, $n = 200$. The two black lines are the obtained estimated optimal treatment rule corresponding to different treatments on each side of the two lines.

change of λ at the first c-step that completes the first iteration. Thus, the second θ-step might change little from $\widehat{\theta}^{(1)}$ because the second θ-step just reapplies the covariate selection procedure with relevant features only, which results in almost redundant subsequent iterations. This empirically justifies the one-step update procedure with sequential tuning as described in Section 2.

In order to explain the optimal treatment assignment rule based on the structured multi-class support vector machines, we use Figure 1 to show how the importance of covariates is reflected in Scenario 1. We first generate the training data set as described in Scenario 1 with sample size 200, including the randomized treatment A, covariates X and the corresponding outcome Y. Applying the proposed structured optimal treatment assignment rule to Scenario 1, we can draw the pathplot for the scaling parameters θ_1 and θ_2 as shown in Figure 1 along with the change of λ_θ. The larger of λ_θ, the smaller the scale of parameters θ_1 and θ_2, i.e., the smaller the number of non-zero parameters. At $\widehat{\lambda}_\theta^{(1)} = 2^{-5}$, the largest minimizer of vector hinge loss, $(\widehat{\theta}_1, \widehat{\theta}_2) = (1, 0)$ with $\widehat{\theta}_1$ being the only non-zero parameter. This indicates that the covariate X_1 indeed is an important factor affecting the optimal treatment assignment rule while the covariate X_2 does not play any role in estimating the optimal assignment rule. By using the structured support vector machine, we can have intuitive grasp of the importance of the covariates which is necessary in practical applications.

We finally get the optimal treatment assignment rule as shown in Figure 2. The two black lines are the obtained estimated optimal treatment rule that corresponds to different treatments on each side of the two lines. The color of dashed lines represents the corresponding estimated treatment assignment and the color of each circle represents the theoretical optimal treatment. The dashed lines and circles sharing the same color indicates accurate assignment estimation, while the different color represents the misclassification. Obviously, the overall assignment is quite good. Almost all the samples can be allocated to the theoretical optimal treatment. Similarly, the estimated optimal treatment assignment rule can be found for each scenario. But it is difficult to visualize intuitively as Figure 2 when the dimension of covariates is more than 2.

For each scenario, the number of simulation runs is 500. The results of misclassification rates and excess risk values can be found in Table 2 and Table 3. Some main conclusions are as follows: our proposed structured approach performs better than the other two methods in terms of smaller misclassification rate, excess risk, and lower standard deviation for each scenario; the improvement by using the proposed method may be substantial, e.g., the sample size $n = 100$ for the case in Scenario 2; the performance of the proposed method increases significantly as the sample size increases and the misclassification error rate and excess risk are quite small when the sample size is 200, which means that our method does not require a large number of samples to achieve high accuracy.

4. ANALYSIS OF A BREAST CANCER SCREENING STUDY

The breast cancer mammography screening for women on a regular manner is a common medical screening in attempt to achieve an earlier stage diagnosis and thus can significantly reduce mortality [1]. The women in the United States at normal risk for breast cancer is recommended to perform the mammography screening every two years in women between the ages of 50 and 74 [17]. The percentage of women in the U.S. who have had at least one mammogram is increasing; however, the rate for routine repeat screening is poor. The data are from a National Institute of Nursing Research (NINR) randomized controlled trial that included female subjects who were non-adherent to mammography screening guidelines at baseline, i.e., no mammogram in the year prior to baseline. One primary interest of the study was to test the efficacy of four tailored interventions to promote mammography screening at different post-baseline. The four

Table 2. Misclassification rates approximated by validation data set of size 1,000, averaged over 500 simulation runs; the numbers in parenthesis are standard deviations over 500 simulation runs

