BIOMETRIC METHODOLOGY

Biometrics WILEY

Variable selection in regression-based estimation of dynamic treatment regimes

Zeyu Bian[1] | **Erica E. M. Moodie[1]** | **Susan M. Shortreed[2,3]** | **Sahir Bhatnagar[1,4]**

[1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

[2] Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA

[3] Department of Biostatistics, University of Washington, Seattle, Washington, USA

[4] Department of Diagnostic Radiology, McGill University, Montreal, Quebec, Canada

Correspondence
Zeyu Bian, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Ave W, Montreal, Quebec H3A 1A2, Canada.
Email: zeyu.bian@mail.mcgill.ca

Abstract

Dynamic treatment regimes (DTRs) consist of a sequence of decision rules, one per stage of intervention, that aim to recommend effective treatments for individual patients according to patient information history. DTRs can be estimated from models which include interactions between treatment and a (typically small) number of covariates which are often chosen a priori. However, with increasingly large and complex data being collected, it can be difficult to know which prognostic factors might be relevant in the treatment rule. Therefore, a more data-driven approach to select these covariates might improve the estimated decision rules and simplify models to make them easier to interpret. We propose a variable selection method for DTR estimation using penalized dynamic weighted least squares. Our method has the strong heredity property, that is, an interaction term can be included in the model only if the corresponding main terms have also been selected. We show our method has both the double robustness property and the oracle property theoretically; and the newly proposed method compares favorably with other variable selection approaches in numerical studies. We further illustrate the proposed method on data from the Sequenced Treatment Alternatives to Relieve Depression study.

KEYWORDS

adaptive treatment strategies, double robustness, LASSO, penalization, precision medicine

1 | INTRODUCTION

Dynamic treatment regimes (DTRs) (Chakraborty and Moodie, 2013), or adaptive treatment strategies, consist of a sequence of decision rules that aim to improve individual patients' health outcomes by tailoring medical treatment to each patient's information. Statistical methods can be used to identify optimal DTRs, constructing treatment rules tailored over time to individual's information that can optimize the expected patient outcome.

DTRs can be estimated from models that include interactions between treatment and covariates, which are often chosen a priori. However, with many covariates and a complex disease process, for which competing treatment choices have heterogeneous effects, it is difficult to know which prognostic factors might be considered relevant in the treatment rule. A more data-driven approach of selecting these covariates might improve the estimated decision rules and simplify models to improve tractability. We are motivated by the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study (Fava *et al.*, 2003), a randomized multistage trial that aimed to determine optimal treatments for patients with major depressive disorder. With many of covariates such as demographic and clinical

characteristics collected throughout the study, it is challenging to select covariates useful for tailoring treatment from among so many based on expert knowledge only. Thus, variable selection with the objective of optimizing individualized treatment decisions becomes important.

Much of the DTR literature focuses on estimation; variable selection with the objective of optimizing treatment decisions has been considered only occasionally. Gunter *et al.* (2011) proposed a ranking method for variable selection in DTRs. Based on this approach, Fan *et al.* (2016) developed the sequential advantage selection approach, which considers variables already in the model when deciding whether to include a new variable by the additional improvement provided by this variable. Lu *et al.* (2013) adopted adaptive LASSO (Zou, 2006) in the context of A-learning (Murphy, 2003), Shi *et al.* (2018) proposed a method which used the Dantzig selector directly to penalize the estimating equations of A-learning and has the double robust property, that is, the estimators are consistent if either one of two nuisance models is correct. The topic of variable selection in a general (not DTR) context has seen many innovations (e.g., Tibshirani, 1996; Fan and Li, 2001). Gunter *et al.* (2011) noted that most variable selection approaches focus on predictive performance, and thus may not perform well in DTRs as these techniques may underestimate the importance of variables that have small predictive ability but that play a significant role in decision making.

In this article, we follow the DTR estimation approach of dynamic ordinary least squares regression (dWOLS) introduced by Wallace and Moodie (2015), an approach which requires only some minor precomputation and the implementation of standard weighted regression. While having similarities to both Q-learning (Watkins, 1989) and G-estimation (Robins, 2004), it provides simplicity and intuitiveness similar to the former and benefits from the double robustness of the latter although it is suitable only for linear decision rules. By adding two penalty terms in the dWOLS model, we perform estimation and variable selection for DTRs simultaneously. The rest of this article is organized as follows. In Section 2, we introduce the proposed penalized dWOLS (pdWOLS) approach, followed by algorithmic details and theoretical properties. Three simulation studies are given in Section 3. Finally, we apply our method to the STAR*D trial data in Section 4.

2 | METHODOLOGY

2.1 | Introductory concepts and notation

We make assumptions to proceed with estimation of DTRs: (1) Stable unit treatment value assumption

(SUTVA) (Rubin, 1980): A patient's potential outcome is not affected by other patients' treatment assignments. (2) Ignorability: Ignorability or no unmeasured confounding (Robins, 1997) specifies that for any possible treatment regimes, the stage k treatment is independent of future potential covariates or outcome conditional on current patient history. (3) No interference, no measurement error, and all the individuals have complete follow-up.

We adopt the setup of Wallace and Moodie (2015). For a K-stages DTR, the following notation is used, with lowercase being used for observed variables and uppercase for their random counterparts: y denotes patient outcome (continuous) which is measured at one point in time. The goal of DTRs is to make treatment decisions that can optimize (typically, maximize) the outcome. The kth binary treatment decision is, for example, $a_k = 1$ for treatment, $a_k = 0$ for standard care. Patient information available at time k and prior to kth treatment decision is denoted x_k. The covariate matrix containing patient history prior to the kth treatment decision is denoted $\boldsymbol{h}_k$; this history can include previous treatments $a_1, \dots, a_{k-1}$. Finally, $\overline{\boldsymbol{a}}_k = (a_1, a_2, \dots, a_k)$ is the vector of the first k treatment decisions, and $\underline{\boldsymbol{a}}_k = (a_{k+1}, a_{k+2}, \dots, a_K)$ is the vector of treatment decisions from stage $k + 1$ onward.

The blip (or contrast) function is defined as the difference in expected potential outcome between patients who received treatment a_k at stage k and patients who received a reference treatment denoted, say, $a_k = 0$, with the same history and assuming they receive optimal treatment after kth stage: $\gamma_k(\boldsymbol{h}_k, a_k) = \mathbb{E}[Y^{\overline{a}_k, \underline{a}_{k+1}^{opt}} - Y^{\overline{a}_{k-1}, a_k=0, \underline{a}_{k+1}^{opt}} | H_k = \boldsymbol{h}_k]$. The regret function (Murphy, 2003) is the expected loss resulting from giving treatment a_k at stage k instead of the optimal treatment a_k^{opt}, assuming optimal treatment is received after kth stage: $\mu_k(\boldsymbol{h}_k, a_k) = \mathbb{E}[Y^{\overline{a}_{k-1}, \underline{a}_k^{opt}} - Y^{\overline{a}_k, \underline{a}_{k+1}^{opt}} | H_k = \boldsymbol{h}_k]$. The blip and regret functions correspond directly: $\mu_k(\boldsymbol{h}_k, a_k) = \gamma_k(\boldsymbol{h}_k, a_k^{opt}) - \gamma_k(\boldsymbol{h}_k, a_k)$. This can be leveraged to simplify some expressions in later sections. Finally, we decompose the expected mean outcome into two components: $\mathbb{E}[Y^a | H = \boldsymbol{h}; \boldsymbol{\beta}, \boldsymbol{\psi}] = f(\boldsymbol{h}_0; \boldsymbol{\beta}) + \sum_{k=1}^{K} \gamma_k(\boldsymbol{h}_k, a_k; \boldsymbol{\psi}_k)$, where $f(\boldsymbol{h}_0; \boldsymbol{\beta})$ and $\gamma_k(\boldsymbol{h}_k, a_k; \boldsymbol{\psi}_k)$ are the so-called treatment-free and blip models, respectively, and $\boldsymbol{h}_0$ are baseline covariates. The function f, being free of any terms relating to the active treatment ($a_k = 1$), is irrelevant for making decisions about optimal treatment selection. For instance, in a simple one-stage setting, we could assume that both f and γ are linear in form: $f(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x$ and $\gamma(x, a; \boldsymbol{\psi}) = a(\psi_0 + \psi_1 x)$, and hence the estimated optimal treatment is $\widehat{a}^{opt} = I(\widehat{\psi}_0 + \widehat{\psi}_1 x > 0)$ where $I(\cdot)$ is the indicator function.

