Journal of the Royal Statistical Society Series C: Applied Statistics, 2024, **73**, 298–313
https://doi.org/10.1093/jrsssc/qlad096
Advance access publication 2 November 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Variable selection for individualised treatment rules with discrete outcomes

Zeyu Bian[1,2], Erica E.M. Moodie[1], Susan M. Shortreed[3,4], Sylvie D. Lambert[5,6] and Sahir Bhatnagar[1]

[1]Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec H3A 0G4, Canada
[2]Miami Herbert Business School, University of Miami, Miami, FL 33146, USA
[3]Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA
[4]Department of Biostatistics, University of Washington, Seattle, Washington, USA
[5]Ingram School of Nursing, McGill University, Montreal, Quebec, Canada
[6]St.Mary's Research Centre, Montreal, Quebec, Canada

Address for correspondence: Zeyu Bian, Miami Herbert Business School, University of Miami, Miami, FL 33146, USA.
Email: zeyu.bian@miami.edu

Abstract

An individualised treatment rule (ITR) is a decision rule that aims to improve individuals' health outcomes by recommending treatments according to subject-specific information. In observational studies, collected data may contain many variables that are irrelevant to treatment decisions. Including all variables in an ITR could yield low efficiency and a complicated treatment rule that is difficult to implement. Thus, selecting variables to improve the treatment rule is crucial. We propose a doubly robust variable selection method for ITRs, and show that it compares favourably with competing approaches. We illustrate the proposed method on data from an adaptive, web-based stress management tool.

Keywords: double robustness, penalisation, precision medicine, variable selection, weighted generalised linear model

1 Introduction

In the precision medicine paradigm, treatment decisions are tailored to individuals rather than relying on a 'one-size-fits-all' approach. This approach to treatment is beneficial when treatment effects are heterogeneous. For example, effective management of stress requires the development of personalised approaches, as patients with different characteristics respond to and engage with treatments differently. With the aim of improving individuals' health outcomes, individualised treatment rules (ITRs) (Chakraborty & Moodie, 2013; Kosorok & Moodie, 2015; Murphy, 2003; Robins, 2004) recommend effective treatments based on each person's specific characteristics. However, collected data often contain many variables that are irrelevant for tailoring treatment. Including all variables in an analysis could reduce statistical efficiency by estimating unnecessary coefficients whose estimates fluctuate around zero for variables that are not useful for tailoring treatment, and yielding an unnecessarily complicated treatment decision rule that is difficult for physicians to interpret or implement. It is therefore important to develop variable selection methods with the objective of optimising individuals' outcomes by identifying useful tailoring variables.

Variable selection for ITRs has been studied in Lu et al. (2013), Jeng et al. (2018), Shi et al. (2018), and Bian et al. (2023), all of which focus on penalised regression-based estimation methods. Jeng et al. (2018) and Lu et al. (2013) considered only a singly robust method in which the propensity score must be correctly specified. Shi et al. (2018) used the Dantzig selector directly to penalise the A-learning (Robins, 2004) estimating equation; Bian et al. (2023) used penalised dynamic weighted ordinary least squares regression to perform variable selection. Zhang and Zhang (2018) and Zhang and Zhang (2022) extended the classification framework for estimating optimal treatment regimes in

Received: November 21, 2022. **Revised:** July 31, 2023. **Accepted:** September 29, 2023

Zhang et al. (2012) to a setting in which variable selection can be performed. In Zhang and Zhang (2018), variables are sequentially selected based on the additional improvement provided by the new variable, while Zhang and Zhang (2022) added a penalised term for the objective function to select the important variables. The methods considered in Zhang and Zhang (2018), Shi et al. (2018), Zhang and Zhang (2022), and Bian et al. (2023) are all doubly robust, i.e. they yield consistent estimators while requiring only one of two nuisance models to be correct.

All of the aforementioned methods focus solely on the case in which the outcome is continuous. Discrete outcomes introduce additional computational challenges to the estimation of ITRs and the variable selection procedure, due to the common use of a non-identity link function. Existing literature focusing on discrete outcomes ITR estimation includes Q-learning (Chakraborty & Moodie, 2013; Linn et al., 2017), Bayesian additive regression trees (Logan et al., 2019), and A-learning (Robins et al., 1992; Tchetgen Tchetgen et al., 2010). However, none of these approaches has been extended to include variable selection. Tian et al. (2014) proposed a straightforward method for estimating ITRs while performing tailoring variable selection by omitting all main effect terms for covariates, and re-scaling the covariates in the interaction terms. In this approach, the binary outcome case was also considered, although only in the randomised-treatment setting. Chen et al. (2017) further generalised the method in Tian et al. (2014) to observational studies for ITR estimation and variable selection. Nevertheless, the proposed approach for binary outcomes requires the propensity score to be correctly specified. A further augmentation of Tian et al. (2014) and Chen et al. (2017) was discussed in Chen et al. (2017) for binary outcomes, yet this augmentation approach still cannot achieve the desired double robustness property, because of the use of the non-linear loss function (see Remark 2 in Tian et al., 2014 for a more detailed explanation). In other words, the augmentation idea for binary outcomes in Chen et al. (2017) is used mainly for the resulting efficiency gain; a correct specification of the treatment model is still needed for consistent ITR estimation even if the outcome model is correctly specified. In this article, we focus on developing doubly robust ITR estimation with variable selection for discrete outcomes (count and binary outcomes).

To provide robustness against model mis-specification, ITRs are often estimated using estimating equations (Murphy, 2003; Robins, 2004). There are at least two ways to achieve sparsity in the use of estimating equations: via a Dantzig selector (Candes & Tao, 2007) or by a regularised estimating equation (REE). Denote by $\boldsymbol{U}(\boldsymbol{\theta}) \in \mathbb{R}^p$ an estimating equation, where $\boldsymbol{\theta} \in \mathbb{R}^p$. The Dantzig estimator $\widehat{\boldsymbol{\theta}}_{\text{dan}}$ can be found by solving the constrained optimisation problem: $\widehat{\boldsymbol{\theta}}_{\text{dan}} = \text{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_1$, subject to $\|\boldsymbol{U}(\boldsymbol{\theta})\|_\infty \leq n\lambda$, where λ is a tuning parameter used to control sparsity, and n is the sample size. Another way to induce sparsity is to solve the REE: $\boldsymbol{U}(\boldsymbol{\theta}) = n\lambda q(|\boldsymbol{\theta}|)$, where $q(|\cdot|)$ is the subgradient of a penalty function $\rho(|\cdot|)$, i.e. $q(|\cdot|) = \partial\rho(|\cdot|)$. For example, lasso (Tibshirani, 1996) regression defined by $\min_{\boldsymbol{\beta}} \{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + n\lambda\|\boldsymbol{\theta}\|_1\}$ is a special case of the REE $\boldsymbol{U}(\boldsymbol{\theta}) = n\lambda\partial\|\boldsymbol{\theta}\|_1$, where $\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix, and $\boldsymbol{Y} \in \mathbb{R}^n$ is the response.

While the Dantzig selector and REE work well for continuous outcomes (Shi et al., 2018), their implementation in ITRs can be difficult for discrete outcomes, which are usually modelled with non-identity link functions. Indeed, the existing doubly robust estimating equations to estimate ITRs for discrete outcomes are non-linear (Robins et al., 1992; Tchetgen Tchetgen et al., 2010, see later in Section 2.2), and hence the Dantzig selector cannot be solved using linear programming (James & Radchenko, 2009). As for REE, it has been studied in Johnson et al. (2008) and Wang et al. (2012) using local quadratic approximation (Fan & Li, 2001) to solve the REE, which is computationally burdensome since it requires the calculation of the inverse of the Hessian matrix. Finally, even if the solution of the Dantzig selector or the REE can be found, selecting the tuning parameter in an ITR context is challenging since the goal is inference about treatment effects rather than just predictive performance. This means that we cannot simply select the tuning parameter that has the lowest prediction error as in the more classical prediction setting.

