Statistica Sinica **33** (2023), 1483-1505
doi:https://doi.org/10.5705/ss.202021.0253

SCORE TESTS WITH INCOMPLETE COVARIATES AND HIGH-DIMENSIONAL AUXILIARY VARIABLES

Kin Yau Wong and Jiahui Feng

The Hong Kong Polytechnic University

Abstract: Analyses of modern biomedical data are often complicated by missing values. When variables of interest are missing for some subjects, it is desirable to use observed auxiliary variables, which are sometimes high dimensional, to impute or predict the missing values in order to improve the statistical efficiency. Although many methods have been developed for prediction using high-dimensional variables, it is challenging to perform a valid inference based on such predicted values. In this study, we develop an association test for an outcome variable and a potentially missing covariate, where the covariate can be predicted using variables selected from a set of high-dimensional auxiliary variables. We establish the validity of the test under data-driven model-selection procedures. We also demonstrate the validity of the proposed method and its advantages over existing methods using extensive simulation studies and an application to a major cancer genomics study.

Key words and phrases: Association test, integrative analysis, missing data, post-selection inference, variable selection.

1. Introduction

In many clinical and epidemiological studies, investigators are interested in testing the presence of an association between an outcome variable and covariates of interest. In practice, such association analyses are often complicated by missing data, arising because of costs or other constraints. The problem of missing data is especially prevalent in large-scale genomic studies, where multiple types of genomic data are collected on a large number of subjects, often for different locations and periods. For example, in The Cancer Genome Atlas (TCGA) (`https://cancergenome.nih.gov/`), over 10,000 subjects with 33 cancer types are measured for multiple types of genomic data, including DNA alterations, RNA expressions, and protein expressions, but protein expressions are not measured for a substantial number of subjects. As another example, in the Trans-Omics for Precision Medicine (TOPMed) program (`https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed`), whole-genome se-

Corresponding author: Kin Yau Wong, Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong. E-mail: kin-yau.wong@polyu.edu.hk.

quencing data are available for hundreds of thousands of subjects, but other types of genomic data, such as RNA sequencing, methylation, and metabolites, are available for only tens of thousands of subjects, or fewer.

A naive approach to handling missing data is to perform a complete-case analysis, where subjects with missing data are discarded. Such an approach is obviously inefficient when subjects with missing data are measured on some relevant auxiliary variables, because information on the auxiliary variables would be discarded. An alternative approach is imputation, where the missing values are imputed by plausible values based on the observed data, and conventional methods are then applied to the imputed data set. However, although estimation based on imputed data may be more efficient than a complete-case analysis, conventional inferential procedures based on (singly) imputed data are, in general, invalid. More sophisticated statistical methods for handling missing data involve modeling the missing-data mechanism or the variables with missing values; see Little and Rubin (2019) for detailed discussions.

Although regression analysis with missing data has been studied extensively, association testing for incomplete data has received relatively less attention. To perform association tests with missing genotype data, Hu et al. (2015) considered a score test based on imputed genotype data, and proposed a variance estimator that properly accounts for the differential quality between the observed and the imputed genotypes. Under outcome-dependent sampling designs, Derkach, Lawless and Sun (2015) and Lawless (2018) proposed modeling the variable with missing values, and studied the score test based on the full likelihood. Under extreme phenotype sampling designs in genetic association studies, Bjørnland et al. (2018) considered a similar model-based score test and a complete-case score test based on the conditional likelihood, given the sampling mechanism. Wong, Zeng and Lin (2019) proposed modeling the variable with missing values semiparametrically, and developed a score test that is robust against a misspecification of the missing-variable model. In all existing works, the observed variables are low dimensional, and the methods cannot be readily extended to accommodate high-dimensional data.

Therefore, we consider an association test between an outcome of interest and an incomplete covariate, where the incomplete covariate may be associated with potentially high-dimensional auxiliary variables. We consider a missing-at-random scenario, where the missing mechanism may depend on the outcome of interest and the observed covariates, and a complete-case analysis or a simple imputation approach is, in general, invalid. We propose selecting a subset of the auxiliary variables, and fitting a regression model of the covariate of interest

against the selected variables. Then, we perform the score test for the covariate effect in the outcome model under the full likelihood, which includes both the outcome and the covariate models. We show that the proposed procedure, though derived by assuming a prespecified covariate model, is valid, even when the selection event of the auxiliary variables is random.

The current problem is inherently a post-selection inference problem, where we wish to perform an inference using a model selected based on the observed data. It is well known that, in general, conventional inferential procedures, such as the F-test and the t-test, on a selected model are invalid, because the parameters to be estimated or tested arise from a data-driven model-selection procedure and are "random." There is a rapidly growing body of literature on post-selection inference. One approach is to perform a conditional inference for the model parameters, given the model-selection event (Lee et al. (2016); Tibshirani et al. (2016); Heller et al. (2018); Tian, Loftus and Taylor (2018)). This approach depends on distributional assumptions, and is applicable only when the model is selected using a prespecified formal selection procedure, such as forward selection or the lasso. An alternative approach is to develop uniformly valid inferential procedures that can be used after arbitrary model selection (Berk et al. (2013); Bachoc, Leeb and Pötscher (2019); Kuchibhotla et al. (2020)). Such procedures are based on uniform tail probability inequalities, and thus are often conservative.