$k = 3$ treatments				
Scenario	Method	$n = 100$	$n = 150$	$n = 200$
1: linear boundary	Proposed	0.07 (0.06)	0.05 (0.04)	0.04 (0.02)
	Lou	0.10 (0.08)	0.06 (0.04)	0.04 (0.03)
	One vs others	0.13 (0.07)	0.09 (0.05)	0.06 (0.04)
2: complex main effect	Proposed	0.15 (0.15)	0.09 (0.07)	0.07 (0.06)
	Lou	0.25 (0.14)	0.15 (0.11)	0.10 (0.09)
	One vs others	0.34 (0.11)	0.24 (0.10)	0.18 (0.09)
3: reduced interaction effect	Proposed	0.14 (0.13)	0.08 (0.07)	0.06 (0.05)
	Lou	0.19 (0.13)	0.10 (0.08)	0.07 (0.06)
	One vs others	0.21 (0.10)	0.14 (0.08)	0.10 (0.07)
4: binary covariate	Proposed	0.21 (0.17)	0.12 (0.11)	0.08 (0.06)
	Lou	0.24 (0.15)	0.16 (0.13)	0.10 (0.09)
	One vs others	0.28 (0.11)	0.22 (0.10)	0.16 (0.08)
5: nonlinear boundary	Proposed	0.11 (0.05)	0.08 (0.03)	0.07 (0.02)
	Lou	0.13 (0.07)	0.09 (0.04)	0.08 (0.02)
	One vs others	0.15 (0.07)	0.11 (0.05)	0.09 (0.03)
6: positive Bayes error	Proposed	0.16 (0.08)	0.12 (0.04)	0.11 (0.03)
	Lou	0.18 (0.09)	0.13 (0.06)	0.11 (0.04)
	One vs others	0.21 (0.08)	0.17 (0.06)	0.14 (0.04)
$k = 4$ treatments				
Scenario	Method	$n = 200$	$n = 300$	$n = 400$
7: linear boundary	Proposed	0.07 (0.04)	0.05 (0.03)	0.04 (0.03)
	Lou	0.08 (0.06)	0.05 (0.03)	0.04 (0.02)
	One vs others	0.18 (0.08)	0.11 (0.05)	0.07 (0.04)
8: nonlinear boundary	Proposed	0.14 (0.05)	0.11 (0.03)	0.11 (0.02)
	Lou	0.14 (0.06)	0.11 (0.02)	0.11 (0.02)
	One vs others	0.21 (0.07)	0.14 (0.05)	0.12 (0.03)

interventions are (i) usual care (control), (ii) phone tailoring, (iii) mail tailoring, and (iv) mail and phone tailoring. These four interventions were based on sound theoretical models of behavior change (e.g., Health Belief Model). Variables in these theoretical models that promote mammography screening are measures of unobserved psychological constructs or beliefs. Specifically, psychological beliefs such as perceived benefits, barriers, self-efficacy, fear, susceptibility, and fatalism are outcomes in this study. And the tailoring interventions for women who are non-adherent to mammography screening guidelines at baseline have been shown to significantly increase mammography screening [4].

This data set has 1,244 women who had no mammogram in the year prior to baseline. After excluding some women with missing observations, we finally use a subset with 870 women for analysis. Among them, 253, 200, 237, and 180 women were assigned to usual care, phone tailoring, mail tailoring, and mail and phone tailoring, respectively. We use the 8 baseline variables, age (Age), race (Race), married or living with partner (Married), number of years had mammogram in last 5 years (Yearmam), doctor/nurse ever said to have a mammogram (Docspoke), currently working (Work), family history of breast cancer (Famhist) and whether more than high school (Educ) as predictors. These eight predictors are all collected in this trial and we want to use proposed method to identify important covariates and improve the interpretability. Since subjects were surveyed once pre-intervention and three times post-intervention about their mammography screening behavior, we independently treat the differences between average at three post baseline time points and the corresponding baseline values as outcomes.

To illustrate the application of our proposed optimal treatment assignment rule, we consider the mentioned NINR data set, which actually motivates our study. Based on our simulation experience, the sample size does not need to be too large. Hence we randomly select 400 observations to construct the treatment rule and use the remaining for validation. We repeat this procedure independently 500 replications to evaluate our method and identify important predictors. The results are presented in Table 4, which provides the screening rate for collected covariates under different outcomes over the 500 splits. We find that the number of years had mammogram in last 5 years (Yearmam) and age (Age) show great importance than others. The percentage of screening by using proposed method is more than 60% for Yearman and Age, which is consistent with the screen-

Table 3. Excess risk values approximated by validation data set of size 1,000, averaged over 500 simulation runs; the numbers in parenthesis are standard deviations over 500 simulation runs