Our work proposing new doubly robust estimating functions for count and binary outcomes is motivated by the desire to evaluate the effectiveness of a web-based stress management intervention for individuals with cardiovascular disease. We use longitudinal data collected as part of a two-stage pilot sequential multiple assignment randomised trial (Lambert et al., 2021) for estimating a stress management ITR. Due to the small sample size of the study (50 observations) and relatively large number of potentially relevant variables collected, selecting useful variables for tailoring treatment solely based on expert knowledge can be an extremely challenging task. Our newly proposed estimating equations allow integration of variable selection approaches. We apply

this variable selection approach with our proposed algorithm for solving the proposed estimating equations to provide valuable insights into the influence of various tailoring variables on patient outcomes, enabling the development of more effective and personalised approaches to the stepped-care approach for web-based stress management.

Specifically, in this work, we propose two new, doubly robust estimating functions for count and binary outcomes, respectively, in the setting of binary treatment with a single stage, to estimate an ITR. A benefit of our proposed estimating function is that it can be easily generalised to a penalised framework, which permits estimating the optimal treatment regimes and selecting important tailoring variables simultaneously. We show that with a suitable choice of weights, a simple penalised regression model for estimating an ITR enjoys the desired double robustness property and is straightforward to implement. The advantage of the newly proposed approach compared to alternative regularised ITR estimation methods is that it can be viewed from a minimisation perspective. Hence, the implementation is simple, various penalty functions can be used, and the solution can be found using existing computationally efficient tools in standard software. We propose a tuning parameter selection procedure to address that the goal of an ITR analysis is estimating a decision rule rather than prediction. To our knowledge, doubly robust variable selection in ITR estimation for discrete outcomes has not been studied in existing literature.

The rest of this article is organised as follows. In Section 2, we present introductory concepts and review existing doubly robust estimation methods for discrete outcomes. In Section 3, we introduce our proposed estimation methods, and we extend them to a penalised framework in Section 4, followed by statements of theoretical properties. A number of simulation studies are in Section 5. Finally, in Section 6, we apply our method to data from an adaptive web-based stress management study.

2 Background

2.1 Notations, assumptions, and introductory concepts

Throughout, we use uppercase letters to denote random variables and lowercase letters to denote observed values. We use non-bold letters to denote individual-level data and bold letters to denote all observations in the data, e.g. $X_i \in \mathbb{R}^p$ are the covariates for subject i, while $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ are covariates for all subjects. In a single stage ITR, $V_i = (X_i, A_i, Y_i)$ consists of the data for the ith subject, where X_i is the subject's baseline covariates, A_i is the binary treatment received, and Y_i is the subject's outcome. Throughout, we consider binary treatment in a static setting (single-stage ITR), while extension to general discrete allocations is discussed in Section 7. In the sequel, we will suppress subscript i where it is clear. We denote the potential outcome under the treatment a as Y^a. The objective of an ITR analysis is to find the optimal treatment $d^{\text{opt}}(X)$ such that the expected potential outcome $\mathbb{E}(Y^d)$ is maximised across the population of individuals. To estimate ITRs, we assume the following standard causal assumptions: (a) the stable unit treatment value assumption (SUTVA) (Rubin, 1980): an individual's potential outcome is not affected by other subjects' treatment assignments; (b) consistency: $Y = AY^1 + (1-A)Y^0$; (c) conditional exchangeability (Robins, 1997): $Y^a \perp\!\!\!\perp A|X = x$; and (d) positivity: $P(A = a|X = x) > 0$ almost surely for all x and $a = 0, 1$.

Finally, we assume that the observations V_i, $i = 1, \ldots, n$ are independent and identically distributed with probability density $h(V)$ with respect to a measure ν. Moreover, we assume the relationship between Y and (X, A) can be captured by a semi-parametric regression model: $g(\mathbb{E}(Y^a|X = x)) = g(\mathbb{E}(Y|X = x, A = a)) = f_0(x; \boldsymbol{\beta}) + \gamma(x, a; \boldsymbol{\psi})$, where g is a known link function, f_0 is an unknown baseline function, and γ is a known function that satisfies $\gamma(x, 0; \boldsymbol{\psi}) = 0$, which is referred to as the blip function (Robins, 2004). A blip function can be interpreted as the difference on the linear predictor scale of the transformed mean potential outcomes

$$\begin{aligned}\gamma(x, a) &= g\big(\mathbb{E}(Y^a|X = x)\big) - g\big(\mathbb{E}(Y^0|X = x)\big)\\ &= g\big(\mathbb{E}(Y^a|X = x, A = a)\big) - g\big(\mathbb{E}(Y^0|X = x, A = 0)\big).\end{aligned}$$

In this modelling paradigm, f_0 is irrelevant for making treatment decisions (a nuisance model).

Hence, our parameter of interest is $\boldsymbol{\psi}$, and the optimal ITR $d^{\text{opt}}(x)$ is given by

$$d^{\text{opt}}(x) = \underset{d}{\text{argmax}}\, \mathbb{E}(Y^d) = \underset{d}{\text{argmax}}\, \mathbb{E}_X\left\{\mathbb{E}\left[g^{-1}(f_0(X;\boldsymbol{\beta}) + \gamma\big(X, d(X);\boldsymbol{\psi}\big))\right]|X\right\}$$
$$= \underset{d}{\text{argmax}}\, \mathbb{E}_X\left[f_0(X;\boldsymbol{\beta}) + \gamma\big(X, d(X);\boldsymbol{\psi}\big)\right] = \underset{d}{\text{argmax}}\, \gamma\big(X, d(X);\boldsymbol{\psi}\big) = \mathbb{I}(\gamma\big(X, 1;\boldsymbol{\psi}\big) > 0),$$

given an increasing link function. Throughout, we assume a log link for count outcomes and a logit link for binary outcomes.

2.2 Existing estimation methods for discrete outcomes

2.2.1 A-learning for count outcomes

Denote by x^{ψ} the covariates in the blip model and by x^{β} the covariates in the baseline model; in what follows, the superscript is omitted if they are identical. We assume that the blip function is of the form of $\gamma(x^{\psi}, a;\boldsymbol{\psi}) = a\boldsymbol{\psi}^T x^{\psi}$ in the sequel. Then the A-learning estimating equation (Robins et al., 1992) for a count outcome, with a log link function, is

$$\boldsymbol{U}_1(\boldsymbol{\psi}) = \frac{1}{n}\sum_{i=1}^{n} x_i^{\psi}(a_i - \widehat{\pi}_i)\exp\{-\gamma(x_i^{\psi}, a_i;\boldsymbol{\psi})\}\Big(y_i - \exp\left(f(x_i^{\beta};\widehat{\boldsymbol{\beta}}) + \gamma(x_i^{\psi}, a_i;\boldsymbol{\psi})\right)\Big),$$

where f is the posited baseline model (not necessarily identical to f_0), $\widehat{\boldsymbol{\beta}}$ is a plug-in estimator, and $\widehat{\pi}$ is the estimated propensity score. The propensity score (Rosenbaum & Rubin, 1983) is defined as the coarsest balancing score $b(x)$ such that $b(x) = P(A = 1|x)$, i.e. the probability of treatment received conditional on confounders. In observational studies, this quantity is unknown and needs to be estimated from the data. It can be shown that $\boldsymbol{U}_1(\boldsymbol{\psi})$ is an unbiased estimating equation (Robins et al., 1992), provided that at least one nuisance model (propensity score model or baseline model) is correctly specified. This property is the so-called double robustness property (Bang & Robins, 2005). Since in observational studies, one can never be sure that either a baseline model or a propensity score model is correct, a double robustness estimator hence is highly desirable, as it provides some safeguards against model misspecification. Furthermore, in settings such as our motivating example, where treatment is randomised, doubly robust methods ensure consistency since the treatment allocation model is known by design.

