HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2017 August 11.

Published in final edited form as:

J Am Stat Assoc. 2017 ; 112(518): 769–778. doi:10.1080/01621459.2016.1166115.

Change-Plane Analysis for Subgroup Detection and Sample Size Calculation

Ailin Fan [graduate student], **Rui Song [Associate Professor]**, and **Wenbin Lu*** **[Professor]**
Department of Statistics, North Carolina State University, Raleigh, NC 27695.

Abstract

We propose a systematic method for testing and identifying a subgroup with an enhanced treatment effect. We adopts a change-plane technique to first test the existence of a subgroup, and then identify the subgroup if the null hypothesis on non-existence of such a subgroup is rejected. A semiparametric model is considered for the response with an unspecified baseline function and an interaction between a subgroup indicator and treatment. A doubly-robust test statistic is constructed based on this model, and asymptotic distributions of the test statistic under both null and local alternative hypotheses are derived. Moreover, a sample size calculation method for subgroup detection is developed based on the proposed statistic. The finite sample performance of the proposed test is evaluated via simulations. Finally, the proposed methods for subgroup identification and sample size calculation are applied to a data from an AIDS study.

Keywords

Change-plane analysis; Doubly robust test; Sample size calculation; Semiparametric model; Subgroup analysis

1 INTRODUCTION

Classical clinical trials are designed to assess therapeutic benefits of treatments for the whole population that has been considered. However, due to patients' heterogeneity in response to treatments, it is likely that a new treatment is effective or has an enhanced effect compared to a standard treatment only for a specific subpopulation. By making use of patient-specific baseline covariates, subgroup analysis aims to identify subgroups of patients with enhanced treatment effects, which can help to narrow down the target population of a treatment. Hence, it provides an important tool for assessing treatment effects and selecting target populations for future studies.

A number of data-driven approaches have been developed for the subgroup identification. Song and Pepe (2004) considered the binary response case and proposed using the selection impact curve (SIC) to evaluate treatment policies dictated by a single covariate. Then, based on the SIC, an optimal division of the population for assigning treatments can be obtained. Bonetti and Gelber (2004) grouped patients by values of a single covariate and estimated

*(lu@stat.ncsu.edu).
(afan@ncsu.edu), (rsong@ncsu.edu)

treatment effects on overlapping subsets of patients using a moving average procedure. Kuk et al. (2010) used recursive subsetting algorithm for identifying subgroups who respond to treatment with high prediction accuracy for clinical outcomes. Foster et al. (2011) developed a "Virtual Twins" method which first predicted the probabilities of response to treatment and control, and then used tree methods to obtain the subgroups with an enhanced treatment effect. Cai et al. (2011) and Zhao et al. (2013) proposed using parametric scoring systems based on multiple baseline covariates to rank treatment effects and then identified patients who benefit more from the new treatment. There are, however, well known risks for undertaking subgroup analysis (Assmann et al., 2000; Wang et al., 2007). For example, subgroup identification may suffer from false positive findings without being performed with a sound statistical hypothesis testing procedure.

Recently, Shen and He (2015) considered a linear logistic-normal mixture model for the response and developed a likelihood-based test for the existence of a subgroup. If a subgroup exists as indicated by the test, the fitted logistic regression model for the subgroup indicator can be used to score patients for treatment selection. The method proposed in Shen and He (2015) provides a valid test for detecting the subgroup with the following two limitations. First, the method relies on some parametric assumptions, such as linear covariate effects and a logistic-normal mixture model for the response, which may be restrictive in applications. Second, since the subgroup is defined by a latent binary variable, the fitted logistic probability for the subgroup indicator is used for treatment selection. This requires selecting a proper threshold parameter, which can be subjective.

In this paper, we consider change-plane analysis for subgroup detection and sample size calculation. Our contribution over the literature can be summarized in the following three folds. First, we consider a semiparametric model with an unspecified baseline function and an interaction between a subgroup indicator and the treatment for the mean response, which greatly improves the flexibility of the response models considered in the literature. In addition, the subgroup indicator is explicitly defined by a change-plane as a function of covariates. Second, adopting techniques similar as those in change-point analysis (Liang et al., 1990; Andrews, 1993; Bai, 1997), we propose a doubly-robust score-type statistic for testing the existence of a subgroup with an enhanced treatment effect. The proposed test is doubly-robust in the sense that it is valid when either the baseline function or the propensity score model is correctly specified. If the null hypothesis that a subgroup does not exist is rejected, the change-plane that defines the subgroup can be estimated by maximizing the score-type statistic. Third, we derive the asymptotic distributions of the proposed statistic under both the null and the local alternative hypotheses. A resampling method is developed to approximate the asymptotic null distribution of the test statistic. Based on the derived asymptotic distributions, we also propose a sample size calculation procedure to design a randomized clinical trial for subgroup detection, which has been seldom studied in the literature.

The remainder of this paper is organized as follows. Section 2 introduces the considered semi-parametric model and the proposed doubly-robust score test statistic for subgroup detection. The asymptotic distributions of the test statistic under both the null and local alternative hypotheses are also presented. Section 3 presents a sample size calculation

procedure based on the proposed test. The numerical performance of the proposed test and the associated sample size calculation method are evaluated by simulation studies in Section 4. An application of the proposed method to a data from the AIDS Clinical Trials Group protocol 175 is illustrated in Section 5. The paper is concluded with some discussions in Section 6. All the technical derivations are given in the Appendix.

2 CHANGE-PLANE ANALYSIS

2.1 The Proposed Model

Let $\boldsymbol{X}$ denote the baseline covariates collected for a subject in an experimental or observational study, A denote the treatment received by the subject, and Y denote his or her response of interest. Here we restrict our attention to a dichotomous treatment coded as 0 and 1, and a continuous response. Let $\boldsymbol{Z} = (\boldsymbol{X}^T, A, Y)^T$. The observed data consist of $\left\{\boldsymbol{Z}_i=\left(\boldsymbol{X}_i^T, A_i, Y_i\right)^T, i=1,\ldots,n\right\}$, which are n independent and identically distributed (i.i.d.) copies of $\boldsymbol{Z}$. Consider the following semiparametric model

$$Y_i=\mu\left(\boldsymbol{X}_i\right)+\tau A_i \mathbf{1}\left(\boldsymbol{\theta}^{\mathbf{T}}\boldsymbol{X}_{\mathbf{i}}\geq \mathbf{0}\right)+\epsilon_{\mathrm{i}}, \quad (1)$$

where $\mu(\mathbf{X})$ is an unknown baseline mean function for patients in treatment 0, $1(\cdot)$ is the indicator function, and $E(\varepsilon_i|A_i, \boldsymbol{X}_i) = 0$. We assume that the first element of $\boldsymbol{X}$ is 1, $\boldsymbol{X}$ is a $(p + 1)$-dimensional vector $(1, X_1, ..., X_p)^T$, and $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_p)^T$, is a $(p + 1)$-dimensional vector of parameters. For the identifiability of $\boldsymbol{\theta}$, let $\|\boldsymbol{\theta}\| = 1$, where $\|\cdot\|$ is the ℓ_2-norm. When $\tau = 0$, treatments do not have an effect on the response and thus there are no subgroups with enhanced treatment effects. When $\tau \neq 0$, a subgroup of patients with an enhanced treatment effect exists and is defined by the change-plane $1(\boldsymbol{\theta}^T\boldsymbol{X} \geq 0)$.

The proposed model is flexible since it puts no assumptions on the baseline mean function. On the other hand, it places constraints on the form of subgroup and the treatment effect for the subgroup, which are directly related to our goal of subgroup detection and identification. Semipara-metric models analogous to model (1) have been considered in the literature for deriving optimal treatment regimes (Murphy, 2003; Robins, 2004). The difference is the way that the interaction between treatment A and covariates $\boldsymbol{X}$ is modeled. For the subgroup identification problem, we consider the interaction term $A\mathbf{1}(\boldsymbol{\theta}^T\boldsymbol{X} \geq 0)$. To test whether there exists a subgroup with an enhanced treatment effect, it is equivalent to test the hypothesis

$$H_0:\tau=0 \quad \text{versus} \quad H_a:\tau \neq 0. \quad (2)$$

2.2 A Doubly-Robust Test

When $\boldsymbol{\theta}$ is known, model (1) fits in the class of semiparametric models considered in Robins and Rotnizky (2001). Based on the semiparametric theory (Tsiatis, 2007), a class of doubly-robust estimating equations for τ is given by

$$\sum_{i=1}^{n}\lambda\left(\boldsymbol{X}_i\right)\left\{A_i-\pi\left(\boldsymbol{X}_i\right)\right\}\left\{Y_i-h\left(\boldsymbol{X}_i\right)-\tau A_i\mathbf{1}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}_{\mathbf{i}}\geq 0\right)\right\}=0, \tag{3}$$

where $\lambda(\boldsymbol{X})$ and $h(\boldsymbol{X})$ are arbitrary functions, and $\pi(\boldsymbol{X}) = P(A = 1|\boldsymbol{X})$ is the propensity score. It can be shown that when either the baseline mean function $h(\boldsymbol{X})$ or the propensity score model $\pi(\boldsymbol{X})$ is correctly specified, (3) is a consistent estimating equation for τ.