$k = 3$ treatments				
Scenario	Method	$n = 100$	$n = 150$	$n = 200$
1: linear boundary	Proposed	0.15 (0.13)	0.11 (0.08)	0.08 (0.05)
	Lou	0.20 (0.16)	0.12 (0.09)	0.09 (0.07)
	One vs others	0.26 (0.15)	0.18 (0.10)	0.13 (0.09)
2: complex main effect	Proposed	0.29 (0.30)	0.17 (0.14)	0.13 (0.12)
	Lou	0.50 (0.29)	0.30 (0.23)	0.19 (0.18)
	One vs others	0.67 (0.23)	0.48 (0.20)	0.36 (0.18)
3: reduced interaction effect	Proposed	0.15 (0.13)	0.08 (0.07)	0.06 (0.05)
	Lou	0.19 (0.13)	0.10 (0.09)	0.07 (0.06)
	One vs others	0.21 (0.11)	0.14 (0.08)	0.10 (0.07)
4: binary covariate	Proposed	0.21 (0.17)	0.12 (0.12)	0.07 (0.07)
	Lou	0.18 (0.15)	0.10 (0.14)	0.04 (0.09)
	One vs others	0.22 (0.11)	0.16 (0.11)	0.10 (0.09)
5: nonlinear boundary	Proposed	0.21 (0.11)	0.17 (0.06)	0.15 (0.05)
	Lou	0.26 (0.15)	0.19 (0.09)	0.16 (0.06)
	One vs others	0.30 (0.15)	0.22 (0.10)	0.17 (0.08)
6: positive Bayes error	Proposed	0.31 (0.15)	0.24 (0.09)	0.22 (0.06)
	Lou	0.35 (0.18)	0.26 (0.12)	0.23 (0.08)
	One vs others	0.42 (0.16)	0.34 (0.13)	0.29 (0.09)
$k = 4$ treatments				
Scenario	Method	$n = 200$	$n = 300$	$n = 400$
7: linear boundary	Proposed	0.14 (0.09)	0.11 (0.07)	0.09 (0.06)
	Lou	0.17 (0.13)	0.10 (0.06)	0.08 (0.05)
	One vs others	0.37 (0.16)	0.22 (0.12)	0.14 (0.08)
8: nonlinear boundary	Proposed	0.27 (0.11)	0.23 (0.07)	0.22 (0.05)
	Lou	0.29 (0.13)	0.22 (0.06)	0.22 (0.06)
	One vs others	0.43 (0.14)	0.28 (0.11)	0.23 (0.08)

Table 4. The screening rate for collected covariates under different outcomes

Outcome	Yearmam	Age	Famhist	Work	Race	Married	Educ	Docspoke
Fear	0.79	0.66	0.46	0.37	0.37	0.39	0.32	0.34
Self efficacy	0.79	0.65	0.52	0.38	0.39	0.30	0.33	0.28
Barriers	0.75	0.70	0.48	0.41	0.37	0.29	0.28	0.32
Susceptibility	0.70	0.53	0.43	0.35	0.34	0.31	0.30	0.27
Benefits	0.69	0.63	0.42	0.35	0.32	0.30	0.29	0.29
Fatalism	0.68	0.60	0.45	0.40	0.34	0.34	0.32	0.32

ing results in [13]. The advantage of proposed method with the l_1 penalty is that we can have a intuitive judgment of the ranking for every covariates.

5. DISCUSSION

In this paper, we propose a structured outcome weighted learning procedure to provide a sparse nonparametric approach to search the optimal individualized treatment rule with multiple treatments, which shows superiority over other existing methods in the simulation studies. By using the structured kernel function and the l_1 penalty term, the structured approach has strong interpretation of covariates and shows which prognostic variables dominate in the assignment rule. To overcome the calculation shortcoming, we use a iterative scheme in terms of two well-defined convex optimization problems and we show that the resulting estimator shares the risk convergence. In practice, we adopt a one-step update procedure with reasonability shown in the simulation.

We can find in the simulation results that a large main effect of covariates may have negative impact on searching the optimal treatment assignment rule. If we try to modify the outcome weight, for example, [26] and [7] considered replacing Y_i by some type of residual which does not change consistency properties, then the finite sample performance of the rule may be better. The corresponding disadvantage is the model fitting to obtain residuals. We need to balance the

accuracy and model assumption in such a case. Furthermore, a dynamic treatment regime is a set of decision rules that determines the next treatment based on each individual's available characteristics and treatment history up to that point. Extension to dynamic treatment assignment rule can be an interesting direction for further research.

ACKNOWLEDGEMENTS

This research was partially supported by the First Class Discipline of Zhejiang – A (Zhejiang University of Finance and Economics – Statistics) (for Lou), by the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education (for Lou), by a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1409-21219 for Shao and Yu), by the US National Science Foundation Grant DMS-1612873 (for Shao). The views in this publication are solely the responsibility of the authors and do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee.