2.2.2 A-learning for binary outcomes

Estimation is more complicated when the outcome is binary; the blip parameter is estimated by solving the following estimating equation, assuming a logit link function:

$$\boldsymbol{U}_2(\boldsymbol{\psi}) = \frac{1}{n}\sum_{i=1}^{n} x_i^{\psi}(a_i - \widehat{\pi}^*)\Big(y_i - \text{expit}(f(x_i^{\beta};\widehat{\boldsymbol{\beta}}) + \gamma(x_i^{\psi}, a_i;\boldsymbol{\psi}))\Big),$$

where

$$\widehat{\pi}^* = \left(1 + \frac{(1 - \text{expit}(u(x;\widehat{\boldsymbol{\tau}}))\text{expit}(f(x;\widehat{\boldsymbol{\beta}}))}{\text{expit}(u(x;\widehat{\boldsymbol{\tau}}))\text{expit}(f(x;\widehat{\boldsymbol{\beta}}) + \gamma(x, a;\boldsymbol{\psi}))}\right)^{-1},$$

$\text{expit}(t) = \frac{\exp(t)}{1+\exp(t)}$, and $u(x;\boldsymbol{\tau})$ is the nuisance treatment model of $\mathbb{E}(A|Y = 0, X)$. Tchetgen Tchetgen et al. (2010) showed that $\boldsymbol{U}_2(\boldsymbol{\psi})$ is an unbiased estimating equation when at least one of $\mathbb{E}(Y|X, A = 0)$ or $\mathbb{E}(A|X, Y = 0)$ is correctly specified. Note that for the logit link, the quantity $\mathbb{E}(A|Y = 0, X)$ is modelled instead of the propensity score to assure the double robustness property, because of the symmetry property of the odds ratio:

$$e^{X^{\top}\boldsymbol{\psi}} = \frac{P(Y = 1|A = 1, X)P(Y = 0|A = 0, X)}{P(Y = 0|A = 1, X)P(Y = 1|A = 0, X)} = \frac{P(A = 1|Y = 1, X)P(A = 0|Y = 0, X)}{P(A = 0|Y = 1, X)P(A = 1|Y = 0, X)}.$$

Chen (2007) showed that there are at least two ways to study the association parameter (in our case, the blip parameter): through the density of Y given X and A or through the density of A given X and Y. This provides an intuitive explanation of why $\mathbb{E}(Y|X, A = 0)$ and $\mathbb{E}(A|X, Y = 0)$ are modelled to assure the double robustness property.

As noted above, the implementation of the Dantzig selector or the REE can be difficult for the A-learning estimating function. In the next section, we propose an alternative estimation method that is also doubly robust and can easily accommodate variable selection.

3 Doubly robust weighted generalised linear model

In this section, we propose two new estimating equations for count and binary outcomes, respectively, and we show that solving these two estimating equations can be reformulated as an iteratively reweighted generalised linear model (IRGLM). The obtained estimators are doubly robust, and the proposed estimating equation can be easily generalised to a variable selection framework. Throughout, we posit a linear model for the baseline function, i.e. $f(x; \boldsymbol{\beta}) = x^T\boldsymbol{\beta}$, which is not necessarily identical to the true baseline model f_0.

3.1 Count outcomes

For count outcomes, we present the following estimating function:

$$\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi}) = \sum_{i=1}^{n} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a; \boldsymbol{\psi})\} \Big(y_i - \exp\left(f(x_i^{\beta}; \boldsymbol{\beta}) + \gamma(x_i^{\psi}, a; \boldsymbol{\psi})\right) \Big).$$

This estimating equation is inspired by the A-learning estimating equation $\boldsymbol{U}_1(\boldsymbol{\psi})$ and the weighted least squares equation using overlap weights $|a_i - \pi_i|$ in Wallace and Moodie (2015). The overlap weights $|a_i - \pi_i|$ ensure that the above estimating equation is unbiased even if the baseline model is mis-specified (under the setting that π is correctly specified). Moreover, Wallace and Moodie (2015) empirically demonstrated that the use of overlap weights can improves efficiency of the resulting estimator over estimators of the same form that use alternative weights such as inverse probability of treatment weights. Note that this equation takes a similar form to $\boldsymbol{U}_1(\boldsymbol{\psi})$, with the leading term $\exp\{-\gamma(x_i^{\psi}, a; \boldsymbol{\psi})\}$, and shares a similar form to Wallace and Moodie (2015) using overlap weights, but is not identical to either.

Assumption 1 When at least one of the two nuisance models π or f is correctly specified, there exists a unique population parameter $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)$ such that $\mathbb{E}[\boldsymbol{U}_3(\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)] = 0$.

Theorem 1 Assume that the SUTVA, ignorability, consistency, and positivity conditions described in Section 2.1 and Assumption 1 hold as described in Section 3.1. If the posited baseline model satisfies $x^{\psi} \subseteq x^{\beta}$, and the link function g is known, then the solution $\boldsymbol{\psi}^*$ to $\mathbb{E}[\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})] = 0$ satisfies $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$, where $\boldsymbol{\psi}_0$ is the underlying true blip parameter.

Theorem 1 states that under standard causal assumptions, the population parameter $\boldsymbol{\psi}^*$ is equivalent to the true data-generating value of the blip (and corresponding ITR) parameter $\boldsymbol{\psi}_0$, if one of two nuisance models, π or f, is correctly specified. This implies that the blip estimator $\widehat{\boldsymbol{\psi}}$ obtained by solving $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$ is a doubly robust estimator.

Remark 1 The condition of the existence of a unique population parameter is similar to the condition of the existence of the quasi-maximum likelihood estimate when the likelihood is mis-specified (White, 1982). The assumption that $x^{\psi} \subseteq x^{\beta}$ in the posited model is referred to as the strong heredity assumption (Chipman, 1996): the corresponding main effects of an interaction term must be included in the model.

Algorithm 1

function $(x_i, a_i, y_i, \widehat{\pi}_i, \varepsilon)$
  Set iteration counter $t \leftarrow 0$
  Initialize $\widetilde{\boldsymbol{\psi}}_0$
  $w_{i0} \leftarrow |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a_i; \widetilde{\boldsymbol{\psi}}_0)\}$ for $i = 1, \ldots, n$
  repeat
    Solve $\boldsymbol{\beta}_t$ and $\boldsymbol{\psi}_t$ such that
    $\sum_{i=1}^{n} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} w_{it}(y_i - \exp(f(x_i^{\beta}; \boldsymbol{\beta}_t) + \gamma(x_i^{\psi}, a_i; \boldsymbol{\psi}_t))) = 0$
    $\widetilde{\boldsymbol{\psi}}_{t+1} \leftarrow \boldsymbol{\psi}_t$
    $w_{i(t+1)} \leftarrow |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a_i; \widetilde{\boldsymbol{\psi}}_{t+1})\}$
    $t \leftarrow t + 1$
  until $\|\boldsymbol{\psi}_t - \boldsymbol{\psi}_{t-1}\| < \varepsilon$

Now we demonstrate that $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$ can be specified as an IRGLM for which efficient computational solutions exist, and thus a penalised estimator can be constructed from the penalised generalised weighted linear model accordingly. We propose Algorithm 1 to solve $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$. The key is to treat the $|a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a; \boldsymbol{\psi})\}$ term in $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$ as a constant in each iteration t. In this way, Step 7 in Algorithm 1 is equivalent to a weighted generalised linear model (GLM) with weights $|a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a; \widetilde{\boldsymbol{\psi}}_t)\}$, where $\widehat{\pi}$ is the estimated propensity score that does not change across iterations and $\widetilde{\boldsymbol{\psi}}$ is the current value of the blip parameter estimate from the most recent iteration update. This can be solved efficiently using, for example, the `glm` function in **R** and specifying the `weights` argument.