Under the assumption that the random errors ε_i's are homoscedastic, the most efficient doubly-robust estimating equation is obtained by setting $\lambda(\boldsymbol{X}) = \mathbf{1}(\boldsymbol{\theta}^T\boldsymbol{X} \geq 0)$ and $h(\boldsymbol{X}) = \mu(\boldsymbol{X})$. As the true baseline function $\mu(\boldsymbol{X})$ and propensity score model $\pi(\boldsymbol{X})$ may not be known in practice, we posit parametric models $h(\boldsymbol{X}, \beta)$ and $\pi(\boldsymbol{X}, \gamma)$ for $h(\mathbf{X})$ and $\pi(\boldsymbol{X})$, respectively. Hereinafter, we assume a linear model for $h(\boldsymbol{X}, \beta)$ and a logistic model for $\pi(\boldsymbol{X}, \gamma)$. However, other parametric models can also be used. Define $\boldsymbol{\eta} = (\beta^T, \gamma^T)^T$. We consider the following score test statistic for testing $H_0 : \tau = 0$:

$$\sum_{i=1}^{n}\psi_1\left(\boldsymbol{Z}_i,\tilde{\eta};\boldsymbol{\theta}\right)\equiv\sum_{i=1}^{n}\mathbf{1}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}_{\mathbf{i}}\geq 0\right)\left\{\mathbf{A_i}-\pi\left(\boldsymbol{X}_{\mathbf{i}},\tilde{\gamma}\right)\right\}\left\{\mathbf{Y_i}-\mathbf{h}\left(\boldsymbol{X}_{\mathbf{i}},\tilde{\beta}\right)\right\},$$

where $\tilde{\eta}=\left(\tilde{\beta}^{\mathrm{T}},\tilde{\gamma}^{\mathrm{T}}\right)^{\mathrm{T}}$, $\tilde{\beta}$ is an estimator of β under the null, and $\tilde{\gamma}$ is an estimator of γ. Specifically, $\tilde{\beta}$ and $\tilde{\gamma}$ are solutions to the following equations

$$\Psi_{2n}\left(\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^{n}\psi_2\left(\boldsymbol{Z}_i,\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^{n}D_{\beta}\left(\boldsymbol{X}_i\right)\left\{Y_i-h\left(\boldsymbol{X}_i,\boldsymbol{\beta}\right)\right\}=0,$$

$$\Psi_{3n}\left(\boldsymbol{\gamma}\right)=\frac{1}{n}\sum_{i=1}^{n}\psi_3\left(\boldsymbol{Z}_i,\boldsymbol{\gamma}\right)=\frac{1}{n}\sum_{i=1}^{n}D_{\gamma}\left(\boldsymbol{X}_i\right)\left\{A_i-\pi\left(\boldsymbol{X}_i,\boldsymbol{\gamma}\right)\right\}=0,$$

where $D_{\beta}(\boldsymbol{X}_i) = \partial h(\boldsymbol{X}_i, \beta)/\partial\beta$ and $D_{\gamma}(\boldsymbol{X}_i) = [\pi(\boldsymbol{X}_i, \gamma)\{1 - \pi(\boldsymbol{X}_i, \gamma)\}]^{-1}\partial\pi(\boldsymbol{X}_i, \gamma)/\partial\gamma$.

Here, although two parametric models are considered for fitting the baseline mean and propensity score functions, we do not require that both models hold in our theoretical derivation. In fact, our theoretical results show that the proposed test is valid when the model for either the baseline mean or propensity score function is correctly specified but not necessarily both, i.e. the so-called doubly robust property. In particular, when the propensity score is known as in randomized clinical trials, the proposed test is valid for any nonparametric baseline mean function, regardless of correctness of the posited linear model. Such theoretical results were also justified by our simulation studies. In this sense, although the proposed test has a parametric form, it is semiparametric in nature.

Note that model (1) does not depend on $\boldsymbol{\theta}$ when $\tau = 0$, hence the parameter $\boldsymbol{\theta}$ is identifiable only under the alternative hypothesis. This makes the testing problem given in (2) non-

regular and standard asymptotic testing framework are not directly applicable. Davies (1977, 1987) consider tests when a nuisance parameter appears under the alternative hypothesis. Andrews (2001) studied such non-regular testing problems for a number of likelihood-based testing procedures under a variety of parametric models. Similar testing problems have also been widely studied for detecting change-points. However, to our knowledge, it remains uninvestigated for detecting the existence of a change-plane based on a semiparametric model. We consider a supremum of squared score test statistics:

$$T_n = \sup_{\theta \in \Theta} \frac{\{\sum_{i=1}^n \psi_1(\boldsymbol{Z}_i, \tilde{\eta}; \boldsymbol{\theta})\}^2}{n\tilde{V}_S(\boldsymbol{\theta})}, \quad (4)$$

where $\Theta = \left\{\boldsymbol{\theta} \in {}^{p+1} : \|\boldsymbol{\theta}\| = 1\right\}$ and $\tilde{V}_S(\boldsymbol{\theta})$ is a consistent estimator for the asymptotic variance of $n^{-1/2}\sum_{i=1}^n \psi_1(\boldsymbol{Z}_i, \tilde{\eta}; \boldsymbol{\theta})$ under the null hypothesis. The definition of $\tilde{V}_S(\boldsymbol{\theta})$ is given in the next section.

To compute the test statistic T_n, we need to find the supremum of squared score test statistics over a unit ball in $^{p+1}$. Since it is infeasible to get the supremum explicitly, we use a numerical method to find the maximum over the space Θ. To incorporate the unit ball constraint, it is natural to consider a sphere coordinates transformation $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)^T \mapsto \boldsymbol{\theta}$, where ϕ_p ranges over $[0, 2\pi)$ and other elements of $\boldsymbol{\phi}$ range over $[0, \pi]$. The transformation is given as follows

$$\begin{cases} \theta_0 = cos(\phi_1), \\ \cdots \\ \theta_{p-1} = sin(\phi_1) sin(\phi_2) \cdots cos(\phi_p), \\ \theta_p = sin(\phi_1) sin(\phi_2) \cdots sin(\phi_p). \end{cases}$$

We consider a set of grid points of $\boldsymbol{\phi}$ over $[0, \pi]^{p-1} \times [0, 2\pi)$ and compute the maximum of squared score statistics over the set of grid points to approximate T_n.

In the next section, we establish the asymptotic distributions of T_n under both the null and the local alternative hypotheses. In addition, we propose a resampling method to compute the critical values of the limiting null distribution. When the null hypotheses is rejected, the change-plane parameter $\boldsymbol{\theta}$ can be estimated by

$$\hat{\boldsymbol{\theta}} = \arg\sup_{\theta \in \Theta} \frac{\left\{\sum_{i=1}^n \psi_1 \frac{(\boldsymbol{Z}_i, \tilde{\eta}; \boldsymbol{\theta})}{}\right\}^2}{n\tilde{V}_S(\boldsymbol{\theta})}. \quad (5)$$

Similar methods for estimating a change point have been studied in the literature (e.g. Bai, 1997). Thus the estimated subgroup with an enhanced treatment effect is $1\left(\hat{\theta}^T \boldsymbol{X} \geq 0\right)$.

2.3 Asymptotic Distributions of T_n

Define $\Psi_2(\beta) = E\{\Psi_{2n}(\beta)\}$ and $\Psi_3(\gamma) = E\{\Psi_{3n}(\gamma)\}$. To establish the asymptotic distributions of T_n, we make the following assumptions:

A1. Equations $\Psi_2(\beta) = 0$ and $\Psi_3(\gamma) = 0$ have unique solutions β_0 and γ_0, respectively, and the solutions $\eta_0 = \left(\beta_0^T, \gamma_0^T\right)^T$ are in a compact set of the parameter space.