Received 6 August 2018

REFERENCES

[1] American Cancer Society (2016). *Breast Cancer Facts & Figures 2015–2016*. American Cancer Society Inc. 250 Williams Street, NW, Atlanta, GA 30303-1002.

[2] Aronszajn, N. (1950). Theory of reproducing kernel. *American Mathematical Society* **68** 337–404. MR0051437

[3] Buzdar, A. U. (2009). Role of Biologic Therapy and Chemotherapy in Hormone Receptor- and HER2-Positive Breast Cancer. *The Annals of Oncology* **20** 993–999.

[4] Champion, V. L. and Huster, G. (1995). Effect of interventions on stage of mammography adoption. *Journal of Behavioral Medicine* **18** 169–187.

[5] Chen, S., Tian, L., Cai, T. and Yu, M. (2017). A General Statistical Framework for Subgroup Identification and Comparative Treatment Scoring. *Biometrics* **73** 1199–1209. MR3744534

[6] Foster, J. C., Taylor, J. M. and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30** 2867–2880. MR2844689

[7] Fu, H., Zhou, J. and Faries, D. E. (2016). Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. *Statistics in Medicine* **35** 3285–3302. MR3528258

[8] Lagakos, S. W. (2006). The challenge of subgroup analyses—reporting without distorting. *New England Journal of Medicine* **354** 1667–1669.

[9] Lee, Y., Kim, S., Lee, S. and Koo, J.-Y. (2006). Structured multicategory support vector machines with analysis of variance decomposition. *Biometrika* **93** 555–571. MR2261442

[10] Lee, Y., Lin, Y. and Wahba, G. (2004). Multicategory Support Vector machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data. *Journal of the American Statistical Association* **99** 67–81. MR2054287

[11] Lin, Y. and Zhang, H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics* **34** 2272–2297. MR2291500

[12] Lipkovich, I., Dmitrienko, A. Denne, J. and Enas, G. (2011). Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine* **30** 2601–2621. MR2815438

[13] Lou, Z., Shao, J. and Yu, M. (2018). Optimal treatment assignment to maximize expected outcome with multiple treatments. *Biometrics* **74** 506–516. MR3825337

[14] Piper, W. E., Boroto, D. R., Joyce, A. S., McCallum, M. and Azim, H. F. A. (1995). Pattern of Alliance and Outcome in Short-Term Individual Psychotherapy. *Psychotherapy* **32** 639–647.

[15] Ruberg, S. J., Chen, L. and Wang, Y. (2010). The mean does not mean as much anymore: finding sub-groups for tailored therapeutics. *Clinical Trials* **7** 574–583.

[16] Schölkoph, B. and Smola, A. (2002). *Learning with Kernels—Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA: MIT Press. MR2498229

[17] Siu, A. L. (2016). Screening for breast cancer: U.S. preventive services task force recommendation statement. *Annals of Internal Medicine* **164** 279–296.

[18] Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M. and Li, B. (2009). Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research* **10** 141–158.

[19] Tibshirani, R. (1996). Regression selection and shrinkage via the LASSO. *Journal of the Royal Statistical Society: Series B* **58** 267–288. MR1379242

[20] Vansteelandt, S., VanderWeele, T. J., Tchetgen, E. J. and Robins, J. M. (2008). Multiply robust inference for statistical interactions. *Journal of the American Statistical Association* **103** 1693–1704. MR2510295

[21] Vapnik, V. N. (1998). *Statistical Learning Theory*. John Weily & Sons, New York. MR1641250

[22] Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM. MR1045442

[23] Xu, Y., Yu, M., Zhao, Y., Li, Q., Wang, S. and Shao, J. (2015). Regularized Outcome Weighted Subgroup Identification for Differential Treatment Effects. *Biometrics* **71** 645–653. MR3402600

[24] Zhang, B., Tsiatis, A., Davidian, M., Zhang, M. and Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat* **1** 103–114.

[25] Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107** 1106–1118. MR3010898

[26] Zhou, X., Mayer-Hamblett, N., Khan, U. and Kosorok, M. R. (2015). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association* **39** 1180–1210. MR3646564

Zhilan Lou
School of Data Sciences
Zhejiang University of Finance and Economics
Key Laboratory of Advanced Theory and Application
in Statistics and Data Science
(East China Normal University)
Ministry of Education
China
E-mail address: louzhilan@126.com

Jun Shao
School of Statistics
East China Normal University
China

Menggang Yu
Department of Biostatistics and Medical Informatics
University of Wisconsin
Madison
USA