3.2 Binary outcomes

A similar framework can be built for binary outcomes using the logit link function. We present estimating equation $\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi})$ for binary outcomes:

$$\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi}) = \sum_{i=1}^{n} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} |a_i - \widehat{\pi}_i^*| \Big(y_i - \text{expit}(f(x_i^{\beta}; \boldsymbol{\beta}) + \gamma(x_i^{\psi}, a; \boldsymbol{\psi})) \Big),$$

Algorithm 2

function $(x_i, a_i, y_i, \widehat{\pi}_i, \varepsilon)$
  Set iteration counter $t \leftarrow 0$
  Initialize: $\widetilde{\boldsymbol{\psi}}_0$
  $w_{i0} \leftarrow |a_i - \widehat{\pi}_i^*(\widetilde{\boldsymbol{\psi}}_0)|$ for $i = 1, \ldots, n$
  repeat
    Solve $\boldsymbol{\beta}_t$ and $\boldsymbol{\psi}_t$ such that
    $\sum_{i=1}^{n} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} w_{it}(y_i - \text{expit}(f(x_i^{\beta}; \boldsymbol{\beta}_t) + \gamma(x_i^{\psi}, a_i; \boldsymbol{\psi}_t))) = 0$
    $\widetilde{\boldsymbol{\psi}}_{t+1} \leftarrow \boldsymbol{\psi}_t$
    $w_{i(t+1)} \leftarrow |a_i - \widehat{\pi}_i^*(\widetilde{\boldsymbol{\psi}}_{t+1})|$
    $t \leftarrow t + 1$
  until $\|\boldsymbol{\psi}_t - \boldsymbol{\psi}_{t-1}\| < \varepsilon$

where

$$\widehat{\pi}^* = \left(1 + \frac{(1 - \text{expit}(u(x; \widehat{\xi}))\text{expit}(f(x; \widehat{\boldsymbol{\beta}}^*))}{\text{expit}(u(x; \widehat{\xi})\text{expit}(f(x; \widehat{\boldsymbol{\beta}}^*) + \gamma(x, 1; \boldsymbol{\psi}))}\right)^{-1},$$

and $u(x; \xi)$ is the nuisance treatment model for $\mathbb{E}(A|Y = 0, X)$. Under mild conditions, the solution of $\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi})$ is a doubly robust estimator. Note that all theoretical properties for count outcomes can be applied equally to binary outcomes; for convenience and space, we include the results for binary outcomes in the online supplementary Appendix (Section A). Algorithm 2 can be used to solve $\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi})$, once again treating the term $|a_i - \widehat{\pi}_i^*|$ as a constant in each iteration.

4 Tailoring variable selection

In this section, we introduce sparsity to our proposed estimating function using the formulation of an REE, and show that this REE is asymptotically equivalent to a penalised weighted GLM given an appropriate initial estimator. Throughout, the main effect of the treatment A is not penalised, as our goal is to select the important tailoring variables.

4.1 Penalised doubly robust method

Due to the non-linear part (log or logit link) of the estimating equation for discrete outcomes, a Dantzig selector with A-learning estimating equation $\boldsymbol{U}_1(\boldsymbol{\psi})$ or $\boldsymbol{U}_2(\boldsymbol{\psi})$ cannot be solved using linear programming (James & Radchenko, 2009). Hence, we pursue an REE approach to introduce sparsity to the proposed estimating equations $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$ and $\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi})$, and once again, reformulate the REE as a penalised weighted GLM. We call this approach the penalised doubly robust (PDR) method, as it will be shown later that the penalised estimator obtained by solving the ITR REE is a doubly robust estimator.

For count and binary outcomes, ITR REE requires finding the solution of, respectively,

$$\sum_{i=1} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a; \boldsymbol{\psi})\} \Big(y_i - \exp\left(f(x_i^{\beta}; \boldsymbol{\beta}) + \gamma(x_i^{\psi}, a; \boldsymbol{\psi})\right)\Big) = n\lambda q(|\boldsymbol{\theta}|), \tag{1}$$

and

$$\sum_{i=1} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} |a_i - \widehat{\pi}_i^*| \Big(y_i - \text{expit}(f(x_i^{\beta}; \boldsymbol{\beta}) + \gamma(x_i^{\psi}, a; \boldsymbol{\psi}))\Big) = n\lambda q(|\boldsymbol{\theta}|). \tag{2}$$

To estimate the blip parameters, $\boldsymbol{\psi}$, consistently, we require that the penalised model satisfies the following properties: (a) no false exclusion of tailoring variables and (b) the selected model has the strong heredity property, i.e. $\widehat{\psi}_j \neq 0 \Longrightarrow \widehat{\beta}_j \neq 0$ (i.e. without loss of generality, assume that x^{ψ} has the same 'ordering' as x^{β}). Many penalty functions can yield a model that has variable selection consistency, i.e. no false inclusion and no false exclusion; for example, lasso, SCAD (Fan & Li, 2001), and adaptive lasso (Zou, 2006). However, these methods all fail to achieve the strong heredity property. Thus, further work is required to implement them in this setting. Bian et al. (2023) used reparametrisation to ensure strong heredity when using penalisation in the context of ITR. Here, we modify the adaptive lasso penalty and show using these modified adaptive weights allows not only the strong heredity constraint to be met, but also the (asymptotically) unbiased estimation of blip parameters.

We omit the subscript for the estimating functions $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$ and $\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi})$ for now, as the properties for both count and binary outcomes can be developed using a general notation $\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\psi})$. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\psi}_0)$ denote the underlying true parameters and recall that $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)$ is the unique population parameter such that $\mathbb{E}[\boldsymbol{U}(\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)] = 0$. Let s be the number of non-zero components of $\boldsymbol{\psi}_0$ (or equivalently, $\boldsymbol{\psi}^*$) and S denote the set of indices of non-zero components for $\boldsymbol{\psi}_0$. Denote by *Sifinmath* the set of indices of non-zero components for $\boldsymbol{\beta}^*$. To satisfy the strong

heredity property, we want the estimated baseline model to satisfy $\widehat{\boldsymbol{\beta}}_{\widetilde{S}} \neq 0$ as n goes to infinity, where $\widetilde{S} = S \cup Sifinmath$ (as such, $S \subseteq \widetilde{S}$ and hence strong heredity holds). The goal is to estimate a targeted indices set S^*, such that $\widehat{\boldsymbol{\theta}}_{S^*} \neq 0$ and $\widehat{\boldsymbol{\theta}}_{S_c^*} = 0$ with probability tending to 1, where S_c^* is the complement of S^* (note that $\boldsymbol{\theta}^*_{S^*} = (\boldsymbol{\beta}^*_{\widetilde{S}}, \boldsymbol{\psi}^*_{\widetilde{S}})$).

Suppose we have an initial estimator $\widehat{\boldsymbol{\theta}}_{ini} = (\widehat{\boldsymbol{\beta}}_{ini}, \widehat{\boldsymbol{\psi}}_{\text{ini}})$, such that $\sqrt{n}\|\widehat{\boldsymbol{\beta}}_{ini} - \boldsymbol{\beta}^*\| = O_p(1)$ and $\sqrt{n}\|\widehat{\boldsymbol{\psi}}_{\text{ini}} - \boldsymbol{\psi}^*\| = O_p(1)$. Following the adaptive lasso (Zou, 2006) principle, we construct our adaptive weights for the corresponding coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ as follows:

$$\widehat{\omega}_j^{\beta} = \left\{\max\left(|\widehat{\beta}_j^{ini}|, |\widehat{\psi}_j^{ini}|\right)\right\}^{-1} \text{ and } \widehat{\omega}_j^{\psi} = |\widehat{\psi}_j^{ini}|^{-1}. \tag{3}$$

We then use the penalty function $\rho(|\boldsymbol{\theta}|) = \rho(|\boldsymbol{\beta}|) + \rho(|\boldsymbol{\psi}|)$, where

$$\rho(|\boldsymbol{\beta}|) = \sum_{j=1}^{p} \widehat{\omega}_j^{\beta}|\beta_j| \quad \text{and} \quad \rho(|\boldsymbol{\psi}|) = \sum_{j=1}^{p} \widehat{\omega}_j^{\psi}|\psi_j|.$$

In this way, for non-zero coefficients of blip variables, the associated weights and those of their corresponding main effects both converge to finite constants, and thus always remain in the model. We refer to our proposed weights in Expression (3) as modified adaptive weights, since these build on the adaptive lasso framework but differ in the choice of $\widehat{\omega}_j^{\beta}$. Theorem 2 establishes the existence of a $\sqrt{n}$-consistent solution to the ITR REE (1) and (2).

Theorem 2 (Existence and Selection Consistency). Assume that conditions in Theorem 1 hold, penalty functions are constructed using the modified adaptive weights described in Expression (3), and the tuning parameter satisfies $\sqrt{n}\lambda = o(1)$ and $n\lambda \to \infty$. There then exists a $\sqrt{n}$-consistent solution $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}})$ of the ITR REE, such that $\widehat{\boldsymbol{\psi}}_S \neq 0$ and $\widehat{\boldsymbol{\psi}}_{S_c} = 0$.