A2. We have

$$\sqrt{n}\left(\tilde{\beta} - \boldsymbol{\beta_0}\right) = -C_1^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_2\left(\boldsymbol{Z}_i, \boldsymbol{\beta}_0\right) + o_p\left(1\right),$$

$$\sqrt{n}\left(\tilde{\gamma} - \boldsymbol{\gamma_0}\right) = -C_2^{-1}\sum_{i=1}^{n}\psi_3\left(\boldsymbol{Z}_i, \gamma_0\right) + o_p\left(1\right),$$

where $C_1 = E\{\partial\psi_2(\boldsymbol{Z}, \beta)/\partial\beta^T|_{\beta=\beta_0}\}$, $C_2 = E\{\partial\psi_3(\boldsymbol{Z}, \gamma)/\partial\gamma^T|_{\gamma=\gamma_0}\}$, and both of them are finite and positive definite deterministic matrices.

A3. The function $\psi_1(\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\eta}$, and has bounded first and second derivatives.

A4. The function $E[\{Y - h(\boldsymbol{X}, \beta)\}^2]$ is uniformly bounded in β.

A5. We have $0 < P(\boldsymbol{\theta}^T\boldsymbol{X} \geq 0) < 1$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Assumptions A1 and A2 ensure the consistency and asymptotic normality of $\tilde{\beta}$ and $\tilde{\gamma}$ These assumptions are satisfied for many commonly used parametric models under mild conditions, such as a linear model for $h(\boldsymbol{X}, \beta)$ and a logistic model for $\pi(\boldsymbol{X}, \gamma)$. The asymptotic distributions of $\tilde{\beta}$ under the null and local alternative hypotheses are similar to those established in Le Cam's third lemma (Van der Vaart (2000, p. 90)). Assumptions A3-A5 are assumed to establish the weak convergence of the process $n^{-1/2}\sum_{i=1}^{n}\psi_1\left(\boldsymbol{Z}_i, \tilde{\eta}; \boldsymbol{\theta}\right)$ indexed by $\boldsymbol{\theta}$.

Theorem 1—Suppose that either the baseline mean function $h(\mathbf{X}, \beta)$ or the propensity model $\pi(\mathbf{X}, \gamma)$ is correctly specified, but not necessarily both. If Assumptions A1-A5 hold, T_n converges in distribution to $\sup_{\theta \in \boldsymbol{\Theta}} G^2(\boldsymbol{\theta})$ under H_0 as n goes to infinity , where$\{G(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is a mean zero Gaussian process with the covariance function

$$\Sigma\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) = E\left\{\psi_{1*}\left(\boldsymbol{Z}, \boldsymbol{\eta}_0; \boldsymbol{\theta}_1\right)\psi_{1*}\left(\boldsymbol{Z}, \boldsymbol{\eta}_0; \boldsymbol{\theta}_2\right)\right\} / \sqrt{E\psi_{1*}^2\left(\boldsymbol{Z}, \boldsymbol{\eta}_0; \boldsymbol{\theta}_1\right)E\psi_{1*}^2\left(\boldsymbol{Z}, \boldsymbol{\eta}_0; \boldsymbol{\theta}_2\right)}$$

for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$, *where*

$$\psi_{1*}(\boldsymbol{Z},\boldsymbol{\eta}_0;\boldsymbol{\theta}) = \psi_1(\boldsymbol{Z},\boldsymbol{\eta}_0;\boldsymbol{\theta}) - K_1^T C_1^{-1}\psi_2(\boldsymbol{Z},\beta_0) - K_2^T C_2^{-1}\psi_3(\boldsymbol{Z},\gamma_0),$$

$K_1 = E\{\partial\psi_1(\boldsymbol{Z}, \boldsymbol{\eta}_0; \boldsymbol{\theta})/\partial\beta\}$, and $K_2 = E\{\partial\psi_1(\boldsymbol{Z}, \boldsymbol{\eta}_0; \boldsymbol{\theta})/\partial\gamma\}$.

Next, we establish the asymptotic distribution of T_n under a sequence of local alternatives H_{1n}: $\tau = n^{-1/2}\delta$.

Theorem 2—Suppose that either the baseline mean function h(**X**, β) or the propensity model $\pi(\boldsymbol{X},\gamma)$ is correctly specified, but not necessarily both. If Assumptions A1-A5 hold, T_n converges in distribution to $sup_{\theta\in\Theta} G_\delta^2(\boldsymbol{\theta})$ *under* H_{1n} *as n goes to infinity, where*$\{G_\delta(\boldsymbol{\theta}) : \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ *is a Gaussian process with the mean function*

$$\mu(\boldsymbol{\theta}) = \delta E\left[\left\{\mathbf{1}\left(\boldsymbol{\theta}_0^{\mathrm{T}}\boldsymbol{X}\geq 0, \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}\geq 0\right)\pi_0(\boldsymbol{X})\{1-\pi(\boldsymbol{X},\gamma_0)\}\right]\right/\sqrt{E\{\psi_{1*}^2(\boldsymbol{Z},\boldsymbol{\eta}_0;\boldsymbol{\theta})\}}$$

and the covariance function $\Sigma(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, *where* $\boldsymbol{\theta}_0$ *is true value of* $\boldsymbol{\theta}$ *and* $\pi_0(\boldsymbol{X})$ *is the true propensity score model.*

To calculate the critical values for the test, we use a resampling method to approximate the limiting null distribution of the test statistic. Define

$$\hat{\psi}_{1*}(\boldsymbol{Z},\tilde{\eta};\boldsymbol{\theta}) = \psi_1(\boldsymbol{Z},\tilde{\eta};\boldsymbol{\theta}) - \hat{K}_1^T\hat{C}_1^{-1}\psi_2\left(\boldsymbol{Z},\tilde{\beta}\right) - \hat{K}_2^T\hat{C}_2^{-1}\psi_3(\boldsymbol{Z},\tilde{\gamma}),$$

where $\hat{K}_1$, $\hat{K}_2$, $\hat{C}_1$, and $\hat{C}_2$, are the empirical estimates of their population counterparts. Specifically, $\hat{K}_1 = \frac{1}{n}\sum_{i=1}^n \partial\psi_1(\boldsymbol{Z}_i,\tilde{\eta};\boldsymbol{\theta})/\partial\boldsymbol{\beta}$, $\hat{K}_2 = \frac{1}{n}\sum_{i=1}^n \partial\psi_1(\boldsymbol{Z}_i,\tilde{\eta};\boldsymbol{\theta})/\partial\boldsymbol{\gamma}$, $\hat{C}_1 = \frac{1}{n}\sum_{i=1}^n \partial\psi_2\left(\boldsymbol{Z}_i,\tilde{\beta}\right)/\partial\boldsymbol{\beta}^T$ and $\hat{C}_2 = \frac{1}{n}\sum_{i=1}^n \partial\psi_3(\boldsymbol{Z}_i,\tilde{\gamma})/\partial\boldsymbol{\gamma}^T$. Then, $\tilde{V}_S(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n\left\{\hat{\psi}_{1*}(\boldsymbol{Z},\tilde{\eta};\boldsymbol{\theta})\right\}^2$. We consider the following perturbed test statistic

$$T_n^* = \sup_{\theta\in\boldsymbol{\Theta}}\frac{\left\{\sum_{i=1}^n \boldsymbol{\xi}_i\hat{\psi}_{1*}(\boldsymbol{Z}_i,\tilde{\eta};\boldsymbol{\theta})\right\}^2}{n\tilde{V}_S(\boldsymbol{\theta})},$$

where $\xi_1, ..., \xi_n$ are i.i.d. standard normal random variables independent of data. By generating a large number of perturbed test statistics, we can use the empirical distribution of T_n^* to compute the critical value C_α, the upper α quantile of the empirical distribution. Then an α-level test reject the null hypothesis when $T_n > C_\alpha$.

3 SAMPLE SIZE CALCULATION

Since most clinical trials are designed to detect the overall treatment effect, they may lack power to detect a subgroup with an enhanced treatment effect (Yusuf et al., 1991; Rothwell, 2005). For example, Brookes et al. (2004) has shown that a trial with 80% power for the

overall effect had only 29% power to detect an interaction effect of the same magnitude. To appropriately conduct a subgroup analysis with targeted power, a careful design and predefined statistical analysis protocol are important (Assmann et al., 2000; Cui et al., 2002). In this section, we provide a sample size calculation method based on the proposed test for subgroup detection in a randomized clinical trial.