2.2 | Dynamic weighted ordinary least squares

Dynamic weighted ordinary least squares uses a sequential regression approach, similar to estimate the blip parameter ψ_k in the model for $\mathbb{E}[Y^a|\boldsymbol{H} = \boldsymbol{h}; \boldsymbol{\beta}, \boldsymbol{\psi}]$, achieving double robustness through weighting by a function of the propensity score (Rosenbaum and Rubin, 1983). The weights must satisfy $\pi(\boldsymbol{x})w(1, \boldsymbol{x}) = (1 - \pi(\boldsymbol{x}))w(0, \boldsymbol{x})$, where $\pi(\boldsymbol{x})$ is the propensity score and $w(a, \boldsymbol{x})$ is the weight for a subject with treatment a and covariates $\boldsymbol{x}$. Wallace and Moodie (2015) suggested to use "absolute value" weights of the form $w(a, x) = |a - \mathbb{E}[A|\boldsymbol{X} = \boldsymbol{x}]|$, as these offered better efficiency than other alternatives considered, while yielding consistent estimators of blip parameters if either the treatment or treatment-free model is correctly specified. Another assumption required by dWOLS is that the treatment-free model must include the main effects for all covariates in the blip model (unlike G-estimation, which can use an intercept-only treatment-free model). Violation of this assumption, known as the strong heredity principle (Chipman, 1996), can lead to biased estimators of blip parameters.

2.3 | Penalized dWOLS

We first introduce our approach in a one-stage setting with a continuous outcome, letting

$$Y = \beta_0 \boldsymbol{1} + \psi_0 \boldsymbol{A} + \sum_{j=1}^{p} X_j \beta_j + \sum_{j=1}^{p} \psi_j (\boldsymbol{A} \circ X_j) + \boldsymbol{\varepsilon}, \quad (1)$$

where $\boldsymbol{1}$ is the vector of 1's, $Y \in \mathbb{R}^n$ is a continuous response measured on n individuals, $X_j \in \mathbb{R}^n$ are the jth covariates, $X_i \in \mathbb{R}^p$ are covariates of ith individual, $\beta_j \in \mathbb{R}$ are the corresponding parameters for the main effects of covariates, $\psi_j \in \mathbb{R}$ are the blip parameters for $j = 0, 1, \dots, p$, $\boldsymbol{A}$ is the binary treatment indicator, "$\circ$" is the elementwise vector multiplication, and $\boldsymbol{\varepsilon}$ is an error term. This model is a simplification of Bhatnagar et al. (2020), which considers an additive interaction regression model. In this posited model, the treatment-free model is $\beta_0 + \sum_{j=1}^{p} X_j \beta_j$ and the blip model is $\psi_0 \boldsymbol{A} + \sum_{j=1}^{p} \psi_j (\boldsymbol{A} \circ X_j)$. To eliminate the intercept β_0, throughout this section, we center the response variable and each input variable in a weighted way, for example, using $Y - \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}$ instead of Y as the outcome.

For a continuous response, we use the weighted squared-error loss: $\mathcal{L}(Y; \theta) = \frac{1}{2n} \| \sqrt{W}(Y -$

$\psi_0 \boldsymbol{A} - \sum_{j=1}^{p} X_j \beta_j - \sum_{j=1}^{p} \psi_j(\boldsymbol{A} \circ X_j)) \|_2^2$, where $\theta = (\beta_1, \dots, \beta_p, \psi_0, \dots, \psi_p)$, and $W = \text{diag}\{w_1(a, x), w_2(a, x), \dots, w_n(a, x)\}$ is a *known* $n \times n$ diagonal matrix with $w_i(a, x)$ the "absolute value" weight for the ith individual. Similar to LASSO, we consider the following objective function that includes the ℓ_1 penalty for variable selection:

$$Q(\theta) = \mathcal{L}(Y; \theta) + \lambda(1 - \alpha)\|\boldsymbol{\beta}\|_1 + \lambda\alpha\|\boldsymbol{\psi}\|_1, \quad (2)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_p)$, $\lambda > 0$ and $\alpha \in (0, 1)$ are tuning parameters, and the solution is given by $\hat{\theta} = \arg\min_\theta Q(\theta)$. The parameter α controls the relative penalties for the main effects and the interaction effects. Other choices of the penalty term include the ℓ_2 penalty, the elastic net (Zou and Hastie, 2005), and the SCAD penalty. The ℓ_2 penalty yields ridge regression and hence cannot produce a sparse solution, and the ℓ_1 penalty cannot handle highly correlated variables very well (Zou and Hastie, 2005); the elastic net combines the ℓ_1 and ℓ_2 penalties, and thus can produce sparsity while offering good performance even when the features are highly correlated. The SCAD is a nonconvex penalty that can produce sparse solutions and nearly unbiased estimators.

An issue with Equation (2) is that since no constraint is placed on the structure of the model, it is possible that an estimated interaction term is nonzero while the corresponding main effects are zero, which violates the strong heredity assumption. To remedy this, our work is built on the strong heredity assumption, a constraint that is often used in practice when estimating interaction effects. Under the strong heredity assumption, an interaction term can be estimated to be nonzero if its corresponding main effects are estimated to be nonzero, whereas a nonzero main effect does not necessarily imply a nonzero interaction term. In DTR analysis, it is most common that there are more confounders than there are potential tailoring variables. Following Choi et al. (2010), we introduce a new set of parameters $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots \tau_p)$ and reparameterize the coefficients for the interaction terms ψ_j as a function of τ_j and the main effect parameters β_j and ψ_0: $\psi_j = \psi_0 \tau_j \beta_j$. In this way, strong heredity can be met, and we consider the following model:

$$\mathcal{L}^*(Y; \theta) = \frac{1}{2n} \| \sqrt{W}(Y - \psi_0 \boldsymbol{A} - \sum_{j=1}^{p} X_j \beta_j - \sum_{j=1}^{p} \underbrace{\psi_0 \tau_j \beta_j}_{\psi_j}$$

$(\boldsymbol{A} \circ X_j)) \|_2^2$, where now $\theta = (\beta_1, \dots, \beta_p, \psi_0, \tau_1, \dots, \tau_p)$. This reparameterized model is nonlinear as it involves products of parameters, and the objective function is expressed as

$$Q(\theta) = \mathcal{L}^*(Y; \theta) + \lambda(1 - \alpha)\|\boldsymbol{\beta}\|_1 + \lambda\alpha\|\boldsymbol{\tau}\|_1. \quad (3)$$

2.4 | Algorithm details

In this section, we describe a blockwise coordinate descent algorithm (Friedman *et al.*, 2007) for fitting the weighted least-squares version of the model in Equation (3). "Block-wise" means we breakdown the optimization problem into subproblems, that is, we fix the interaction terms τ and solve for the main effects ψ_0 and β and vice versa. Following Hastie *et al.* (2010), we fix the value for the tuning parameter α and minimize the objective function over a decreasing sequence of λ values ($\lambda_{max} > \dots > \lambda_{min}$).