By Lemma 1 in the online supplementary Appendix B.2, to establish the existence of the REE solution, it suffices to show that for sufficiently large n, there exists a constant r, such that on the boundary of a ball around $\boldsymbol{\theta}^*$ with radius $n^{-1/2}r$, the variational inequality holds for function $\boldsymbol{U}(\boldsymbol{\theta}) - n\lambda q(|\boldsymbol{\theta}|)$ with high probability. That is, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|=n^{-1/2}r} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T[\boldsymbol{U}(\boldsymbol{\theta}) - n\lambda q(|\boldsymbol{\theta}|)] > 0\right) > 1 - \varepsilon.$$

This technique has been adopted in Portnoy (1984) and Wang (2011) to prove the existence of the M-estimator and generalised estimated equations estimator when the number of predictors is large. Theorem 3 establishes the asymptotic normality of the ITR REE estimators under standard regularity conditions (see online supplementary Appendix B for details).

Theorem 3 (Asymptotic Normality). For any $\sqrt{n}$-consistent solution $\widehat{\boldsymbol{\theta}}$ of ITR REE,

$$\sqrt{n}\boldsymbol{J}(\boldsymbol{\psi}^*_S)\{\widehat{\boldsymbol{\psi}}_S - \boldsymbol{\psi}^*_S + \boldsymbol{J}(\boldsymbol{\psi}^*_S)^{-1}\lambda q(|\boldsymbol{\psi}^*_S|)\} \to_d \boldsymbol{N}\big(\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\psi}^*_S)\big),$$

where $\boldsymbol{I}(\boldsymbol{\theta}) \in \mathbb{R}^{2p\times 2p}$ is the variance of the estimating equation $\boldsymbol{U}(V_i, \boldsymbol{\theta})$, $\boldsymbol{J}(\boldsymbol{\theta}) \in \mathbb{R}^{2p\times 2p}$ is the quantity $\mathbb{E}_{\boldsymbol{\theta}}[-\frac{\partial \boldsymbol{U}(V_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}]$, p is the length of the full covariate vector X_i, and $\boldsymbol{I}(\boldsymbol{\psi}^*_S)$ and $\boldsymbol{J}(\boldsymbol{\psi}^*_S)$ are the corresponding $s \times s$ sub-matrices of $\boldsymbol{I}$ and $\boldsymbol{J}$ evaluated at the truth.

A detailed proof of Theorems 2 and 3 are in the online supplementary Appendix (Sections B.4 and B.5). To illustrate the double robustness property of our proposed estimators, we borrow the idea of the oracle estimator (Fan & Li, 2001). Define the oracle estimator $\widehat{\boldsymbol{\psi}}_{\text{ora}} \in \mathbb{R}^s$ as the solution of $\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\psi})$ using $f(x_{\widetilde{S}})$ and $\gamma(x_S, a)$ (i.e. assume that the zero and non-zero coefficients are known in advance). Since we do not know the truly important variables in the application, the oracle

estimator is just a concept to help establish the theoretical properties in variable selection. Due to the double robustness of $\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\psi})$, $\widehat{\boldsymbol{\psi}}_{\text{ora}}$ is a consistent asymptotically normal estimator of $\boldsymbol{\psi}_S^*$ under standard regularity conditions for M-estimators. The properties of $\widehat{\boldsymbol{\psi}}$ in Theorems 2 and 3 are referred to as the oracle property (Fan & Li, 2001), i.e. $\widehat{\boldsymbol{\psi}}$ performs as well as the oracle estimator $\widehat{\boldsymbol{\psi}}_{\text{ora}}$.

Corollary (Double Robustness). The oracle estimator $\widehat{\boldsymbol{\psi}}_{\text{ora}}$ constructed above is a doubly robust estimator of $\boldsymbol{\psi}_0$. Since the resulting estimator $\widehat{\boldsymbol{\psi}}$ mimics the oracle estimator $\widehat{\boldsymbol{\psi}}_{\text{ora}}$, $\widehat{\boldsymbol{\psi}}$ is also a doubly robust estimator. That is to say, the resulting estimator $\widehat{\boldsymbol{\psi}}$ is a consistent estimator of $\boldsymbol{\psi}_0$ if either of two nuisance models is correct.

4.2 A one-step estimator

For settings in which the number of variables, p, is fixed, we present an approximation to solve the ITR REE (1) in one step. Suppose that we can find an initial estimator $\widehat{\boldsymbol{\psi}}_{\text{ini}}$ of the blip parameter, such that $\sqrt{n}\|\widehat{\boldsymbol{\psi}}_{\text{ini}} - \boldsymbol{\psi}^*\|_2 = O_p(1)$. Then we can plug $\widehat{\boldsymbol{\psi}}_{\text{ini}}$ into the weight term of Expression (1) and solve it directly, which is equivalent to maximising a weighted penalised likelihood. Taking the count outcomes as an example, we can use the solution of the unpenalised estimating equation $\boldsymbol{U}_1(\boldsymbol{\theta})$ or $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\theta})$ as the initial estimator. Then under mild conditions, using $\widehat{\boldsymbol{\psi}}_{\text{ini}}$ as a plug-in estimator will have a negligible effect on the resulting estimator $\widehat{\boldsymbol{\psi}}$. That is, the solution of

$$\sum_{i=1}\begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a_i; \widehat{\boldsymbol{\psi}}_{\text{ini}})\}\Big(y_i - \exp\big(f(x_i^{\beta}; \boldsymbol{\beta}) + \gamma(x_i^{\psi}, a_i; \boldsymbol{\psi})\big)\Big) = n\lambda\partial\rho(|\boldsymbol{\theta}|)$$

is asymptotically equivalent to the solution of equation (1). In high-dimensional settings in which an unpenalised initial estimator cannot easily be computed, the ridge penalty can be used to obtain the initial estimator.

4.3 Tuning parameter selection

The choice of the tuning parameter λ in Expressions (1) and (2) plays an important role in the performance of the REE: An inappropriately large or small value of λ will greatly weaken the performance of the resulting estimator in generating the estimation error and variable selection results. As previously noted, our proposed method can be viewed from a minimisation perspective, i.e. $\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}}\{\mathcal{L}_n(\boldsymbol{\theta}; \boldsymbol{y}) + n\lambda\rho(|\boldsymbol{\theta}|)\}$. Following the idea used in classical information criteria (Akaike, 1974; Schwarz, 1978), we propose to select the tuning parameter by choosing the model that has the smallest value of $n^{-1}[D_\lambda(\widehat{\boldsymbol{\theta}}, \boldsymbol{y}) + \kappa_n s_\lambda]$, where $D_\lambda(\widehat{\boldsymbol{\theta}}, \boldsymbol{y}) = 2[\mathcal{L}_n^{\text{sat}}(\widehat{\boldsymbol{\theta}}; \boldsymbol{y}) - \mathcal{L}_n(\widehat{\boldsymbol{\theta}}; \boldsymbol{y})]$ is the quasi-deviance, $\mathcal{L}_n^{\text{sat}}$ is the quasi-log-likelihood of the saturated model, κ_n is some positive sequence, and s_λ is the number of non-zero components in the model for a given λ. We suggest setting κ_n as $\log(\log n)\log p$ following Fan and Tang (2013), as this can achieve model selection consistency in a penalised likelihood setting. In practice, we could also use cross-validation to choose the tuning parameter that corresponds to the lowest average loss $\mathcal{L}_n^{\text{cv}}(\widehat{\boldsymbol{\theta}}; \boldsymbol{y})$.

In a penalised likelihood, where the goal is prediction, the optimal λ is often chosen so the corresponding model has the lowest information criterion, usually estimated by a measure of model fit (e.g. negative log-likelihood) with an extra penalty term such as the Akaike information criterion (Akaike, 1974) or the Bayesian information criterion (Schwarz, 1978). However, using an Akaike or Bayesian information criterion to select the tuning parameter will fail if the likelihood is misspecified (i.e. outcome model is mis-specified). Thus, these classic methods of tuning parameter selection are not appropriate to the doubly robust setting where a likelihood is not positive and the mean model is not assumed to be correctly specified. Our proposed approach to selecting the tuning parameter outlined above requires that only one of the nuisance models is correctly specified.