To derive the sample size formula, we first calculate the asymptotic power of the test under the local alternatives $H_{1n}: \tau = n^{-1/2}\delta$, where n is the sample size. The sample size formula can then be derived at a pre-specified power $1 - \beta$. Here we are interested in sample size calculation for a randomized trial, therefore the propensity score is given and there is no need to estimate γ. In addition, we assume that the errors ε_i's in model (1) are i.i.d. with mean 0 and variance σ^2. Under this case, the asymptotic covariance function of the test statistic T_n can be simplified as

$$\Sigma(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{E\left\{\mathbf{1}\left(\boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{X} \geq 0, \boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{X} \geq 0\right)\mathrm{g}(\boldsymbol{X})\right\}}{\sqrt{E\left\{\mathbf{1}\left(\boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{X} \geq 0\right)\mathrm{g}(\boldsymbol{X})\right\} E\left\{\mathbf{1}\left(\boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{X} \geq 0\right)\mathrm{g}(\boldsymbol{X})\right\}}},$$

where $g(\boldsymbol{X}) = \pi(1 - \pi)[\{\mu(\boldsymbol{X}) - h(\boldsymbol{X}, \beta_0)\}^2 + \sigma^2]$ and $\pi = P(A = 1)$. In addition, under the local alternatives, the asymptotic mean function of T_n is given by

$$\mu(\boldsymbol{\theta}) = \delta\pi(1 - \pi)\frac{E\left\{\mathbf{1}\left(\boldsymbol{\theta}_0^{\mathrm{T}}\boldsymbol{X} \geq 0, \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X} \geq 0\right)\right\}}{\sqrt{E\left\{\mathbf{1}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X} \geq 0\right)\mathrm{g}(\boldsymbol{X})\right\}}}.$$

For an α-level test to have $1 - \beta$ power in detecting an enhanced treatment effect of size τ_0, we need to find δ_0 such that $P\left(sup_{\theta\in\Theta} G^2_{\delta_0}(\boldsymbol{\theta}) > q_\alpha\right) = 1 - \beta$, where q_α is the upper α-quantile of the distribution of $\sup_{\theta\in\Theta} G^2(\boldsymbol{\theta})$ is the Gaussian process defined in Theorem 1. Based on the relationship $\tau_0 = n^{-1/2}\delta_0$, the required sample size is given by $n = (\delta_0/\tau_0)^2$. To find δ_0, we take the following three steps. In Step 1, we compute the mean function $\mu(\boldsymbol{\theta})$ and the covariance function $\Sigma(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ via numerical integration, for which we need to specify the true value $(\boldsymbol{\theta})$ in the change-plane, the distribution of covariates $\boldsymbol{X}$, the difference between the true baseline mean function and the posited mean function, $\mu(\boldsymbol{X}) - h(\boldsymbol{X}, \beta_0)$, and σ_2, the variance of ε. These quantities can be estimated from historical data or a pilot study. In Step 2, for any given δ, we compute the probability $P\left(sup_{\theta\in\Theta} G^2_{\delta}(\boldsymbol{\theta}) > q_\alpha\right)$ via Monte Carlo simulations detailed as follows. We first approximate $sup_{\theta\in\Theta} G^2_{\delta}(\boldsymbol{\theta})$ by $max_{k=1,\ldots,K} G^2_{\delta}(\boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ is a set of fine grids of $\boldsymbol{\theta} \in \Theta$. This can be done by the same sphere coordinates transformation used previously. Next, we generate $(W_1, \ldots, W_K)^T$ from a multivariate normal distribution with the mean $(\mu(\boldsymbol{\theta}_1), \ldots, \mu(\boldsymbol{\theta}_K))^T$ and the variance-covariance matrix $\{\Sigma(\boldsymbol{\theta}_{k1}, \boldsymbol{\theta}_{k2}) : k_1, k_2 = 1, \ldots, K\}$. Finally, we compute the probability $P\left(sup_{\theta\in\Theta} G^2_{\delta}(\boldsymbol{\theta}) > q_\alpha\right)$ based on the empirical distribution of $max_{k=1,\ldots,K} W_k^2$. Note that q_α can be calculated similarly by generating $(W_1, \ldots, W_K)^T$ from a multivariate normal

distribution with mean 0 and the same variance-covariance matrix. In practice, we generate a large set, say 10, 000 of $max_{k=1,\ldots,K} W_k^2$ to compute the probability. In Step 3, we find δ_0 via a grid search.

4 SIMULATION STUDIES

4.1 Test and Estimation

We conducted extensive simulation studies to investigate the empirical performance of the proposed test for subgroup detection and the estimation for the change-plane parameter $\boldsymbol{\theta}$ under the alternative hypotheses. In particular, we considered various settings to examine the robustness of the test against the misspecification of the baseline mean function in both randomized and observational studies.

Simulated data with sample sizes $n = 500$ and 1000 were generated based on model (1), where two covariates $\boldsymbol{X} = (X_1, X_2)^T$ were considered. Here, X_1 follows a Bernoulli distribution with the success probability 0.5 and X_2 follows a uniform distribution on (−1, 1). The random noise ε is normally distributed with mean zero and variance 0.25. For the treatment assignment indicator A, we considered the following two settings for the propensity score model $\pi(\boldsymbol{X})$ (In short as P-Model hereinafter):

- P-Model I: $\pi(\boldsymbol{X}) = 0.5$;

- P-Model II: $\pi(\boldsymbol{X}) = \frac{exp(\gamma_0+\gamma_1 X_1+\gamma_2 X_2)}{1+exp(\gamma_0+\gamma_1 X_1+\gamma_2 X_2)}$, $\gamma_0 = 0$, $\gamma_1 = \gamma_2 = 0.5$.

The two settings represent a randomized clinical trial and an observational study, respectively. We also considered three baseline mean functions for $\mu(\boldsymbol{X})$ (In short as B-Model hereinafter):

- B-Model I: $\mu(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, $\beta_0 = \beta_1 = \beta_2 = 1$;

- B-Model II: $\mu(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_3 X_2^2$, $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = 0$, $\beta_3 = 1$;

- B-Model III: $\mu(\boldsymbol{X}) = \beta_0 + \beta_1 \sin(\beta_2 X_1 + \beta_3 \pi X_2)$, $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$.

The proposed test was implemented on each simulated dataset. When calculating the test statistic in (4), we fit a linear model $h(\boldsymbol{X}, \beta)$ for the baseline mean function and a logistic model $\pi(\boldsymbol{X}, \gamma)$ for the propensity score. Therefore, the baseline mean function is correctly specified for the setting with B-Model I and is misspecified for the settings with B-Models II and III, while the propensity score model is correctly specified for both P-Model I and II. When calculating the test statistic, we used the spherical coordinates transformation and searched the supremum over $K = 100 \times 100$ grid points, with 100 grid values for each angular coordinate. For each test, we used 1000 resamplings to obtain the critical values of the test. We reported the empirical type I errors and powers of the test. Simulation results are summarized below for both the null ($\tau \neq 0$) and the alternative ($\tau \neq 0$), respectively.

4.1.1 Type I Errors—For each setting, we simulated 5000 data sets to compute type I errors of the test with the significance level of 0.05 and 0.1. The results in Table 1 show that the empirical type I errors are all close to their nominal values, which demonstrate the validity and robustness of the proposed test for subgroup detection.

4.1.2 Powers and Estimates of Change-Plane Parameters—Under alternative hypotheses, the enhanced treatment effect in the subgroup is set to be $\tau = \pm 0.1, \pm 0.25$ and ± 0.5. In addition, the true change-plane parameter is chosen as $\theta_0 = (-0.15, 0.3, 0.942)^T$ for all settings. With this choice of θ_0, the subgroup with an enhanced treatment effect contains approximately 50% of the population, and includes subjects with $X_1 = 1$ and $X_2 \geq -0.159$ or $X_1 = 0$ and $X_2 \geq 0.159$.

Empirical powers based on 1000 simulated datasets for all settings are shown in Table 2. As expected, the powers for detecting the subgroup increase as the sample size n or the magnitude of treatment effect τ increases. When the magnitude of treatment effect τ increases to 0.5, the powers are almost 100% for all settings. The powers for B-Model II are comparable to those for B-Model I, while the powers for B-Model III are relatively smaller than those of B-Models I and II. One explanation may be that since B-Model III has a very nonlinear baseline mean function, a posited linear model may not be a good fit and thus lost some efficiency.