Denote the n-dimensional residual column vector $\mathbf{R} = Y - \widehat{Y}$, where $\widehat{Y}$ is the current fitted value of $\mathbb{E}(Y)$ under the posited model. The subgradient equations are given by

$$\frac{\partial Q}{\partial \psi_0} = -\frac{1}{n}\left(\mathbf{A} + \sum_{j=1}^{p} \tau_j \beta_j \mathbf{A} \circ X_j\right)^{\top} W\mathbf{R} = 0, \qquad (4)$$

$$\frac{\partial Q}{\partial \beta_j} = -\frac{1}{n}\left(X_j + \tau_j \psi_0 \mathbf{A} \circ X_j\right)^{\top} W\mathbf{R} + \lambda(1-\alpha)s_1 = \mathbf{0}, \quad (5)$$

$$\frac{\partial Q}{\partial \tau_j} = -\frac{1}{n}\left(\psi_0 \beta_j \mathbf{A} \circ X_j\right)^{\top} W\mathbf{R} + \lambda \alpha s_2 = 0, \qquad (6)$$

where s_1 and s_2 are subgradients of the ℓ_1-norm, that is, $s_1 \in \text{sign}(\beta_j)$ if $\beta_j \neq 0$, $s_1 \in [-1, 1]$ if $\beta_j = 0$; $s_2 \in \text{sign}(\tau_j)$ if $\tau_j \neq 0$, $s_2 \in [-1, 1]$ if $\tau_j = 0$.

Define the partial residuals, without the jth predictor for $j = 1, \dots, p$, as

$$\mathbf{R}_{(-j)} = Y - \sum_{\ell \neq j} X_\ell \beta_\ell - \psi_0 \mathbf{A} - \sum_{\ell \neq j} \tau_\ell \psi_0 \beta_\ell (\mathbf{A} \circ X_\ell), \quad (7)$$

the partial residual without A as $\mathbf{R}_{(-A)} = Y - \sum_{j=1}^{p} X_j \beta_j$ and the partial residual without the jth interaction for $j = 1, \dots, p$, as

$$\mathbf{R}_{(-jA)} = Y - \sum_{j=1}^{p} X_j \beta_j - \psi_0 \mathbf{A} - \sum_{\ell \neq j} \tau_\ell \psi_0 \beta_\ell (\mathbf{A} \circ X_\ell). \quad (8)$$

From the subgradient equations (4)–(6), we see that

$$\widehat{\psi}_0 = \frac{\left(\mathbf{A} + \sum_{j=1}^{p} \tau_j \beta_j (\mathbf{A} \circ X_j)\right)^{\top} W\mathbf{R}_{(-A)}}{\left(\mathbf{A} + \sum_{j=1}^{p} \tau_j \beta_j (\mathbf{A} \circ X_j)\right)^{\top} W\left(\mathbf{A} + \sum_{j=1}^{p} \tau_j \beta_j (\mathbf{A} \circ X_j)\right)}, \qquad (9)$$

$$\widehat{\beta}_j = \frac{S\left(\left(X_j + \tau_j \psi_0 (\mathbf{A} \circ X_j)\right)^{\top} W\mathbf{R}_{-j}, n \cdot \lambda(1-\alpha)\right)}{\left(X_j + \tau_j \psi_0 (\mathbf{A} \circ X_j)\right)^{\top} W\left(X_j + \tau_j \psi_0 (\mathbf{A} \circ X_j)\right)}, \qquad (10)$$

$$\widehat{\tau}_j = \frac{S\left(\left(\psi_0 \beta_j (\mathbf{A} \circ X_j)\right)^{\top} W\mathbf{R}_{(-jA)}, n \cdot \lambda \alpha\right)}{\left(\psi_0 \beta_j (\mathbf{A} \circ X_j)\right)^{\top} W\left(\psi_0 \beta_j (\mathbf{A} \circ X_j)\right)}, \qquad (11)$$

where $S(x, u)$ is the soft-thresholding operator defined as $S(x, u) = \text{sign}(x)(|x| - u)_+$ (x_+ is the maximum value of x and 0).

The strong heredity assumption means that finding the λ which shrinks all coefficients to 0, is reduced to finding the smallest λ such that all *main effect* coefficients are shrunk to 0. From the subgradient equation (5), we see that $\beta_j = 0$ is a solution if $|\frac{1}{n}(X_j + \tau_j \psi_0(\mathbf{A} \circ X_j))^{\top} \mathbf{R}_{(-j)}| \leq \lambda(1-\alpha)$. From the subgradient equation (6), we see that $\tau_j = 0$ is a solution if $|\frac{1}{n}(\psi_0(\mathbf{A} \circ X_j)\beta_j)^{\top} \mathbf{R}_{(-jA)}| \leq \lambda \alpha$. Thus, the strong heredity assumption implies that the parameter vector $(\beta_1, \dots, \beta_p, \psi_1, \dots, \psi_p)$ will be entirely equal to $\mathbf{0}$ if $(\beta_1, \dots, \beta_p) = \mathbf{0}$. Therefore, the smallest value of λ for which the entire parameter vector reduces to $\lambda_{max} = \frac{1}{n(1-\alpha)} \max_j \{|(X_j)^{\top} \mathbf{R}_{(-j)}|\}$. The computational algorithm to fit all the parameters in a sequence of loops is further detailed in the Supplementary Material (Algorithm 1).

2.5 | Multiple intervals estimation

Knowing how to estimate the blip parameters in a one-stage setting, we now describe how the pdWOLS approach works in a K-stages setting. Starting from the last stage, the estimation procedure is applied to the Kth stage observed outcome y_K, treatment $\mathbf{a}_K$, and covariates x_K. The estimated blip parameters are obtained by maximizing the objective function in Equation (3) and the estimated rules $\widehat{a}_K^{opt} = I(\widehat{\psi}_{0K} + x_K \widehat{\psi}_K > 0)$, where I is the indicator function. The $(K-1)$th stage outcome is based on "optimal responses," that is, the estimation procedure is applied to the pseudo-outcome $\widetilde{y}_{k-1} = y_K + \mu_K(x_K, a_K; \widehat{\psi}_K)$, treatment $\mathbf{a}_{K-1}$, and covariates x_{K-1}, where $\mu_K(x_K, a_K; \widehat{\psi}_K) = \gamma_K(x_K, \widehat{a}_K^{opt}; \widehat{\psi}_K) - \gamma_K(x_K, a_K; \widehat{\psi}_K)$ is the regret function at stage K. The pseudo-outcome, $\widetilde{y}_{K-1}$, is optimal since the regret is added to the observed outcome y_K. The same procedure continues, recursively working backward, until stage 1 estimation, such that the blip parameters across all the stages are obtained and all treatment decisions can be made.

2.6 | Asymptotic properties of the pdWOLS estimator

We now show that when the number of predictors, p, is fixed and the sample size n approaches infinity, the pdWOLS estimator has both the double robustness and oracle properties (Fan and Li, 2001) under several assumptions. Following the adaptive LASSO (Zou, 2006), we add adaptive weights (or penalty factors) to the objective function (3) to obtain

$$\mathcal{L}^*(Y; \theta) + \lambda(1-\alpha) \sum_{j=1}^{p} w_j^{main} |\beta_j| + \lambda\alpha \sum_{j=1}^{p} w_j^{int} |\tau_j|, \quad (12)$$

where w_j^{main} and w_j^{int} are adaptive weights of main effect and interaction terms, respectively, in this way, the coefficients are not forced to be equally penalized in the ℓ_1 penalty. For instance, we can choose $w_j^{main} = |\widehat{\beta}_j^{wls}|^{-1}$ and $w_j^{int} = |\frac{\widehat{\beta}_j^{wls} \widehat{\psi}_0^{wls}}{\widehat{\psi}_j^{wls}}|$ for penalty factors, where $\widehat{\beta}_j^{wls}$ and $\widehat{\psi}_j^{wls}$ are unpenalized weighted least square estimates of the pdWOLS model. As n goes to infinity, the weights corresponding to unimportant variables go to infinity, which puts a large penalty on those variables, and the weights corresponding to important variables converge to a finite constant. Thus, small coefficients are removed, and large coefficients are unbiasedly estimated. Without loss of generality, we can rewrite Equation (12) as $\mathcal{L}^*(Y; \theta) + \sum_{j=1}^{p} \lambda_j^{\beta} |\beta_j| + \sum_{j=1}^{p} \lambda_j^{\tau} |\tau_j|$, where $\lambda_j^{\beta} = \lambda(1-\alpha)w_j^{main}$ and $\lambda_j^{\tau} = \lambda\alpha w_j^{int}$.