5 Numerical studies

In this section, we illustrate the double robustness of our proposed method and show how the choice of the initial estimator can impact the resulting estimators.

5.1 Competing methods and implementation

We compare our proposed method with three different methods: unpenalised A-learning, Zhang and Zhang (2018) and Zhang and Zhang (2022), where the last two competing methods were established based on the binary classification framework proposed in Zhang et al. (2012). The **R** package `drgee` (Zetterqvist & Sjölander, 2015) is implemented to obtain the A-learning estimates; in addition, the sample code to conduct methods in Zhang and Zhang (2018) and Zhang and Zhang (2022) can be found in the supplementary material for the latter article.

Recall that in Section 4.2, the initial estimator can be obtained from A-learning or our proposed IRGLM. We now evaluate the performance of our proposed PDR method using two different initial estimators for the variable selection rate and the resulting error rate in the estimated treatment decision, as well as for the value function (expected outcome) of the estimated decision rules. The error rates and the average value function were calculated over a testing set of size 10,000. The data generation procedure for count outcomes is

- Step 1: Generate 15 independent multivariate normal covariates $(X_1, \ldots, X_{15})$ with mean equal to 0.5 and unit variance.
- Step 2: Generate treatment such that $P(A = 1|x_1, x_2) = \text{expit}(-0.2 + \sum_{j=1}^{2} x_j)$.
- Step 3: Set the blip function as $\gamma(x, a; \boldsymbol{\psi}) = a(\psi_0 + \psi_1 x_1)$ for $\psi_0 = 1$ and $\psi_1 = -2$.
- Step 4: Set the baseline model to $f(\boldsymbol{x}; \boldsymbol{\beta}) = \exp(-x_1^2 - x_2^2 + x_3 - x_4) + x_1 - 0.2x_2$.
- Step 5: Generate the outcome $Y \sim \text{Poisson}(\exp(f(\boldsymbol{x}; \boldsymbol{\beta}) + \gamma(\boldsymbol{x}, a; \boldsymbol{\psi})))$.

Under this data generation procedure, the optimal treatment is $\mathbb{I}(1 - 2x_1 > 0)$, which corresponds to treatment $A = 1$ for about 50% of subjects, and the marginal mean of the outcome under observed (rather than optimal) treatment is 1.21.

The data generation procedure for binary outcomes is the same for Steps 1–3 above. In Step 4, we now set the nuisance treatment model as $\mathbb{E}(A|Y = 0, X = x) = \exp(-x_1^2 - x_2^2 + x_3 - x_4) + x_1 - 0.2x_2$, and marginalise the conditional expectation over the distribution of Y to obtain the propensity score model $\mathbb{E}(A|X = x)$. Lastly, we generate the outcome $Y \sim \text{Bernoulli}(\text{expit}(f(\boldsymbol{x}; \boldsymbol{\beta}) + \gamma(\boldsymbol{x}, a; \boldsymbol{\psi})))$. Under this data generation procedure, the optimal treatment corresponds to treatment $A = 1$ for about 50% of subjects, and the marginal mean of the outcome under observed (rather than optimal) treatment is 0.47.

For both count outcomes and binary outcomes, we consider two scenarios with two sample sizes (500 and 1,000). The baseline model is mis-specified in scenario 1 (a linear working model is used), and the treatment model is mis-specified in scenario 2 (the propensity score is setting to 0.5 for all the observations). For PDR, we consider two alternative initial estimators: In the first case, referred to as PDR1, the estimator is obtained from A-learning, and in the second, PDR2, from our proposed IRGLM approach. Finally, we refer to unpenalised A-learning and the methods in Zhang and Zhang (2018) and Zhang and Zhang (2022) as UA, ZZ1, and ZZ2, respectively.

Tables 1 and 2 present the error rate (proportion of times the estimated optimal ITR fails to coincide with the true optimal ITR); value; false-negative rate (i.e. setting a tailoring variable's coefficient to 0 when it should be non-zero); false-positive rate (i.e. selecting a tailoring variable when the coefficient should in fact be zero) of the blip parameter estimates of the three methods for binary and count outcomes, respectively. In summary, all four methods have similar and good performance; this is expected as they are all doubly robust methods. For count outcomes, in scenario 1 with sample size 500, ZZ2 has the smallest error rate and the largest value; as the sample size increases to 1,000, our proposed PDR1 outperforms other methods with respect to error rate and the value. As for scenario 2, our proposed PDR1 and PDR2 outperform all other competing methods regardless of the sample size. For example, when $n = 1{,}000$, PDR2 has the smallest error rate as well as the largest value; moreover, the FP and FN are both 0. The results of methods evaluated here for binary outcomes generally exhibit similarities to those for count outcomes.

Table 1. Error rate (ER), value, false-negative (FN), and false-positive (FP) rate of variable selection results, with $n = 500$ and 1,000, for 400 simulations and test size 10,000 in three scenarios for a count outcome

	Scenario 1					Scenario 2				
	UA	ZZ1	ZZ2	PDR1	PDR2	UA	ZZ1	ZZ2	PDR1	PDR2
$n = 500$										
ER	0.13	0.07	0.06	0.07	0.08	0.09	0.05	0.06	0.03	0.03
Value	3.28	3.34	3.35	3.34	3.33	3.33	3.35	3.35	3.36	3.36
FN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FP	1.00	0.03	0.17	0.16	0.19	1.00	0.00	0.22	0.04	0.01
$n = 1{,}000$										
ER	0.09	0.06	0.05	0.04	0.04	0.06	0.04	0.05	0.03	0.03
Value	3.33	3.35	3.35	3.36	3.35	3.35	3.36	3.36	3.36	3.36
FN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FP	1.00	0.00	0.16	0.07	0.08	1.00	0.00	0.24	0.01	0.00

Note. For comparison, the value function of the true optimal regime, and the strategies of always treat and never treat are 3.36, 1.82, and 2.08, respectively. PDR = penalised doubly robust; UA = unpenalised A-learning; ZZ1 = Zhang and Zhang (2018); ZZ2 = Zhang and Zhang (2022).

Table 2. Error rate (ER), value, false-negative (FN), and false-positive (FP) rate of variable selection results, with $n = 500$ and 1,000, for 400 simulations and a test size 10,000 in three scenarios for a binary outcome

	Scenario 1					Scenario 2				
	UA	ZZ1	ZZ2	PDR1	PDR2	UA	ZZ1	ZZ2	PDR1	PDR2
$n = 500$										
ER	0.18	0.08	0.08	0.07	0.07	0.18	0.07	0.07	0.07	0.08
Value	0.61	0.64	0.63	0.64	0.64	0.61	0.63	0.64	0.64	0.64
FN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FP	1.00	0.00	0.21	0.05	0.05	1.00	0.00	0.21	0.04	0.08
$n = 1{,}000$										
ER	0.13	0.07	0.07	0.05	0.05	0.13	0.06	0.05	0.04	0.05
Value	0.63	0.64	0.64	0.64	0.64	0.63	0.63	0.64	0.64	0.64
FN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FP	1.00	0.00	0.24	0.03	0.02	1.00	0.00	0.25	0.03	0.05

Note. For comparison, the value function of the true optimal regime, and the strategies of always treat and never treat are 0.64, 0.48, and 0.48, respectively. PDR = penalised doubly robust; UA = unpenalised A-learning; ZZ1 = Zhang and Zhang (2018); ZZ2 = Zhang and Zhang (2022).

We make some final remarks on the simulation results here. First, no obvious difference in the error rate, value, and variable selection performance were observed between PDR1 and PDR2 in the simulations. Second, the penalisation-based methods (PDRs and ZZ2) have a larger FP rate than the sequentially selection-based method ZZ1 in general. Specifically, ZZ1 has the best variable selection performance: for example, it achieves 0 FP rate for binary outcomes in both scenarios. Our proposed PDR approach has a slightly higher FP rate than ZZ1, however, it still can yield a larger value and a smaller error rate than ZZ1 in many settings. Moreover, our PDR approach has a much smaller FP rate than ZZ2, although they both are based on the ℓ_1 penalty, PDR takes advantage of using the data-dependant adaptive weights and hence achieve a better variable selection performance than ZZ2.