Next, we estimated the change-plane parameter θ by (5). We report the bias and the empirical standard deviation of the estimates $\hat{\theta}$ in Figure 1. We also report the misclassification rate for identifying the true subgroup in Table 3. The misclassification rate is the proportion of subjects who are misidentified either as members in the subgroup or as members not in the subgroup, and is calculated by

$$\frac{1}{n}\sum_{i=1}^{n} \left| 1\left(\hat{\theta}^{\mathrm{T}} X_{\mathrm{i}} \geq 0\right) - 1\left(\theta_0{}^{\mathrm{T}} X_{\mathrm{i}} \geq 0\right) \right|.$$

Based on the results, it is observed that the biases and standard deviations of the estimates decrease as the sample size or the magnitude of treatment effect increase. In particular, when the magnitude of treatment effect increases to 0.5, all the estimate are nearly unbiased. For small treatment effects, the estimators for θ are underestimated. This may be because that the true $\theta_2 = 0.942$, which is likely to be underestimated due to the upper limit of 1 for θ. Similar to the power results, the estimates for B-Model III have larger biases and standard deviations compared to those for B-Model I and II. In addition, Table 3 shows that the misclassification rates also decrease as the sample size n or the magnitude of treatment effect τ increases. When the magnitude of the treatment effect increases to 0.5, most misclassification rates are less than 5% except for those under B-Model III.

For practical use, we also report the computational time for conducting the proposed test for different sample sizes n and number of grid points K on the unit ball Θ. Table 4 summarizes the average computational time (in seconds) and its standard deviation over 1000 simulation runs for the setting with B-Model I, P-Model I, $\tau = 0.5$ and $M = 1000$. The average computational time increases almost linearly in n and K. But even with $n = 2000$ and $K = 10000$, the average computational time is less than 2 minutes.

Furthermore, we compare the proposed method with the test of Shen and He (2015) (named EM test) in terms of power for detecting the subgroup and with the method of Zhao et al. (2013) in terms of accuracy for identifying the true subgroup. Specifically, when comparing with the EM test of Shen and He (2015), we consider the simulation settings with P-Model I

and B-Model I/II/III at $\tau = 0$, 0.1 and 0.5. Note that under B-Model I, since the null model is a linear model, the EM test is a valid test. The type-I error and power results are reported in Table 5. We only present the results for B-Model I. Based on the results, we observe that when the baseline mean model is linear (B-Model I), the EM test has correct type I error at both levels, however, its empirical power is smaller than that of the proposed test, especially for $\tau = 0.1$. When the baseline mean model is not linear (B-Model II and III), the model assumption for the EM test of Shen and He (2015) is violated. Our results (not presented here) show that the EM test has severely inflated type I error and hence is not valid as expected, while the proposed test still has correct type I error due to the doubly robust property.

In Zhao et al. (2013), the subgroup of interest is defined by $\left\{\hat{D}(\boldsymbol{X}) \geq c\right\}$, where $\hat{D}(\boldsymbol{X})$ is the estimated treatment difference and c is a threshold that will be determined based on data. Specifically, they propose an estimate of the average treatment effect in the subgroup defined by $\left\{\hat{D}(\boldsymbol{X}) \geq c\right\}$ and denote it by $\widehat{AD}(c)$. Then, under the monotonicity assumption, the threshold c is a chosen to solve the equation $\widehat{AD}(c) = \tau$, where τ is a desired treatment effect in the subgroup. In our implementation, following Zhao et al. (2013), we first fit a linear model: $E(Y|\boldsymbol{X}, A) = \beta_1^T \boldsymbol{X} + \left(\beta_2^T \boldsymbol{X}\right) A$. Note that in our notation, the covariates $\boldsymbol{X}$ include a column of 1 as the first column. The estimated treatment difference is $\hat{D}(\boldsymbol{X}) = \hat{\beta}_2^T \boldsymbol{X}$ with the subgroup defined by $\left\{\hat{\beta}_2^T \boldsymbol{X} \geq c\right\}$ Then, we obtain the estimated average treatment difference $\widehat{AD}(c)$ in the subgroup and determine the threshold c using the same method as in Zhao et al. (2013). We consider the simulation settings with P-Model I and B-Model I/II/III at $\tau = 0.1$ and 0.5. Misclassification rates of the identified subgroups using Zhao et al. (2013)'s method and our method compared with the true subgroup are reported in Table 6. The results show that the identified subgroups using Zhao et al. (2013)'s method have much higher misclassification rates compared with the proposed method.

4.2 Sample Size Calculation Examples

In this section, we conducted a simulation study to evaluate the proposed sample size calculation procedure for a randomized trial with the equal treatment assignment probability, i.e. $\pi = 0.5$. In this simulation study, we considered a single covariate X from a uniform distribution on (−1,1) in all settings. The subgroup of interest is defined by $X \geq \theta_0$, where θ_0 was chosen as: −0.5, 0 and 0.5, corresponding to the scenarios that 75%, 50% and 25% of the population are in the subgroup with an enhanced treatment effect. The variance of the random noise ε is set as $\sigma^2 = 0.25$. We considered three levels of treatment effects: $\tau = 0.1$, 0.25 and 0.5, which represent small, medium and large effects, respectively. In addition, we considered three baseline mean functions: $\mu(X) = 1 + X$, $1 - X^2$ and $1 + sin(\tilde{\pi} X)$, where $\tilde{\pi}$ is the circumference to diameter ratio. In our test statistics, we fit a linear model $h(X, \beta)$ for $\mu(X)$. After some calculation, it can be shown that when $\mu(X) = 1 - X^2$, $h(X, \beta_0) = 2/3$; while when $\mu(X) = 1 + sin(\tilde{\pi} X)$, $h(X, \beta_0) = 1 + (3/\tilde{\pi}) X$. Therefore, the difference $\mu(X) - h(X, \beta_0)$ can be calculated accordingly. We calculate the required sample size n for the test with

level $\alpha = 0.05$ and power $1 - \beta = 90\%$. For each setting, based on the calculated sample size n, we generated 1000 data sets and computed the empirical power of the proposed test statistic. Table 7 summarizes the results. We observe that the empirical powers under all settings are close to the nominal level 90%, which shows the validity of the proposed sample size formula.

For comparison, we also compute the required sample size for subgroup analysis based on a simple method proposed by Brookes et al. (2004). The main idea of Brookes et al. (2004) is to inflate the original sample size for testing the overall treatment effect such that the interaction test between treatment and subgroup can achieve the nominal level. For example, when the subgroup contains half of the population, the required sample size for testing the treatment-subgroup interaction is $n = 4n_{overall}$, where $n_{overall}$ is the required sample size for testing the overall treatment effect. In our considered simulation settings, $\theta_0 = 0$ corresponds to the case that the subgroup contains half of the population. We only consider this setting in our comparison. Based on their formula, we need to calculate the sample size $n_{overall}$ for testing the overall treatment effect based on model $Y = \beta_0 + \beta_1 X + \tau A + \varepsilon$. Here, we use the sample size formula for analysis of covariance proposed in Borm et al. (2007). Specifically, $n_{overall} = (1 - \rho^2)n_t$, where $n_t = \frac{4(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\tau^2}$ is the required sample size for a standard two-sample t-test, σ^2 is the variance of ε, and ρ is the correlation between X and Y.

Table 8 shows the calculated sample size using the above formula for level $\alpha = 0.05$ and power $1 - \beta = 0.9$. We also report the corresponding empirical power of our proposed test under the calculated sample size. Based on these results, the sample sizes calculated using the method of Brookes et al. (2004) are all smaller than those obtained using the proposed sample size formula given in Table 7, and the corresponding power is much lower than the desired 90% level. This demonstrates that simply inflating the sample size for testing the overall treatment effect may not work well for detecting the subgroup and the proposed delicate sample size formula is necessary.

5 APPLICATION TO AIDS DATA

We illustrated the proposed method with a data from the AIDS Clinical Trials Group (ACTG) protocol 175 (Hammer et al., 1996), a study that randomized subjects to four different daily regimens: zidovudine (ZDV) monotherapy, ZDV + didanosine (ddI), ZDV + zalcitabine (zal) and ddI monotherapy. We focused on comparing two treatments: ZDV+ddI (treatment 1) and ZDV+zal (treatment 0). There are 522 subjects in treatment 1 and 524 subjects in treatment 0. Following Lu et al. (2013), we considered the CD4 counts (cells/mm^3) at 20±5 weeks after randomization as the response and used two covariates for subgroup identification: age (years) and homosexual activity (0=no, 1=yes), denoted as homo.