We assume that the true model follows the strong heredity assumption described above and regularity conditions detailed in the Supplemental Material hold. Note that the regularity conditions of pdWOLS are for quasi-likelihood since the loss function contains data-dependent weights and the treatment-free model may be misspecified. We describe the asymptotic properties of pdWOLS in the following theorems; proofs are given in the Supplemental Material. Assume that the observations $V_i, i = 1, \dots, n$ are independent and identically distributed with probability density $g(V)$ with respect to a measure ν. Denote the negative quasi-log-likelihood as $L_n^*(V; \theta) = -\sum_{i=1}^{n} \log h(V_i, \theta)$ (i.e., the dWOLS loss function), where h is the posited family of densities. Let θ^* be the underlying true parameters, and θ_* the minimizer of the Kullback–Leibler divergence between h and g (i.e., θ_* is the closest point to θ^* in the posited family of densities). Define B_1 as the indices of nonzero components for main effects and B_2 as the indices of nonzero components for interaction terms such that $B_1 = \{j : \beta_{*j} \neq 0\}, B_2 = \{j + p + 1 : \tau_{*j} \neq 0\}, B = B_1 \cup B_2$, where we define τ_* in a way

such that $\tau_{*j} = \frac{\psi_{*j}}{\psi_{*0}\beta_{*j}}$ if $\beta_{*j} \neq 0$ and 0 otherwise, since we assume the strong heredity property holds. Let na_n be the maximum value of the tuning parameters $(\lambda^{\beta}, \lambda^{\tau})$ such that the corresponding coefficients are nonzero and nb_n be the minimum value of the tuning parameters such that the corresponding coefficients are zero. For λ^{τ}, we only consider the index m such that $\beta_{*m} \neq 0$ and $\psi_{*m} = 0$ (i.e., $m \in B_1$): $a_n = \frac{1}{n}\max\{\lambda_j^{\beta}, \lambda_m^{\tau} : j \in B_1, m + p + 1 \in B_2\}$ $b_n = \frac{1}{n}\min\{\lambda_j^{\beta}, \lambda_m^{\tau} : j \in B_1^c, m + p + 1 \in B_2^c$ such that $\beta_{*m} \neq 0\}$.

Theorem 1. *Correct Sparsity: Assume that $\sqrt{n}a_n = O(1)$ and $\sqrt{n}b_n \to \infty$, then there exists a local minimizer $\widehat{\theta}_n$ of Equation (12) such that $\|\widehat{\theta}_n - \theta_*\| = O_p(n^{-\frac{1}{2}} + a_n)$. Moreover, we have $P(\widehat{\theta}_{B^c} = 0) \to 1$.*

Theorem 2. *Asymptotic Normality: Assume that $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$, then $\sqrt{n}(\widehat{\theta}_B - \theta_{*B}) \to_d N(0, \boldsymbol{J}^{-1}(\theta_{*B})\boldsymbol{I}(\theta_{*B})\boldsymbol{J}^{-1}(\theta_{*B}))$, where $\boldsymbol{J}(\theta) = -E_\theta[\frac{\partial^2 \log h(V; \theta)}{\partial \theta \partial \theta^T}]$ and $\boldsymbol{I}(\theta) = E_\theta[(\frac{\partial \log h(V; \theta)}{\partial \theta})(\frac{\partial \log h(V; \theta)}{\partial \theta})^T]$.*

Remark 1. Oracle properties of $\widehat{\theta}_n$ are established such that the estimator converges to some population parameter instead of the underlying true parameter θ^*. Also, the asymptotic covariance matrix no longer equals the inverse of the Fisher's information matrix. If the treatment-free model is correctly specified, then $\widehat{\theta}_n$ will converge to θ^*. To mimic the oracle, we further assume that all the observational weights are 1 (e.g., as in a randomized study).

Corollary 1. *Double Robustness: Assume that the blip function is correctly specified and SUTVA and ignorability described in Section 2.1 hold, then the resulting blip parameter estimators of pdWOLS are doubly robust; the estimators are consistent (i.e., $\psi_* = \psi^*$) if either the treatment model or the treatment-free model is correct. Note that correct specification of the blip model permits overspecification—that is, the true blip model may be contained within the analyst-specified model. From Theorems 1 and 2, pdWOLS has the same performance as dWOLS, and hence it has the double robustness property.*

Remark 2. There are no consistency guarantees for the first-stage estimator if an important confounder is missing in the second-stage model, as this violates an assumption at the second stage such that the estimator of second-stage parameters (subsequently plugged into the first-stage estimating function) may be biased. However, if estimation at the second stage is consistent (no unmeasured confounding, at least one of the nuisance models correct, etc.), then

double robustness at the first stage can be assured under key assumptions.

3 | SIMULATION STUDIES

In this section, we first illustrate the double robustness of pdWOLS and compare its performance to competing approaches through a number of simulations; then we implement the proposed method in a high-dimensional setting where $p > n$. Finally, we present simulation results for a two-stage setting. The tuning parameter α was set to 0.5 for all simulations, and λ was selected using fourfold cross-validation to reduce the computational burden.

In addition to assuming that there are no unmeasured confounders, we assume that the number of confounders is relatively small, so that the propensity score model can be fitted using logistic regression with the entire vector X. The propensity score is used to ensure balance between treatment groups. If model misspecification is a concern, one can use data-adaptive techniques, however, care must be taken in using data-adaptive approaches to estimating the propensity score to avoid the risk of selecting instruments, that is, variables that only predict treatment (Shortreed and Ertefaie, 2017). To consider a general framework, main effects are penalized in Equation (3). However, in a low-dimensional setting, we may want to retain all available covariates in the outcome model to ensure no weak confounders are erroneously omitted. In such cases, we can choose to not penalize the main effects, setting the corresponding penalty factors in Equation (12) to zero.

3.1 | Competing methods

We compare the variable selection results, error rate (in terms of the estimated rules as compared to the true optimal treatment), and out-of-sample value (i.e., expected outcome) under the estimated rules of pdWOLS with Q-learning combined with LASSO (Blatt *et al.*, 2004) and penalized A-Learning (PAL) (Shi *et al.*, 2018). Q-learning is a sequential regression approach to DTR estimation; relying only on outcome models; it is not doubly robust. PAL first estimates the treatment-free and propensity score models, then uses the Dantzig selector (Candes and Tao, 2007) to penalize the estimating equations of A-learning: $\widehat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}} \|\boldsymbol{\psi}\|_1$ subject to $\|X^T \text{diag}(\boldsymbol{A} - \widehat{\boldsymbol{\pi}})(Y - f(x;\widehat{\boldsymbol{\beta}}) - \gamma(x,a;\boldsymbol{\psi}))\|_\infty \leq n\lambda_{pal}$, where λ_{pal} is the tuning parameter and $\widehat{\boldsymbol{\pi}}$ is the estimated propensity score.