In this section, we have focused exclusively on settings where assumptions are met. For a demonstration of the impact of violations of the assumption of correct specification of the blip model function, please see the online supplementary Appendix C (Table C1). As anticipated, performance deteriorates significantly when this key assumption is not met.

6 Application to an adaptive web-based stress management study

We illustrate the newly proposed approach on a dataset from a two-stage pilot of a sequential multiple assignment randomised trial (Lambert et al., 2021). The trial aimed to assess a web-based, stress management intervention adapted across time using a stepped-care approach for people with cardiovascular disease. We focus our analysis on the first stage only, in which 50 participants were randomised into two treatment groups, each with probability 0.5, stratified by recruitment source and stress level. The two treatment groups were: website only ($A = 0$) and website plus weekly telephone coaching ($A = 1$).

The primary outcome in this analysis is the stress subscale from the Depression Anxiety Stress Scales (DASS) (Lovibond & Lovibond, 1996), which is a count outcome measured at 6 weeks after stage 1 randomised allocation. A lower DASS-stress subscale score suggests the presence of fewer symptoms of stress, so the optimal treatment decision minimises the DASS-stress subscale score. The aims of our analysis were to determine the tailoring variables related to the decision rule and to obtain the estimated ITR for individuals with cardiovascular disease. We restricted our analysis to eight variables: mental component score, age, DASS-stress subscale score at baseline, sex, marital status, stomach condition, physical component score, and vision. These were previously found to be useful for tailoring treatment using Bian et al. (2023).

A logistic regression model was posited to estimate the propensity score adjusted for the recruitment source and stress level. We applied PDR to this study with A-learning as the initial estimator (referred to as PDR1 in Section 5); both the baseline model and the blip model are posited to be linear. We found that five variables were relevant for tailoring treatment: DASS at baseline, sex, marital status, stomach condition, and vision. The estimated treatment rule is

$$\begin{aligned}\widehat{a}^{\text{opt}} = \mathbb{I}\{ &- 0.78 + 0.09\mathbb{I}(\text{male}) + 0.45\mathbb{I}(\text{unmarried}) + 0.01\text{DASS} \\ &+ 0.45\mathbb{I}(\text{stomach=yes}) - 0.08\mathbb{I}(\text{vision=yes}) < 0\}.\end{aligned}$$

For example, a married woman who does not have either a vision problem or a stomach ailment and who has a DASS greater than 13 would be recommended for website plus weekly telephone coaching ($A = 1$). We compared our estimated treatment rule with results using the approach in Bian et al. (2023), treating the DASS as a continuous measure. We found that 74% of subjects' recommended treatments were the same under the two strategies. Moreover, all five non-zero, estimated blip parameters had the same sign as the estimated blip parameters using Bian et al. (2023).

We also considered, for illustrative purposes, an analysis that dichotomises the outcome Y at its median, using our proposed binary outcome approach. However, due to the small sample size, neither A-learning nor standard logistic regression yielded a solution, due to lack of convergence.

Finally, we applied our newly proposed approach to data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study (Fava et al., 2003). The STAR*D data are considered a benchmark dataset for ITR analyses and were analysed in Chakraborty et al. (2013), Shi et al. (2018), Wallace et al. (2019), and Bian et al. (2023), among others. While these data are less novel, we considered the comparison relevant and provide results in the online supplementary Appendix D. In summary, the findings in the current analysis, using the methods we propose for both count and binary outcomes, align well with the results found in Chakraborty et al. (2013), Wallace et al. (2019), and Bian et al. (2023).

7 Discussion

We proposed new, doubly robust estimating functions to estimate an ITR when the outcome is discrete and the log or logit link functions are used to model the outcome. The newly proposed approach can be solved using a weighted GLM iteratively, given a suitable choice of observational weights. The benefit of our proposed estimating function is that it is easily generalised to a

penalised framework, which permits estimating a parsimonious ITR and selecting important tailoring variables simultaneously. Based on this finding, we also present a doubly robust criterion to select the tuning parameter. Numerical studies indicated that the newly proposed PDR method compares favourably with other competing approaches in the context of ITRs. To our knowledge, doubly robust variable selection approach for ITRs with binary or count outcomes has not previously been studied.

We applied our proposed variable selection method to a sequential multiple assignment randomised trial (Lambert et al., 2021) to evaluate the effectiveness of a web-based stress management intervention for individuals with cardiovascular disease. We found that five variables were relevant for tailoring treatment: DASS at baseline, sex, marital status, stomach condition, and vision. Furthermore, we derived a linear decision rule that may assist physicians in effectively recommending the web-based stress management intervention for patients with cardiovascular disease. This analysis yielded important insights into the influence (or lack thereof) of potential tailoring variables on patient primary outcomes, thus aiding to develop more effective and personalised approaches to care.

One limitation of our proposed method is that we require that the parametric form of the blip function is known (i.e. that the blip is correctly specified). This requirement is slightly stronger than the assumption that the parametric form of the treatment regimes is correctly specified (see, e.g. Zhang & Zhang, 2022), since assuming that the blip function is correct implies that the treatment regime is correct, but not the converse. An interesting avenue for future research would be to consider imposing smoothness assumptions on the blip function and estimating it using off-the-shelf non-parametric variable selection tools, for instance using splines with a penalty to control overfitting.

In this paper, for simplicity, we focus on a binary treatment setting. The extension to general discrete allocations, in which $a = \{0, 1, \ldots, l\}$, is straightforward: a multinomial model analogous to the generalised propensity score could be fit in place of π. Taking the outcomes to be counts, for example, the estimating function now is

$$\sum_{a\neq 0}\sum_{i=1}^{n}\begin{pmatrix}\mathbb{I}(A_i=a)x_i^{\psi}\\ x_i^{\beta}\end{pmatrix}|\mathbb{I}(A_i=a)-\mathbb{P}(A_i=a)|\exp\{-\gamma(x_i^{\psi},a;\boldsymbol{\psi})\}$$
$$\times\left(y_i-\exp\left(f(x_i^{\beta};\boldsymbol{\beta})+\gamma(x_i^{\psi},a;\boldsymbol{\psi})\right)\right)=0.$$

As such, the estimation procedure and the theoretical results can be adapted without extra difficulty. Similarly, continuous exposure densities can be modelled directly, or approximated using quantile binning and modelled via a multinomial regression. Both of these approaches rely on a generalised propensity score (Imbens, 2000) and were implemented in a continuous outcome setting for individualised treatment by Schulz and Moodie (2021).

To obtain a doubly robust estimator, a well-behaved initial estimator is needed, which can be estimated using an unpenalised doubly robust approach. When the number of predictors is larger than the sample size, we recommend using the ridge estimator to acquire the initial estimate. In future work, we could also build on an idea in Huang et al. (2008), which used the marginal regression approach to obtain the initial estimator for the adaptive lasso (i.e. the outcome is regressed separately on each variable). However, this technique is more challenging in our setting, as it violates the assumption that the blip model is correctly specified. This is partial identification problem has been studied in van der Laan and Robins (2003), and this work may shed light on how to use marginal regression to obtain a valid initial estimator. It also may be of interest, in future work, to investigate the algorithm to directly solve the REE instead of using the approximation. As this alternative does not require an initial estimator, and it might perform better in a large p, small n scenario.

The extension of the single-stage estimation approach to a multistage setting also requires further investigation. In a multistage setting, the estimation procedure is conducted recursively using backward induction, and the 'outcome' at each stage is set to be a predicted or estimated optimal response. For discrete outcomes, the optimal outcome is usually modelled by multiplicative effects, e.g. the optimal outcome at the $(k-1)$th stage for a count outcome is computed by

$\widehat{y}_{k-1}^{\text{opt}} = y \times \prod_k^K \exp\{\gamma_k(x_k^{\psi}, \widehat{a}_k^{\text{opt}}; \boldsymbol{\psi}_k) - \gamma_k(x_k^{\psi}, a_k; \boldsymbol{\psi}_k)\}$, where K is the total number of stages. A challenge under the multistage scenario is that the estimated optimal outcome at any stage for subjects with zero-valued outcome will always remain zero, unless adjustments are made (Wallace et al., 2019), which may lead to a loss of efficiency.