We applied the proposed method to detect whether there is a subgroup with an enhanced treatment effect. A linear model was used for the baseline function. The value of the test statistic is 21.25, which is calculated based on 200 × 50 grid points of the sphere coordinates. The p-value based on 1000 resamplings is less than 0.001, showing a strong evidence for the existence of a subgroup with an enhanced treatment effect. The estimated

change-plane parameter $\hat{\boldsymbol{\theta}}=(-0.576, 0.037, -0.816)^T$. The identified subgroup includes subjects with *age* > 37.64 if *homo* = 1 or with *age* > 15.58 if *homo* = 0. There are 622 subjects included in this subgroup, among which 315 subjects received treatment 1 and 307 received treatment 0. Given the estimated change-plane and the fitted linear model for the baseline mean function, we can estimate the enhanced treatment effect τ, which is $\hat{\tau}=41.6$. Therefore, for patients in the identified subgroup, treatment 1 is better than treatment 0. This agrees with the findings in Lu et al. (2013) that treatment 1 is better than treatment 0 for older patients.

Next, based on the AIDS data, we calculated the required sample size for subgroup detection in future balanced randomized trials. As an illustration, we considered the test with size α = 0.05 and power $1-\beta = 0.9$. From the AIDS data, we estimated the standard deviation of ε as $\hat{\sigma}=145.9$. Therefore, we set $\sigma=\hat{\sigma}$ and the true change-plane parameter as $\boldsymbol{\theta}_0$ = (−0.576, 0.037, −0.816)^T. Covariate *age* is assumed from a normal distribution with estimated mean 35.33 and standard deviation 8.75 and covariate *homo* is from a Bernoulli distribution with the success probability 0.66, similar to those in the AIDS data. For simplicity, we set the difference $\mu(\boldsymbol{X}) - h(\boldsymbol{X}, \beta_0) = 0$. The estimated sample sizes for different treatment effect sizes are given in Table 9. The estimated sample size has a wide range, which is common in practice since the weaker the treatment effect is, the larger the sample size is required. For this AIDS study, there were 1046 subjects receiving either treatment 1 or treatment 0. Therefore, the proposed test can approximately achieve 90% power for identifying a subgroup with an enhanced treatment effect $\tau = 60$.

6 DISCUSSION

In this paper, based on a change-plane analysis technique, we developed a doubly-robust testing procedure for detecting a subgroup with an enhanced treatment effect. We established the asymptotic distributions of the proposed test statistic under both the null and the local alternative hypotheses. We also developed its associated sample size calculation method, which is useful for sizing a clinical trial with desired power for subgroup detection.

In our current work, the subgroup with an enhanced treatment effect is defined by a change-plane, which may be restrictive sometimes. It is feasible to extend the way for defining a subgroup from a change-plane to more general forms, for example, a combination of multiple change-planes $1(\theta_{10} + \theta_{11}X_1 + \theta_{12}X_2 \geq 0)$ and $1(\theta_{20} + \theta_{21}X_3 \geq 0)$. However, a more complicated form for defining a subgroup requires a more comprehensive way of searching the supremum of squared score-type statistics over the possible space, which may be challenging. One assumption made in the considered semiparametric model is that the enhanced treatment effect in the subgroup is constant. It is also interesting to study a more general situation that the magnitude of the enhanced treatment effect varies for subjects in the subgroup. Lastly, it is likely that many covariates are collected at the baseline but not all of them are useful for subgroup detection. Therefore, a built-in variable selection for subgroup detection will be very helpful, which warrants further investigation.

Acknowledgments

The authors would like to thank an associate editor and two referees for their thoughtful and constructive comments, which help to improve an earlier version of the paper. This work was partly supported by a NIH grant P01 CA142538.

Appendix: Proofs of Theorems

Proof of Theorem 1

By Taylor expansion and assumptions A1-A2, we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\boldsymbol{Z}_i,\tilde{\eta};\boldsymbol{\theta}\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\psi_1\left(\boldsymbol{Z}_i,\boldsymbol{\eta}_0;\boldsymbol{\theta}\right)-K_1^TC_1^{-1}\psi_2\left(\boldsymbol{Z},\beta_0\right)-K_2^TC_2^{-1}\psi_3\left(\boldsymbol{Z},\boldsymbol{\gamma}_0\right)\right\}+o_p\left(1\right)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{1*}\left(\boldsymbol{Z}_i,\boldsymbol{\eta}_0;\boldsymbol{\theta}\right)+o_p\left(1\right),\end{aligned}$$

where $\psi_{1*}(\boldsymbol{Z}_i, \boldsymbol{\eta}_0; \boldsymbol{\theta})$'s are i.i.d. with mean 0 under the null when either the propensity score model or the baseline mean function is correctly specified.

In addition, by assumptions A3-A5, we can show that the class $\mathcal{F}=\{\psi_{1*}\left(\boldsymbol{Z},\boldsymbol{\eta}_0;\boldsymbol{\theta}\right):\|\boldsymbol{\theta}\|=1\}$ is P-Donsker. Therefore, $n^{-1/2}\sum_{i=1}^{n}\psi_1\left(\boldsymbol{Z}_i,\tilde{\eta};\boldsymbol{\theta}\right)$ converges weakly to a mean zero Gaussian process with the convariance function $E\{\psi_{1*}\left(\boldsymbol{Z},\boldsymbol{\eta}_0;\boldsymbol{\theta}_1\right)\psi_{1*}\left(\boldsymbol{Z},\boldsymbol{\eta}_0;\boldsymbol{\theta}_2\right)\}$, where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$. Finally, it is easy to show that the variance estimator $\tilde{V}_s\left(\boldsymbol{\theta}\right)$ converges uniformly to $E\left\{\psi_{1*}^2\left(\boldsymbol{Z},\boldsymbol{\eta}_0;\boldsymbol{\theta}\right)\right\}$ for $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ under both the null and the local alternative hypotheses. Therefore, the results established in Theorem 1 hold.

Proof of Theorem 2

Under the local alternatives, we have the same asymptotic representation (6). In addition,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\boldsymbol{Z}_i,\boldsymbol{\eta}_0;\boldsymbol{\theta}\right)\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{1}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}_{\mathbf{i}}\geq 0\right)\left\{\mathbf{A_i}-\pi\left(\boldsymbol{X}_{\mathbf{i}},\boldsymbol{\gamma}_{\mathbf{0}}\right)\right\}\left\{\mathbf{Y_i}-\mathbf{h}\left(\boldsymbol{X}_{\mathbf{i}},\boldsymbol{\beta}_{\mathbf{0}}\right)-\frac{\delta}{\sqrt{n}}\mathbf{A_1 1}\left(\boldsymbol{\theta}_{\mathbf{0}}^{\mathrm{T}}\boldsymbol{X}_{\mathbf{i}}\geq 0\right)\right\}\\ +&\frac{1}{n}\sum_{i=1}^{n}\delta\mathbf{1}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}_{\mathbf{i}}\geq 0,\boldsymbol{\theta}_{\mathbf{0}}^{\mathrm{T}}\boldsymbol{X}_{\mathbf{i}}\geq 0\right)\mathbf{A_i}\left\{\mathbf{A_i}-\pi\left(\boldsymbol{X}_{\mathbf{i}},\boldsymbol{\gamma}_{\mathbf{0}}\right)\right\}.\end{aligned}$$

The terms in the first summation are i.i.d. with mean 0 under the local alternatives when either the propensity score mode or the baseline mean function is correctly specified. As in Theorem 1, it can be show that the first summation term converges weakly to the same mean zero Gaussian process as $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\boldsymbol{Z}_i,\boldsymbol{\eta}_0;\boldsymbol{\theta}\right)$ does under the null. In addition, it can be shown that the second summation term converges uniformly to $\delta E\left[\left\{\mathbf{1}\left(\boldsymbol{\theta}_{\mathbf{0}}^{\mathrm{T}}\boldsymbol{X}\geq 0,\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}\geq 0\right)\pi_{\mathbf{0}}\left(\boldsymbol{X}\right)\left\{\mathbf{1}-\pi\left(\boldsymbol{X},\boldsymbol{\gamma}_{\mathbf{0}}\right)\right\}\right]$ for $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Therefore, the results established in Theorem 2 hold.

References

Andrews DW. Tests for parameter instability and structural change with unknown change point. Econometrica: Journal of the Econometric Society. 1993; 61(4):821–856.

Andrews DW. Testing when a parameter is on the boundary of the maintained hypothesis. Econometrica. 2001; 69(3):683–734.

Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis) uses of baseline data in clinical trials. The Lancet. 2000; 355(9209):1064–1069.

Bai J. Estimation of a change point in multiple regression models. Review of Economics and Statistics. 1997; 79(4):551–563.