LASSO was implemented using the R package glmnet (Hastie *et al.*, 2010) with λ_{LASSO} selected via fourfold cross-validation. PAL was implemented using the R package ITRSelect (Shi *et al.*, 2018) with the tuning parameter

λ_{pal} selected via the Bayesian information criteria (BIC) (Schwarz, 1978). The main effect of treatment A is not penalized in any of the three methods. We also present unpenalized estimates of the blip parameters from a two-step approach: that is, after variable selection, the blip parameters are recalculated by solving the unpenalized weighted least squares via Q-learning, dWOLS, and A-learning with the selected variables, which we term *refitted* procedure.

3.2 | Experiments examining double robustness property

We begin with a simple one-stage example with the following data generation procedure:

Step 1: Generate 10 covariates ($X_1 - X_{10}$) where X are multivariate normal with zero mean, unit variance, and correlation $\text{Corr}(X_j, X_k) = 0.25^{|j-k|}$ for $j, k = 1, 2, \ldots, 10$.

Step 2: Generate treatment according to the model:
$$P(A = 1|X_1, X_2) = \frac{\exp(1 + x_1 + x_2)}{1 + \exp(1 + x_1 + x_2)}.$$

Step 3: Set the blip function, and hence the optimal treatment strategy, to depend only on X_1: $\gamma(x, a; \boldsymbol{\psi}) = a(\psi_0 + \psi_1 x_1)$ for $\psi_0 = 1, \psi_1 = -1.5$.

Step 4: Set the treatment-free model to $f(x; \boldsymbol{\beta}) = 0.5 - 0.6e^{x_1} - 2x_1 - 2x_2$.

Step 5: Generate the outcome $Y \sim N(f(x; \boldsymbol{\beta}) + \gamma(x, a; \boldsymbol{\psi}), 1)$.

We apply estimation and variable selection approaches with a variety of sample sizes (100, 500, and 2000) in four scenarios, where neither, one, or both of the treatment and treatment-free models is correctly specified. Specifically, the scenarios are: *Scenario 1 (neither treatment nor treatment-free is correct):* Regress Y on $(\mathbf{1}, X, A, AX)$, and set all observational weights to 1 (similar to assuming a null propensity score model). As this scenario fails to meet the assumptions of correct model specification, consistency is not assured for any approach. *Scenario 2 (treatment correct, treatment-free incorrect):* Regress Y on $(\mathbf{1}, X, A, AX)$, but fit a correctly specified propensity score model whose parameters are estimated via logistic regression. *Scenario 3 (treatment incorrect, treatment-free correct):* Regress Y on $(\mathbf{1}, e^{X_1}, X, A, Ae^{X_1}, AX)$, so that the treatment-free model is correctly specified but—as in Scenario 1—set all observational weights to 1. *Scenario 4 (both treatment and treatment-free are correct):* Regress Y on $(\mathbf{1}, e^{X_1}, X, A, Ae^{X_1}, AX)$, and estimate the parameters using a correctly specified propensity score.

Since Q-learning does not incorporate any propensity score adjustments, Scenarios 1 and 2 yield identical estimates, as do Scenarios 3 and 4. All the three methods have the same treatment-free models and the same blip

TABLE 1 Variable selection rate (%) of the blip parameters, error rate (ER, %) and value function over a testing set of size 10,000 under the estimated decision rules using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions ($n = 500$, 400 simulations)

	Scenario 2			Scenario 3			Scenario 4		
	pdWOLS	QL	PAL	pdWOLS	QL	PAL	pdWOLS	QL	PAL
Ae^{X_1}	-	-	-	72	14	72	42	14	0
AX_1[a]	100	100	99	100	100	33	100	100	100
AX_2	53	51	2	73	44	3	52	44	1
AX_3	2	26	2	6	23	0	2	23	1
AX_4	4	28	2	5	24	1	3	24	1
AX_5	4	29	4	4	26	2	2	26	2
AX_6	2	26	2	4	21	1	1	21	1
AX_7	3	25	2	5	22	1	2	22	0
AX_8	3	27	3	6	23	1	2	23	0
AX_9	2	27	3	6	24	1	2	24	1
AX_{10}	2	28	2	5	22	1	1	22	1
ER	3.9	9.8	22.9	5.5	3.4	12.0	4.2	3.4	23.4
ER (Refitted)	4.5	8.5	4.9	3.4	3.6	8.7	3.6	3.6	3.8
Value	0.6	0.6	0.5	0.6	0.7	0.6	0.6	0.7	0.5
Value (Refitted)	0.6	0.6	0.6	0.6	0.7	0.6	0.7	0.7	0.6

Note: The main effect of treatment is not penalized (and hence is always selected). Note that Ae^{X_1} was not included in the blip model for Scenario 2.
aTerm with a nonzero coefficient in the data-generating model.

functions to be estimated in the four scenarios. Across all scenarios where at least one nuisance model was correctly specified, refitted estimators performed better than their penalized counterparts in terms of bias (see Figure S1). When at least one of the treatment or treatment-free models was correctly specified, the blip parameter estimators were consistent for refitted pdWOLS. When the treatment-free model was correct (Scenarios 3 and 4), the refitted Q-learning (LASSO) estimators were consistent, as expected. Surprisingly, PAL failed when the treatment model was incorrect (Scenario 3). This result was not anticipated since PAL is a double robust method, although previous simulations have not considered its performance in terms of parameter estimates (Shi *et al.*, 2018).

The variable selection results for optimal treatment decisions are presented in Table 1. In Scenarios 2–4, the important tailoring variable was correctly selected by both pdWOLS and Q-learning (LASSO). PAL failed in Scenario 3. However, the false positive rates of pdWOLS and Q-learning (LASSO) were higher than that of PAL in all scenarios: for example, in Scenario 3, both LASSO and pdWOLS falsely selected the variable Ae^{X_1} 72% of the time.

Table 1 also summarizes the error rates (i.e., $\frac{1}{n}\sum_{i=1}^{n} I(a_i^{opt} \neq \hat{a}_i)$) of the estimated optimal treatment regimes for treatment decision making and value functions. The average value function and the error rates were computed over a testing set of size 10,000 (i.e., a data set generated according to the process described above in all respects except that treatment was allocated according

to the estimated rule). Both the error rate and the value of pdWOLS and Q-learning with LASSO were very close; pdWOLS outperformed other methods in Scenario 2, while Q-learning with LASSO had the best performance in Scenarios 3 and 4. The performance of the refitted versions of pdWOLS and Q-learning were similar; the performance of PAL was uniformly worse than the other methods performed without refitting, however refitting PAL substantially improved its performance.

3.3 | Simulations evaluating performance in a high-dimensional setting

Here, we present the performance of the new procedure in a high-dimensional setting with $p = 400$ and $n = 200$. The data generation procedure is the same as in Section 3.2, except that we now set $P(A = 1)$ to 0.5 for everyone such that no confounding is present. The blip function is $\gamma(x, a; \psi) = a(1 - 1.5x_1)$ where $\psi_0 = 1, \psi_1 = -1.5$ and the treatment-free model is $f(x; \beta) = 0.5 - 0.6e^{x_1} - 2x_1 - 2x_2$. We regress Y on $(1, X, A, AX)$ where the treatment-free model is misspecified.