Conflict of interest: S.M.S. has worked on grants awarded to Kaiser Permanente Washington Health Research Institute (KPWHRI) by Bristol Meyers Squibb and by Pfizer. She was also a co-investigator on grants awarded to KPWHRI from Syneos Health, which represented a consortium of pharmaceutical companies carrying out U.S. Food and Drug Administration-mandated studies on the safety of extended-release opioids.

Funding

The research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number R01 MH114873 (co-PIs S.M.S. and E.E.M.M.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. E.E.M.M. is a Canada Research Chair (Tier 1) in Statistical Methods for Precision Medicine and acknowledges the support of a chercheur de mérite career award from the Fonds de Recherche du Québec, Santé. S.B. acknowledges funding via a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, RGPIN-2020-05133. The adaptive web-based stress management study was funded by the Canadian Institutes of Health Research, and S.D.L. is a Canada Research Chair (Tier 2) in self-management interventions.

Data availability

The Stress Management data that support the findings of this study in the main paper are available from S.D.L., upon reasonable request. The STAR*D data that support the findings in the online supplementary Appendix are available in the National Institute of Mental Health Data Archive at https://nda.nih.gov/, collection ID number 2148.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

References

Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Bang H., & Robins J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Bian Z., Moodie E. E., Shortreed S. M., & Bhatnagar S. (2023). Variable selection in regression-based estimation of dynamic treatment regimes. *Biometrics*, *79*(2), 988–999. https://doi.org/10.1111/biom.v79.2

Candes E., & Tao T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, *35*(6), 2313–2351. https://dx.doi.org/10.1214/009053607000000532

Chakraborty B., Laber E. B., & Zhao Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, *69*(3), 714–723. https://doi.org/10.1111/biom.v69.3

Chakraborty B., & Moodie E. E. M. (2013). *Statistical methods for dynamic treatment regimes*. Springer.

Chen H. (2007). A semiparametric odds ratio model for measuring association. *Biometrics*, *63*(2), 413–421. https://doi.org/10.1111/biom.2007.63.issue-2

Chen S., Tian L., Cai T., & Yu M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, *73*(4), 1199–1209. https://doi.org/10.1111/biom.v73.4

Chipman H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, *24*(1), 17–36. https://doi.org/10.2307/3315687

Fan J., & Li R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Fan Y., & Tang C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *75*(3), 531–552. https://doi.org/10.1111/rssb.12001

Fava M., Rush A. J., Trivedi M. H., Nierenberg A. A., Thase M. E., Sackeim H. A., Quitkin F. M., Wisniewski S., Lavori P. W., Rosenbaum J. F., & Kupfer D. J. (2003). Background and rationale for the sequenced treatment alternatives to relieve depression (STAR* D) study. *Psychiatric Clinics of North America*, *26*(6), 457–494. https://doi.org/10.1016/S0193-953X(02)00107-7

Huang, J., Ma, S., & Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, *18*, 1603–1618. https://www.jstor.org/stable/24308572

Imbens G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710. https://doi.org/10.1093/biomet/87.3.706

James G. M., & Radchenko P. (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika*, *96*(2), 323–337. https://doi.org/10.1093/biomet/asp013

Jeng X. J., Lu W., & Peng H. (2018). High-dimensional inference for personalized treatment decision. *Electronic Journal of Statistics*, *12*(1), 2074. https://doi.org/10.1214/18-EJS1439

Johnson B. A., Lin D., & Zeng D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, *103*(482), 672–680. https://doi.org/10.1198/016214508000000184

Kosorok M. R., & Moodie E. E. M. (2015). *Adaptive treatment strategies in practice: Planning trials and analyzing data for personalized medicine*. Society for Industrial and Applied Mathematics.

Lambert, S. D., Grover, S., Laizner, A. M., McCusker, J., Belzile, E., Moodie, E. E., Kayser, J. W., Lowensteyn, I., Vallis, M., & Walker, M. (2021). Adaptive web-based stress management programs among adults with a cardiovascular disease: A pilot sequential multiple assignment randomized trial (SMART). *Patient Education and Counseling*, *104*, 1608–1635. https://doi.org/10.1016/j.pec.2021.01.023

Linn K. A., Laber E. B., & Stefanski L. A. (2017). Interactive Q-learning for quantiles. *Journal of the American Statistical Association*, *112*(518), 638–649. https://doi.org/10.1080/01621459.2016.1155993

Logan B. R., Sparapani R., McCulloch R. E., & Laud P. W. (2019). Decision making and uncertainty quantification for individualized treatments using Bayesian additive regression trees. *Statistical Methods in Medical Research*, *28*(4), 1079–1093. https://doi.org/10.1177/0962280217746191

Lovibond S. H., & Lovibond P. F. (1996). *Manual for the depression anxiety stress scales*. Psychology Foundation of Australia.

Lu W., Zhang H. H., & Zeng D. (2013). Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, *22*(5), 493–504. https://doi.org/10.1177/0962280211428383

Murphy S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(2), 331–355. https://doi.org/10.1111/1467-9868.00389

Portnoy, S. (1984). Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large. I. Consistency. *Annals of Statistics*, *12*, 1298–1309. https://doi.org/10.1214/aos/1176346793

Robins J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent variable modeling and applications to causality*. Lecture Notes in Statistics (pp. 69–117). Springer.

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

Robins J. M., Mark S. D., & Newey W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, *48*(2), 479–495. https://doi.org/10.2307/2532304

Rosenbaum P. R., & Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, *75*(371), 591–593. https://doi.org/10.2307/2287653

Schulz J., & Moodie E. E. M. (2021). Doubly robust estimation of optimal dosing strategies. *Journal of the American Statistical Association*, *116*(533), 256–268. https://doi.org/10.1080/01621459.2020.1753521

Schwarz G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shi C., Fan A., Song R., & Lu W. (2018). High-dimensional A-learning for optimal dynamic treatment regimes. *The Annals of Statistics*, *46*(3), 925. https://doi.org/10.1214/17-AOS1570

Tchetgen Tchetgen E. J., Robins J. M., & Rotnitzky A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, *97*(1), 171–180. https://doi.org/10.1093/biomet/asp062

Tian L., Alizadeh A. A., Gentles A. J., & Tibshirani R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532. https://doi.org/10.1080/01621459.2014.951443

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288. https://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x

van der Laan M. J., & Robins J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

Wallace M. P., & Moodie E. E. M. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, *71*(3), 636–644. https://doi.org/10.1111/biom.v71.3

Wallace M. P., Moodie E. E. M., & Stephens D. A. (2019). Model selection for G-estimation of dynamic treatment regimes. *Biometrics*, *75*(4), 1205–1215. https://doi.org/10.1111/biom.v75.4

Wang L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *Annals of Statistics*, *39*(1), 389–417. https://doi.org/10.1214/10-AOS846

Wang L., Zhou J., & Qu A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, *68*(2), 353–360. https://doi.org/10.1111/biom.2012.68.issue-2

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, *51*, 1–25. https://doi.org/10.2307/1912004

Zetterqvist J., & Sjölander A. (2015). Doubly robust estimation with the R package drgee. *Epidemiologic Methods*, *4*(1), 69–86. https://doi.org/10.1515/em-2014-0021

Zhang B., Tsiatis A. A., Davidian M., Zhang M., & Laber E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat*, *1*(1), 103–114. https://doi.org/10.1002/sta.411

Zhang B., & Zhang M. (2018). Variable selection for estimating the optimal treatment regimes in the presence of a large number of covariates. *The Annals of Applied Statistics*, *12*(4), 2335–2358. https://dx.doi.org/10.1214/18-AOAS1154

Zhang B., & Zhang M. (2022). Subgroup identification and variable selection for treatment decision making. *The Annals of Applied Statistics*, *16*(1), 40–59. https://doi.org/10.1214/21-AOAS1468

Zou H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735