Bonetti M, Gelber RD. Patterns of treatment effects in subsets of patients in clinical trials. Biostatistics. 2004; 5(3):465–481. [PubMed: 15208206]

Borm GF, Fransen J, Lemmens WA. A simple sample size formula for analysis of covariance in randomized clinical trials. Journal of clinical epidemiology. 2007; 60(12):1234–1238. [PubMed: 17998077]

Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. Journal of clinical epidemiology. 2004; 57(3):229–236. [PubMed: 15066682]

Cai T, Tian L, Wong PH, Wei L. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011; 12(2):270–282. [PubMed: 20876663]

Cui L, James Hung H, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. Journal of biopharmaceutical statistics. 2002; 12(3):347–358. [PubMed: 12448576]

Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika. 1977; 64(2):247–254.

Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika. 1987; 74(1):33–43.

Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. Statistics in medicine. 2011; 30(24):2867–2880. [PubMed: 21815180]

Hammer SM, Katzenstein DA, Hughes MD, Gundacker H, Schooley RT, Haubrich RH, Henry WK, Lederman MM, Phair JP, Niu M, et al. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. New England Journal of Medicine. 1996; 335(15):1081–1090. [PubMed: 8813038]

Kuk A, Li J, Rush AJ. Recursive subsetting to identify patients in the star* d: a method to enhance the accuracy of early prediction of treatment outcome and to inform personalized care. The Journal of clinical psychiatry. 2010; 71(11):1502–1508. [PubMed: 21114950]

Liang K-Y, Self SG, Liu X. The cox proportional hazards model with change point: An epidemiologic application. Biometrics. 1990; 46(3):783–793. [PubMed: 2242414]

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

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

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

Robins JM, Rotnizky A. Comment on the Bickel and Kwon article, “Inference for semiparametric models: some questions and an answer”. Statistica Sinica. 2001; 11(4):920–936.

Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. The Lancet. 2005; 365(9454):176–186.

Shen J, He X. Inference for subgroup analysis with a structured logistic-normal mixture model. Journal of the American Statistical Association. 2015; 110(509):303–312.

Song X, Pepe MS. Evaluating markers for selecting a patient's treatment. Biometrics. 2004; 60(4):874–883. [PubMed: 15606407]

Tsiatis, A. Semiparametric theory and missing data. Springer Science & Business Media; 2007.

Van der Vaart, AW. Asymptotic statistics, volume 3. Cambridge university press; 2000.

Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicinereporting of subgroup analyses in clinical trials. New England Journal of Medicine. 2007; 357(21):2189–2194. [PubMed: 18032770]

Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. Jama. 1991; 266(1):93–98. [PubMed: 2046134]

Zhao L, Tian L, Cai T, Claggett B, Wei L-J. Effectively selecting a target population for a future comparative study. Journal of the American Statistical Association. 2013; 108(502):527–539. [PubMed: 24058223]

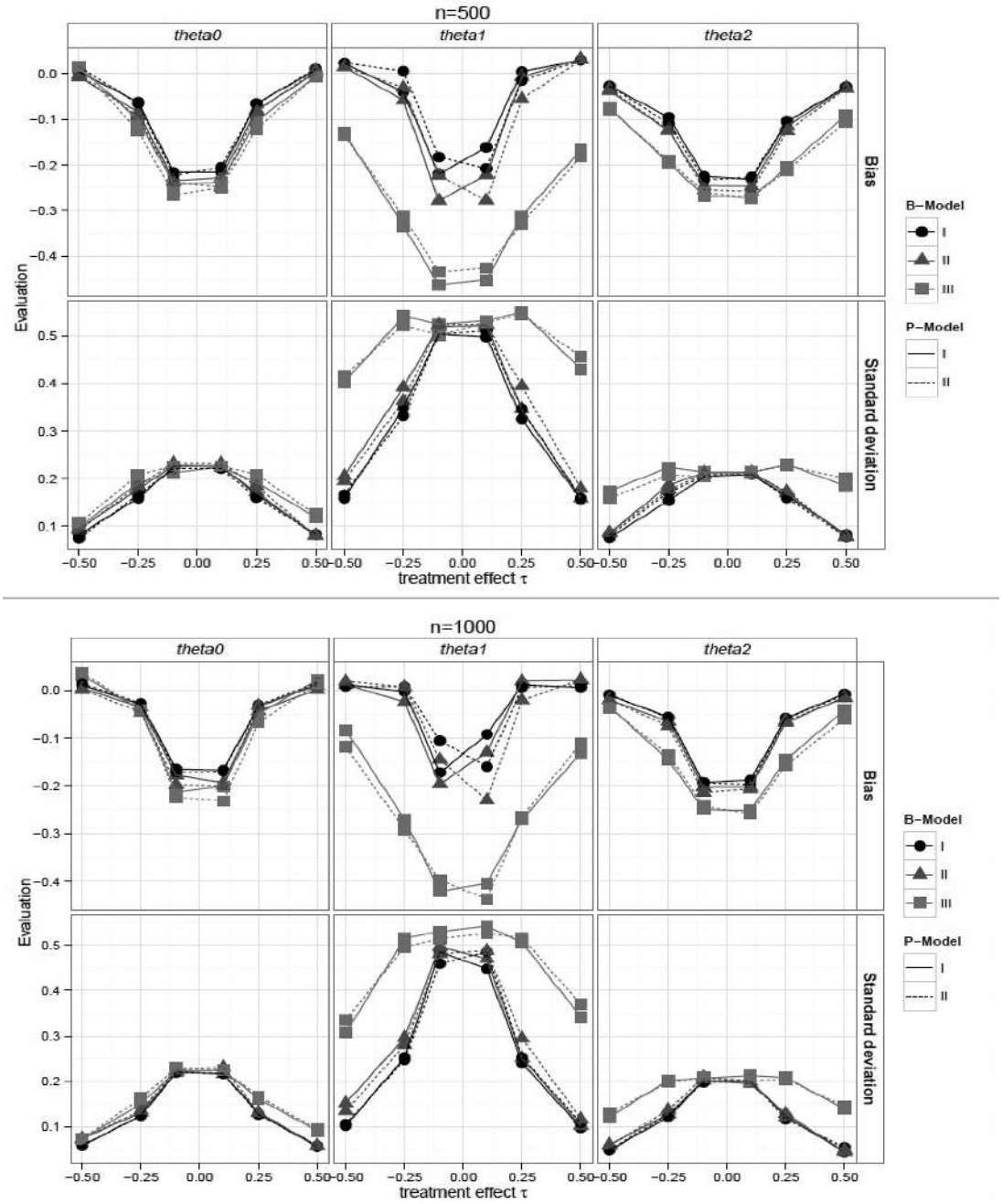

Figure 1.
Bias and standard deviation of estimated change-plane parameter θ.

Table 1

Type I errors of the proposed test based on resampling. (Corresponding standard errors for type I errors with size 0.05 and 0.1 are 0.003 and 0.004.)

		B-Model I		B-Model II		B-Model III	
n	**P-Model**	**size 0.05**	**size 0.1**	**size 0.05**	**size 0.1**	**size 0.05**	**size 0.1**
100	I	0.052	0.104	0.054	0.107	0.050	0.099
	II	0.050	0.105	0.052	0.110	0.051	0.106
500	I	0.052	0.100	0.045	0.102	0.047	0.092
	II	0.055	0.105	0.051	0.106	0.054	0.101
1000	I	0.050	0.101	0.049	0.100	0.053	0.108
	II	0.051	0.105	0.044	0.102	0.053	0.108

Table 2

Power (%) of the proposed test at 0.05 and 0.1 levels. (Standard errors are shown in parenthesis.)