Figure 1 summarizes the blip parameter estimates in the high-dimensional setting. Like before, for all the methods, refitted estimators improved the performance of their penalized counterparts. For ψ_0, Q-learning with LASSO and its refitted estimator had the smallest bias; as for ψ_1, pdWOLS and its refitted version had the smallest bias.
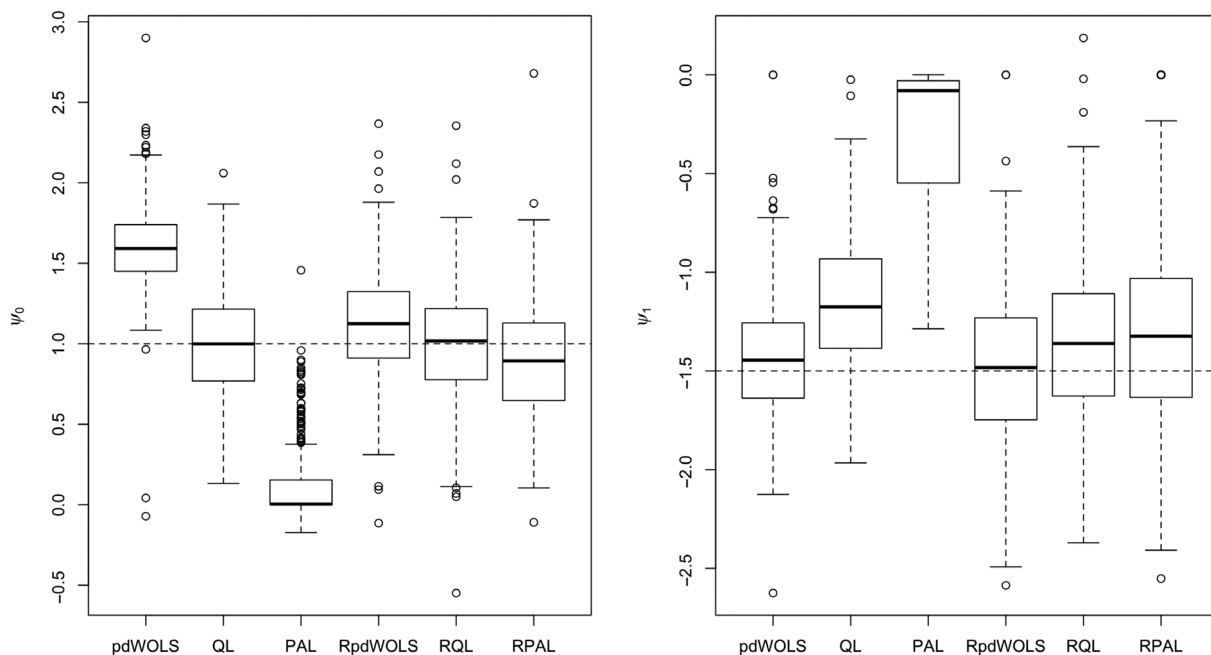
FIGURE 1 Estimates of blip parameters using pdWOLS, Q-learning (LASSO), PAL and their refitted versions with sample size 200 (400 simulations) in a high-dimensional ($p = 400$) setting. The true value is represented by the dotted line

TABLE 2 False negative (FN, %) rate and false positive (FP, %) rate of variable selection results of the blip parameters, error rate (ER, %) and value using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions with sample size 200 (400 simulations) in a high-dimensional ($p = 400$) setting

	FN	FP	ER	Value
pdWOLS	0.3	0.2	12.8	0.7
QL (LASSO)	0.0	1.4	11.3	0.7
PAL	2.6	0.4	24.6	0.6
RpdWOLS	0.3	0.2	9.9	0.7
RQL (LASSO)	0.0	1.4	16.8	0.6
RPAL	2.6	0.4	25.1	0.5

Note: The main effect of treatment is not penalized (and hence is always selected).

Table 2 shows false negative rates (the proportion of times a method wrongly removed a truly important variable), false positive rates (the proportion of times a method wrongly included a nonimportant variable), error rates, and the value under the estimated rules of the three methods. The average value function and the error rates were computed over a testing set of size 10,000. Q-learning with LASSO achieved a zero false negative rate; pdWOLS and refitted pdWOLS had the lowest false positive rate, error rate, and the highest value, which indicates favorable performance of the newly proposed method. However, unlike before, even the refitted PAL estimator had a smaller bias than the PAL estimator; refitted PAL did not improve the performance of PAL with respect to value and error rate, which shows that smaller bias in estimation of blip param-

eters does not necessarily translate into a better performance of the estimated regime.

3.4 | Simulations evaluating performance in multistage setting

In this subsection, we demonstrate the performance of the proposed pdWOLS approach when treatment decisions are made at multiple stages. We consider two different data generation procedures in order to follow previous literature. Setting 1, in which the true treatment-free model does not have an analytical closed form (misspecified treatment-free model) is presented here. Setting 2, in which the treatment-free models can be computed analytically, is available in the Supplemental Material.

We follow the data generation procedure in Wallace and Moodie (2015) with a sample size of 1000:

Step 1: Generate 10 covariates at stage 1: $X_{j1} \sim N(0, 1)$ for $j = 1, 2, \dots 10$.

Step 2: Generate treatment at stage k according to $P(A_k = 1 | X_{1k}, X_{2k}) = \frac{\exp(x_{1k} - x_{2k})}{1 + \exp(x_{1k} - x_{2k})}$ for $k = 1, 2$.

Step 3: Generate covariates at stage 2, such that $X_{12} \sim N(0.5A_1 + 0.8X_{11}, 1)$ and $X_{j2} \sim N(0.8X_{j1}, 1)$ for $j = 2, 3, \dots 10$.

Step 4: Set the blip functions to be $\gamma_1(x_1, a_1; \psi_1) = a_1(0.8 - 2x_{11})$ and $\gamma_2(x_2, a_2; \psi_2) = a_2(1 - 1.5x_{12})$, so that $\psi_{01} = 0.8, \psi_{11} = -2, \psi_{02} = 1$ and $\psi_{12} = -1.5$.

Step 5: Generate the outcome under optimal treatment according to $y^{opt} = 0.5 + 2x_{11} + 2x_{12}$. The observed
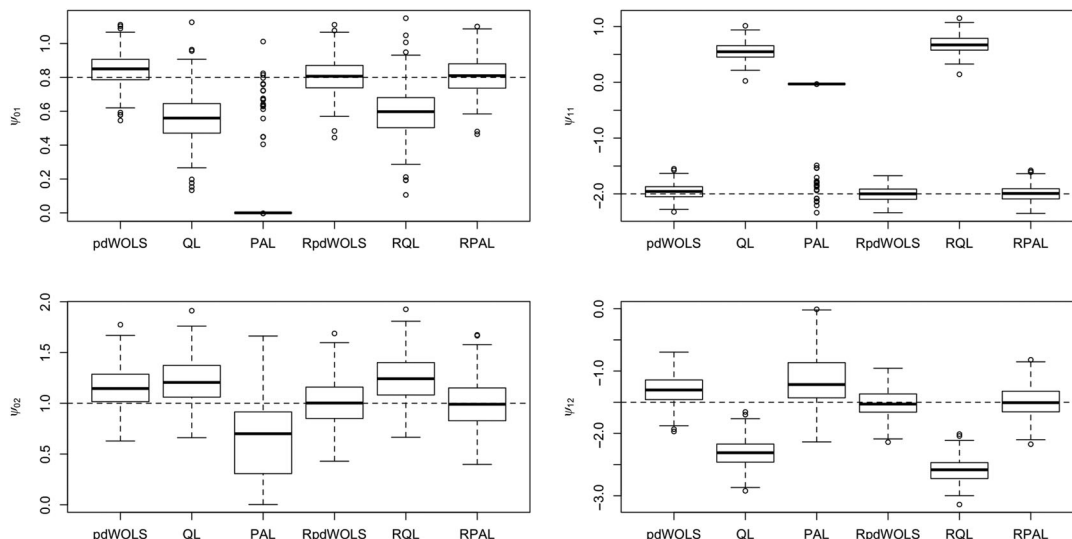
FIGURE 2 Estimates of blip parameters using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions with sample size 1000 (400 simulations) in two-stage Setting 1. The true value is represented by the dotted line

outcome is generated such that $Y \sim N(y^{opt} - \mu_1 - \mu_2, 1)$, where μ_1 and μ_2 are regret function at stages 1 and 2, defined through the blip functions in step 4.