The proposed approach is akin to the uniform approach, which is not restricted to a specific model-selection procedure, and we make no assumptions on the correctness of the selected model. Nevertheless, an essential difference between the current framework and those considered in the literature is that the selected model, that is, the covariate model, is only of secondary interest, and the parameter of interest does not vary with the selected model. As a result of this special structure, the variability of the model selection does not affect the asymptotic distribution of the score statistic. Therefore, unlike existing methods that are potentially conservative, the proposed score test is as efficient as the standard score test that treats the selected covariate model as prespecified.

There is a related body of literature on high-dimensional inference based on debiased estimators or decorrelated score functions (van de Geer et al. (2014); Zhang and Zhang (2014); Ning and Liu (2017)). These approaches conduct inferences on the parameters in the full, high-dimensional regression model, in contrast to post-selection approaches, which focus on an inference of a selected model. One may prefer these high-dimensional approaches because the full model is considered to be more scientifically relevant than the selected model. Nevertheless, in the current framework, this potential advantage is not pertinent, because

the high-dimensional model is only of secondary interest, and the parameter of interest remains the same, regardless of whether the full or reduced model is fit. Because these high-dimensional approaches require sparsity assumptions on the true model and involve selecting multiple tuning parameters, they are not considered in this paper.

The rest of this paper is structured as follows. In Section 2, we formulate the model and develop a post-selection score test. In Section 3, we establish the asymptotic properties of the proposed score test. In Section 4, we report the results from our simulation studies. In Section 5, we provide an application to a data set from TCGA. Section 6 concludes the paper. Technical details are relegated to the Appendix.

2. Model and the Post-Selection Score Test

Consider an outcome of interest Y, a covariate of interest S, a vector of other covariates $\boldsymbol{X}$, and a potentially high-dimensional vector of auxiliary variables $\boldsymbol{A}$. For example, in genomic studies, Y may represent a disease phenotype, S may represent a genomic variable of interest, $\boldsymbol{X}$ may represent clinical or demographic variables, and $\boldsymbol{A}$ may represent other types of genomic or environmental variables collected in the study. The vector of covariates $\boldsymbol{X}$ includes a constant component of one. Assume that

$$Y \mid (\boldsymbol{X}, S) \sim F_Y(; \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S), \tag{2.1}$$

where $\boldsymbol{\alpha}$ and β are regression parameters, and F_Y is a distribution function such that, for some known function $\mu(\cdot)$, $\mathrm{E}[\{Y - \mu(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S)\}(\boldsymbol{X}^{\mathrm{T}}, S)^{\mathrm{T}}] = \mathbf{0}$ at the true values of $\boldsymbol{\alpha}$ and β. This formulation includes as special cases the linear regression model, with $\mu(x) = x$, and the logistic regression model, with $\mu(x) = e^x/(1 + e^x)$. The parameter β captures the effect of S on Y, given $\boldsymbol{X}$. In cancer genomic studies, we typically set $\boldsymbol{X}$ to be clinical or demographic variables, and do not include mediator variables in the effect of S on Y (such as downstream variables of S) in $\boldsymbol{X}$. In this case, β represents the total effect of S after accounting for the clinical/demographic covariates. We do not assume an explicit model for S, but allow an arbitrary association structure with $(\boldsymbol{X}, \boldsymbol{A})$. Because the major purpose of fitting the model of S is to predict missing S values, we can set $\boldsymbol{A}$ to be (potential) predictive variables of S.

Suppose that S may be missing, and let R be the indicator of whether S is observed. Specifically, $R = 1$ if S is observed, and $R = 0$ otherwise. We assume that R is conditionally independent of $(S, \boldsymbol{A})$, given $(Y, \boldsymbol{X})$. This missing

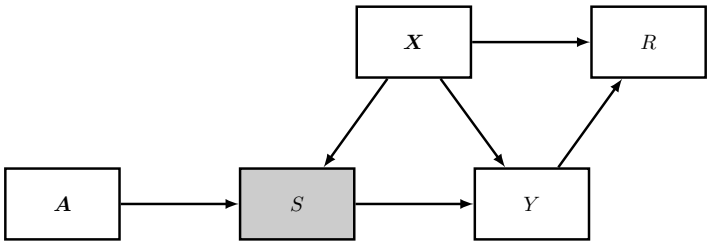

Figure 1. Relationships among the observed and incomplete variables.

mechanism is common in two-phase studies, where the outcome Y and the basic covariates $\boldsymbol{X}$ are measured for all subjects in the first phase, and subjects with a certain outcome or covariate values are selected to be measured for an expensive covariate S in the second phase. We do not allow R to depend directly on $\boldsymbol{A}$, because the auxiliary variables, though completely observed, may not be selected into the model of S. If R depends on a component of $\boldsymbol{A}$ that is associated with S and is not selected, then the missing mechanism becomes not at random. For a sample of size n, the observed data consist of $\{(Y_i, \boldsymbol{X}_i, \boldsymbol{A}_i, R_i S_i, R_i) : i = 1, \ldots, n\}$. The assumed relationships among these variables are illustrated in Figure 1.