			B-Model I		B-Model II		B-Model III	
n	**P-Model**	***τ***	**size 0.05**	**size 0.1**	**size 0.05**	**size 0.1**	**size 0.05**	**size 0.1**
500	I	0.1	21.2 (1.3)	31.1 (1.5)	17.5 (1.2)	27.5 (1.4)	9.9 (0.9)	16.1 (1.2)
		0.25	90.3 (0.9)	95.1 (0.7)	81.3 (1.2)	88.4 (1.0)	45.9 (1.6)	57.7 (1.6)
		0.5	100 (0)	100 (0)	100 (0)	100 (0)	97.5 (0.5)	99.1 (0.3)
		-0.1	19.9 (1.3)	30.1 (1.5)	16.9 (1.2)	27.0 (1.4)	9.1 (0.9)	16.5 (1.2)
		-0.25	88.7 (1.0)	94.0 (0.7)	74.5 (1.4)	84.3 (1.2)	47.6 (1.6)	60.4 (1.5)
		-0.5	100 (0)	100 (0)	100 (0)	100 (0)	99.4 (0.2)	99.7 (0.2)
	II	0.1	18.8 (1.2)	29.5 (1.4)	21.4 (1.3)	30.8 (1.5)	11.1 (1.0)	18.6 (1.2)
		0.25	84.6 (1.1)	90.2 (0.9)	76.6 (1.3)	83.3 (1.2)	42.9 (1.6)	58.2 (1.6)
		0.5	100 (0)	100 (0)	99.9 (0.1)	100 (0)	97.1 (0.5)	98.8 (0.3)
		-0.1	20.1 (1.3)	29.6 (1.4)	18.3 (1.2)	26.4 (1.4)	12.7 (1.1)	20.8 (1.3)
		-0.25	84.5 (1.1)	91.5 (0.9)	73.9 (1.4)	82.5 (1.2)	46.2 (1.6)	58.0 (1.6)
		-0.5	100 (0)	100 (0)	99.9 (0.1)	100 (0)	98.4 (0.4)	98.8 (0.3)
1000	I	0.1	41.3 (1.6)	52.5 (1.6)	30.4 (1.5)	43.3 (1.6)	19.5 (1.3)	27.1 (1.4)
		0.25	99.8 (0.1)	99.9 (0.1)	99.0 (0.3)	99.5 (0.2)	77.1 (1.3)	86.0 (1.1)
		0.5	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
		-0.1	39.3 (1.5)	51.3 (1.6)	29.9 (1.4)	43.1 (1.6)	17.3 (1.2)	24.3 (1.4)
		-0.25	99.7 (0.2)	99.8 (0.1)	98.2 (0.4)	99.4 (0.2)	78.8 (1.3)	86.4 (1.1)
		-0.5	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	II	0.1	36.2 (1.5)	48.5 (1.6)	29.7 (1.4)	42.1 (1.6)	15.5 (1.1)	23.7 (1.3)
		0.25	99.2 (0.3)	99.8 (0.1)	97.4 (0.5)	99.4 (0.2)	71.9 (1.4)	81.2 (1.2)
		0.5	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
		-0.1	37.6 (1.5)	50.3 (1.6)	27.9 (1.4)	40.1 (1.6)	18.9 (1.2)	29.9 (1.4)
		-0.25	99.6 (0.2)	99.7 (0.2)	96.4 (0.6)	98.0 (0.4)	79.5 (1.3)	88.2 (1.0)
		-0.5	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)

Table 3

Misclassification rate (%) of identified subgroup based on estimated change-plane parameter θ. (Standard errors are shown in parentheses.)

		P-Model I			P-Model II		
n	τ	B-Model I	B-Model II	B-Model III	B-Model I	B-Model II	B-Model III
500	0.1	26.3 (1.4)	28.3 (1.4)	33.3 (1.5)	26.2 (1.4)	30.6 (1.5)	33.1 (1.5)
	0.25	11.8 (1.0)	12.7 (1.1)	22.1 (1.3)	12.2 (1.0)	14.6 (1.1)	22.9 (1.3)
	0.5	4.8 (0.7)	4.7 (0.7)	12.4 (1.0)	4.8 (0.7)	5.2 (0.7)	12.3 (1.0)
	–0.1	26.6 (1.4)	29.2 (1.4)	33.0 (1.5)	26.6 (1.4)	29.0 (1.4)	34.1 (1.5)
	–0.25	12.0 (1.0)	14.1 (1.1)	21.7 (1.3)	11.8 (1.0)	14.2 (1.1)	23.6 (1.3)
	–0.5	4.6 (0.7)	6.3 (0.8)	9.9 (0.9)	4.9 (0.7)	6.1 (0.8)	11.2 (1.0)
1000	0.1	20.9 (1.3)	23.0 (1.3)	30.0 (1.5)	22.9 (1.3)	25.9 (1.4)	32.2 (1.5)
	0.25	7.6 (0.8)	8.5 (0.9)	17.4 (1.2)	8.1 (0.9)	9.4 (0.9)	18.0 (1.2)
	0.5	2.6 (0.5)	3.0 (0.2)	9.1 (0.9)	3.0 (0.5)	3.3 (0.5)	9.3 (0.9)
	–0.1	22.6 (1.3)	23.9 (1.3)	31.0 (1.5)	21.2 (1.3)	24.0 (1.4)	31.0 (1.5)
	–0.25	7.9 (0.9)	9.6 (0.9)	17.1 (1.2)	7.7 (0.8)	9.2 (0.9)	17.4 (1.2)
	–0.5	2.9 (0.5)	4.7 (0.7)	7.2 (0.8)	2.8 (0.5)	4.0 (0.6)	7.9 (0.9)

Table 4

Average computational time (in seconds).

K	$n = 500$	$n = 1000$	$n = 2000$
1000	4.10 (0.29)	7.14 (0.79)	10.20 (1.22)
10000	36.55 (5.40)	62.74 (1.24)	102.72 (12.57)

Table 5

Type I error and power (%) of the EM test and proposed test for B-Model I at 0.05 and 0.1 levels. (Standard errors are shown in parenthesis.)

		EM Test		Proposed Test	
n	***τ***	**size 0.05**	**size 0.1**	**size 0.05**	**size 0.1**
500	0	4.6 (0.7)	9.4 (0.9)	5.2 (0.3)	10.0 (0.4)
	0.1	6.9 (0.8)	12.6 (1.0)	21.2 (1.3)	31.1 (1.5)
	0.5	89.6 (0.9)	93.3 (0.8)	100 (0)	100 (0)
1000	0	5.4 (0.7)	10.4 (1.0)	5.0 (0.3)	10.1 (0.4)
	0.1	10.4 (1.0)	18.6 (1.2)	41.3 (1.6)	52.5 (1.6)
	0.5	99.8 (0.1)	100 (0)	100 (0)	100 (0)

Table 6

Misclassification rates (%) of the identified subgroups using Zhao et al. (2013)'s method and our method. (Standard errors are shown in parenthesis.)

		Zhao et al. (2013)			Proposed Method		
n	***τ***	**B-Model I**	**B-Model II**	**B-Model III**	**B-Model I**	**B-Model II**	**B-Model III**
500	0.1	36.3 (0.6)	39.4 (0.6)	42.6 (0.6)	26.3 (1.4)	28.3 (1.4)	33.3 (1.5)
	0.5	13.5 (0.3)	13.5 (0.3)	17.6 (0.4)	4.8 (0.7)	4.7 (0.7)	12.4 (1.0)
1000	0.1	32.2 (0.5)	34.8 (0.6)	37.0 (0.5)	20.9 (1.3)	23.0 (1.3)	30.0 (1.5)
	0.5	10.9 (0.2)	11.2 (0.2)	13.3 (0.3)	2.5 (0.5)	3.0 (0.2)	9.1 (0.9)

Table 7

Results for sample size calculation. Here, n is the required sample size given by our procedure. Empirical power of the test under the calculated sample size is reported based on 1000 data replications.

$\mu(X)$		$1+X$		$1-X^2$		$1+\sin(\widetilde{\pi}X)$	
τ	θ_0	n	Power	n	Power	n	Power
0.1	0	2992	91.3	4054	90.0	5440	91.1
	0.5	6034	91.8	8972	94.3	10924	90.6
	−0.5	2042	90.8	2726	90.2	3514	91.7
0.25	0	480	91.4	650	88.9	872	88.5
	0.5	966	91.8	1436	94.2	1748	89.7
	−0.5	328	90.8	436	89.1	564	88.7
0.5	0	120	87.2	164	85.6	218	85.4
	0.5	242	88.4	360	92.7	438	88.5
	−0.5	82	87.6	110	85.3	142	89.0

Table 8

Results for sample size calculation based on the method in Brookes et al. (2004). Here, n is the required sample size. Empirical power of the proposed test under the calculated sample size is reported based on 1000 data replications.

$\mu(X)$		$1+X$		$1-X^2$		$1+\sin(\widetilde{\pi}X)$	
τ	θ_0	n	Power	n	Power	n	Power
0.1	0	1580	63.9	1840	57.0	2318	52.3
0.25	0	254	61.1	296	51.0	372	48.2
0.5	0	64	52.4	74	41.6	94	38.2

Table 9

Required sample sizes for detecting a subgroup with an enhanced treatment effect τ based on the AIDS study data.

treatment effect τ	sample size n
40	2392
60	1064
80	598
100	384