Recall that a backward recursive approach can be used to make the treatment decision. Starting from the last stage, the estimation procedure is applied to the observed outcome y. The estimated blip parameters and the estimated rules, $\hat{a}_2^{opt}$, are obtained. Estimation then proceeds to stage 1, where again the estimation procedure is applied to a pseudo-outcome which represents the expected effect of the observed stage 2 treatment with the optimal stage 2 treatment. In pdWOLS, the pseudo-outcome is $\tilde{y}_1 = y + \gamma_2(x_2, \hat{a}_2^{opt}; \hat{\psi}_2) - \gamma_2(x_2, a_2; \hat{\psi}_2)$, whereas for Q-learning with LASSO, the pseudo-outcome is $\tilde{y}_1^Q = f(x_2; \hat{\beta}_2) + \gamma_2(x_2, \hat{a}_2^{opt}; \hat{\psi}_2)$.

In this setting, the treatment-free model in the second stage of estimation aims to represent $y^{opt} - \mu_1 - a_2^{opt}(1 - 1.5x_{12})$ which depends on a_2^{opt}, which in turn is a function of second-stage parameters ψ_2 and covariate x_2. The treatment-free model in this setting cannot be computed analytically. We nevertheless assumed that the treatment-free models were linear in the covariates measured at their respective stages, and thus in these simulations, it is always the case that the treatment-free models were misspecified. For those methods relying on a propensity score, the treatment models were fit using correctly specified logistic regression models at each stage using all covariates measured at that stage.

Figure 2 summarizes the estimates of blip parameters using the three methods in the two-stage Setting 1. As expected, pdWOLS and PAL work when at least one of the treatment or treatment-free models is correctly specified

(in this case, the treatment model is correctly specified), and Q-learning with LASSO failed, since the treatment-free model at both stages are misspecified. For pdWOLS and PAL, refitted estimators were nearly unbiased, and they performed better than their penalized counterparts. At stage 1, the bias of PAL estimators decreased to almost zero after refitting. Thus, PAL exhibits excellent performance in variable selection but requires the additional step of refitting for accurate estimation. Unlike PAL, pdWOLS can have small bias even without the refitting procedure.

Table 3 presents the variable selection results for optimal treatment decisions. The important tailoring variables were selected by all methods at both stages. At stage 2, the false positive rate of pdWOLS was much smaller than other two methods. For instance, the selection frequency of $AX_2 - AX_{10}$ were all less than 5%. Note that at stage 1, because the pseudo-outcomes were different for refitted version and their penalized counterparts, the variables selected by the procedures may differ between penalized and unpenalized implementations.

Table 4 summarizes the error rates of the estimated optimal treatment decisions and value functions, computed over a testing set of size 10,000. As before, refitted methods had lower error rate and higher value functions than their penalized counterparts. Penalized dynamic ordinary least squares outperformed other methods at both stages with respect to the error rate and value function; refitting greatly improved the performance of PAL.

In addition, we compared the choice of tuning parameter α, in order to assess sensitivity of the results to this choice; we considered values of 0.2, 0.5 (as in the analyses above), and 0.8. The results are presented in the Supplemental Material (Figure S4, Tables S3 and S4). To briefly

TABLE 3 Variable selection rate (%) of the blip parameters using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions with sample size 1000 (400 simulations) in two-stage Setting 1

	Stage 1						Stage 2		
	pdWOLS	QL	PAL	RpdWOLS	RQL	RPAL	pdWOLS	QL	PAL
AX_1[a]	100	100	100	100	100	100	100	100	100
AX_2	49	32	1	45	33	2	22	44	33
AX_3	4	28	2	2	34	2	2	41	38
AX_4	3	30	0	2	34	2	3	45	37
AX_5	3	25	1	2	29	2	2	40	40
AX_6	4	25	1	2	29	1	3	40	40
AX_7	4	27	0	2	33	2	3	40	38
AX_8	4	28	0	2	30	1	2	42	38
AX_9	4	26	1	2	32	4	2	41	36
AX_{10}	3	29	2	2	32	2	2	44	38

Note: The main effect of treatment is not penalized (and hence is always selected).
aTerm with a nonzero coefficient in the data-generating model.

TABLE 4 Error Rate (ER, %) and value function using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions with sample size 1000 (400 simulations) in two-stage Setting 1

	TER	ER (Stage 1)	ER (Stage 2)	Value
pdWOLS	9.2	2.2	7.2	0.4
QL (LASSO)	52.8	50.4	4.6	−0.6
PAL	22.4	14.5	9.8	0.4
RpdWOLS	6.5	2.0	4.6	0.5
RQL (LASSO)	58.0	54.8	6.1	−0.7
RPAL	11.3	2.1	9.5	0.4

Note: The total error rate (TER, %) in the estimated optimal treatment across both stages as well as the stagewise error rates are shown.

summarize, among all the α's, the bias and the variance of the estimators, the error rate and the estimated value were virtually identical. However, for variable selection, as α increased, the false positive rate decreased notably (See Table S3), as a larger α will put more penalty on the interaction terms.

4 | APPLICATION TO STAR*D STUDY

In this section, we apply pdWOLS to STAR*D data ([dataset] National Institute of Mental Health (NIMH), 2001) from the NIMH Data Archive, a multistage randomized trial that aimed to determine effective treatments for patients with major depressive disorder, where severity was measured using the Quick Inventory of Depressive Symptomatology (QIDS) score (Rush *et al.*, 2003). The study was divided into four levels (one of which had two sublevels); patients had different treatments at each level and would exit the study upon achieving remission. See the Supplemental Materials for details.

We follow Wallace *et al.* (2019) and Chakraborty *et al.* (2013) to perform two-stage analysis based on the use of a selective serotonin reuptake inhibitor (SSRI), with negative QIDS score as the outcome. Three tailoring variables were considered: (1) the QIDS score measured at the beginning of each level (denoted by q_k at stage k); (2) change in QIDS score divided by the time in the previous level (QIDS slope, denoted by s_k at stage k); and (3) patient preference measured prior to receiving treatment, which is a binary variable (denoted by p_k at stage k). We also generated d iid noise variables at each stage: noise variables at stage 1 were generated using $X_{j1} \sim N(0, 1)$ and at stage 2, $X_{j2} \sim N(\log|X_{j1}|, 1)$ for $j = 1, 2, \ldots, d$. We consider three scenarios for the analysis where $d = 5, 10, 20$, respectively.

Logistic regression was used to estimate the treatment model adjusting for patient preference only, following the trial design, and weights $w = |A - E(A|X)|$ were used in the analysis. As in Wallace *et al.* (2019), the treatment-free models were linear in (q_1, s_1, p_1) at stage 1 and (a_1, q_2, s_2, p_2) at stage 2. Linear blip models with covariates (q_1, s_1, p_1) at stage 1 and (a_1, q_2, s_2, p_2) at stage 2 were considered. Note in Wallace *et al.* (2019), a_1 and p_2 were not included in the blip models to avoid the multicollinearity; this is not necessary in pdWOLS, and hence our model specifications differ.

As in our simulations, the main effect of treatment was not penalized. In all three scenarios and both stages, pdWOLS returned the intercept-only blip model, suggesting that the optimal treatments are treat with SSRI ($A_1 = 1$) and treat with a non-SSRI ($A_2 = 0$) at stages 1 and 2, respectively, for all patients. PAL, in contrast, was sensitive to the number of noise variables: when $d = 5$, PAL selected a_j, q_j, s_j, p_j for both stages $j = 1, 2$. When $d = 10$, PAL selected a_2 at stage 2 and a_1, q_1, s_1, p_1 at stage 1, and when $d = 20$, PAL selected a_2 at stage 2 and a_1, p_1 at stage

1. Chakraborty *et al.* (2013) and Wallace *et al.* (2019) found that no stage 2 blip covariates were statistically significant (consistent with pdWOLS), while at stage 1, they found only treatment preference was significant.

The false positive rates of PAL at stages 2 and 1 were $100\%, 40\%(d = 5), 10\%, 50\%(d = 10)$, and $10\%, 45\%(d = 20)$, respectively; for pdWOLS, the rate was 0% for all d.

5 | DISCUSSION

In this article, we extended dWOLS to a penalized estimation framework for variable selection and estimating the optimal treatment regimes simultaneously. The proposed method inherits the double robustness property from dWOLS. Our simulations indicated that pdWOLS compares favorably with other variable selection approaches in the context of DTRs.

Our method automatically enforces strong heredity through a simple reparameterization, which guarantees an assumption required by dWOLS. The idea of reparameterization is simple, however, one limitation is that the objective function is nonconvex. Hence, it may be of interest, in future work, to investigate approaches that use convex constraints to achieve strong heredity. See, for example, Bien *et al.* (2013); Zhao *et al.* (2009), and Haris *et al.* (2016).

The standard errors for the estimated blip parameters can be obtained directly; a sandwich formula for computing the covariance of the estimates of the nonzero components can be derived (Fan and Li, 2001). How to derive the standard errors for the estimated blip parameters under the use of refitted pdWOLS requires further investigation. Postselection inference (Lee *et al.*, 2016) should also be addressed.

The proposed method is, fundamentally, based on prediction, selecting any variables that can improve predictive ability. As such, in finite samples, pdWOLS may underestimate the importance of variables that have small predictive ability but that play a significant role in DTRs. Besides, the application of predictive methods directly to causal models may result in inflated variances and self-inflicted bias (Hernán and Robins, 2020). The importance of the distinction between DTRs (causal inference) and prediction must be kept in mind. Variable selection in causal inference is a tough problem: on the one hand, we want to adjust for enough covariates in the analysis to achieve ignorability; on the other hand, adjustment for some other irrelevant variables could induce bias and losses of statistical efficiency (Rotnitzky *et al.*, 2010). Hence, a thoughtful selection of confounders is needed, using expert knowledge to guide variable selection is encouraged. Other discussions about confounder selection can be found in Short-reed and Ertefaie (2017), Robins and Greenland (1986), and Schneeweiss *et al.* (2009). For pdWOLS, if we are worried about confounding and our focus is on building simple rules, we may want to do minimal selection on main effects but lots of selection on interaction effects, which can be implemented by setting small adaptive weights w_j for the main effects or setting α to a large value. How to choose the tuning parameter λ and α in a DTR framework is an open and intriguing problem worthy of further investigation.

DATA AVAILABILITY STATEMENT
The data that support the findings in this paper are available in the National Institute of Mental Health Data Archive at https://nda.nih.gov/, collection ID number 2148.

ORCID
Zeyu Bian https://orcid.org/0000-0001-9849-9601
Erica E. M. Moodie https://orcid.org/0000-0002-7225-3977
Susan M. Shortreed https://orcid.org/0000-0001-7918-601X
Sahir Bhatnagar https://orcid.org/0000-0001-8956-2509

REFERENCES
Bhatnagar, S.R., Lu, T., et al. (2020) A sparse additive model for high-dimensional interactions with an exposure variable. *BioRxiv*, 445304.
Bien, J., Taylor, J. and Tibshirani, R. (2013) A lasso for hierarchical interactions. *Annals of Statistics*, 41, 1111–1141.
Blatt, D., Murphy, S.A., and Zhu, J. (2004) A-learning for approximate planning. *Ann Arbor*, 1001, 48109–2122.

Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35, 2313–2351.

Chakraborty, B., Laber, E.B. and Zhao, Y. (2013) Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69, 714–723.

Chakraborty, B. and Moodie, E.E.M. (2013) *Statistical Methods for Dynamic Treatment Regimes*. New York Boca Raton: Springer.

Chipman, H. (1996) Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24, 17–36.

Choi, N.H., Li, W. and Zhu, J. (2010) Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105, 354–364.

[dataset] NIMH. (2001) Sequenced treatment alternatives to relieve depression. NIMH Data Archive. Available at: https://nda.nih.gov/edit_collection.html?id=2148.

Fan, A., Lu, W. and Song, R. (2016) Sequential advantage selection for optimal treatment regime. *Annals of Applied Statistics*, 10, 32–53.

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

Fava, M., Rush, A.J., Trivedi, M.H., Nierenberg, A.A., Thase, M.E., Sackeim, H.A., et al. (2003) Background and rationale for the sequenced treatment alternatives to relieve depression (STAR* D) study. *Psychiatric Clinics of North America*, 26, 457–494.

Friedman, J., Hastie, T., et al. (2007) Pathwise coordinate optimization. *Annals of Applied Statistics*, 1, 302–332.

Gunter, L., Zhu, J. and Murphy, S. (2011) Variable selection for qualitative interactions. *Statistical Methodology*, 8, 42–55.

Haris, A., Witten, D. and Simon, N. (2016) Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25, 981–1004.

Hastie, T., Tibshirani, R. and Friedman, J. (2010) Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

Hernán, M.A. and Robins, J.M. (2020) *Causal Inference: What If*. Chapman & Hall/CRC.

Lee, J.D., Sun, D.L., Sun, Y. and Taylor, J.E. (2016) Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44, 907–927.

Lu, W., Zhang, H.H. and Zeng, D. (2013) Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, 22, 493–504.

Murphy, S.A. (2003) Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 65, 331–355.

Robins, J.M. (1997) Causal inference from complex longitudinal data. In Berkane, M. (Ed.) *Latent Variable Modeling and Applications to Causality: Lecture Notes in Statistics*. New York: Springer, pp. 69–117.

Robins, J.M. (2004) Optimal structural nested models for optimal sequential decisions. In Lin, D.Y. and Heagerty, P. (Eds.) *Proceedings of the Second Seattle Symposium in Biostatistics*. Springer, pp. 189–326.

Robins, J.M. and Greenland, S. (1986) The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology*, 123, 392–402.

Rosenbaum, P.R. and Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Rotnitzky, A., Li, L. and Li, X. (2010) A note on overadjustment in inverse probability weighted estimation. *Biometrika*, 97, 997–1001.

Rubin, D. (1980) Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, 75, 591–593.

Rush, A.J., Trivedi, M.H., et al. (2003) The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, 54, 573–583.

Schneeweiss, S., Rassen, J.A., et al. (2009) High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20, 512–522.

Schwarz, G.E. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Shi, C., Fan, A., Song, R. and Lu, W. (2018) High-dimensional A-learning for optimal dynamic treatment regimes. *Annals of Statistics*, 46, 925–957.

Shortreed, S.M. and Ertefaie, A. (2017) Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*, 73, 1111–1122.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Wallace, M.P. and Moodie, E.E.M. (2015) Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, 71, 636–644.

Wallace, M.P., Moodie, E.E.M. and Stephens, D.A. (2019) Model selection for G-estimation of dynamic treatment regimes. *Biometrics*, 75, 1205–1215.

Watkins, C.J.C.H. (1989) *Learning from delayed rewards*. PhD Thesis, King's College, Cambridge, UK.

Zhao, P., Rocha, G. and Yu, B. (2009) The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37, 3468–3497.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodological)*, 67, 301–320.

SUPPORTING INFORMATION

A Web Appendix containing the algorithm referenced in Section 2.4, regularity conditions and proofs of Theorems in Section 2.6, additional simulation results in Sections 3.2, 3.4, STAR*D details in Section 4 and an example of pdWOLS implemented in the R programming language are available with this paper at the Biometrics website on Wiley Online Library.