We wish to test the null hypothesis $H_0 : \beta = 0$. Because of the missing data, we propose fitting a working model of S on $(\boldsymbol{X}, \boldsymbol{A})$, and adopt a score test based on the full model (including both models of Y and S). Fitting a working model of S against $(\boldsymbol{X}, \boldsymbol{A})$ allows us to use information about the missing S-values contained in the auxiliary variables and, in general, is more efficient than ignoring the auxiliary variables. We consider the score test rather than the Wald test or likelihood-ratio test, because it involves estimation only under the null hypothesis, whereas the other two tests involve estimation under the alternative hypothesis. Note that estimation under the alternative hypothesis is more challenging, because the likelihood generally involves an integration without a closed-form expression.

Because $\boldsymbol{A}$ is potentially high dimensional, maximum likelihood estimation for the model of S may be infeasible. In addition, the model of S is only of secondary interest, so a full specification of the model may not be necessary. Therefore, we propose a two-step approach. In the first step, we select a low-dimensional subset of $\boldsymbol{A}$ into the model of S, and in the second step, we perform a score test based on the model of Y and a working model of S.

In the first step, we perform variable selection on $\boldsymbol{A}$. Let $\mathcal{K}^*$ be a general model-selection operator, such that for an m-vector of outcome variables $\mathcal{Y}$ and

an $(m \times p)$-matrix of covariates $\mathcal{Z}$, $\mathcal{K}^*(\mathcal{Y}, \mathcal{Z}) : \mathbb{R}^m \times \mathbb{R}^{m\times p} \to \mathcal{C}_p$, where $\mathcal{C}_p$ is the collection of subsets of $\{1, \ldots, p\}$. For example, for marginal screening (Fan and Lv (2008); Fan and Song (2010)) with a quantitative outcome variable and standardized $\mathcal{Z}$, $\mathcal{K}^*$ can be defined as $\mathcal{K}^* : (\mathcal{Y}, \mathcal{Z}) \mapsto \{j : |\mathcal{Y}^{\mathrm{T}}\mathcal{Z}_j| > \lambda\}$, where λ is a tuning parameter, and $\mathcal{Z}_j$ is the jth column of $\mathcal{Z}$. Likewise, for the lasso (Tibshirani (1996)),

$$\mathcal{K}^* : (\mathcal{Y}, \mathcal{Z}) \mapsto \left\{ j : \widehat{\boldsymbol{\gamma}}_j \neq 0, \text{ where } (\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_p)^{\mathrm{T}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left(\|\mathcal{Y} - \mathcal{Z}\boldsymbol{\gamma}\|^2 + \lambda \|\boldsymbol{\gamma}\|_1 \right) \right\}.$$

We use this operator to select a model for S based on the residual $S - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}} \boldsymbol{X}$ and $\boldsymbol{A}$, where $\widehat{\boldsymbol{\gamma}}_X \equiv (\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}})^{-1} \sum_{i=1}^n R_i \boldsymbol{X}_i S_i$ is the least-squares estimator of S on $\boldsymbol{X}$ using the subjects with $R = 1$. The selected components of $\boldsymbol{A}$ are $\mathcal{K}^*(\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, \mathcal{A})$, where $\mathcal{S}$ is a vector that consists of $\{S_i : R_i = 1\}$, and $\mathcal{X}$ and $\mathcal{A}$ are matrices that consist of rows of $\{\boldsymbol{X}_i : R_i = 1\}$ and $\{\boldsymbol{A}_i : R_i = 1\}$, respectively. For simplicity of presentation, we write $\mathcal{K}^* = \mathcal{K}^*(\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, \mathcal{A})$ and let $\mathcal{K}$ be the observed value of $\mathcal{K}^*$.

Let $\boldsymbol{W}_{\mathcal{K}}$ denote a vector consisting of $\boldsymbol{X}$ and the components of $\boldsymbol{A}$ indexed by $\mathcal{K}$. In the second step, we fit model (2.1) and the working model $S = \boldsymbol{\gamma}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}} + \delta$ under the null hypothesis H_0, where δ is a mean-zero error term, and $\boldsymbol{\gamma}_{\mathcal{K}}$ is a vector of regression parameters. In particular, we estimate $\boldsymbol{\gamma}_{\mathcal{K}}$ by $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}} \equiv (\sum_{i=1}^n R_i \boldsymbol{W}_{\mathcal{K},i} \boldsymbol{W}_{\mathcal{K},i}^{\mathrm{T}})^{-1} \sum_{i=1}^n R_i \boldsymbol{W}_{\mathcal{K},i} S_i$, the least-squares estimator using the subjects with observed S-values. Let $\widehat{\boldsymbol{\alpha}}$ be the Z-estimator of $\boldsymbol{\alpha}$ under H_0, such that $\sum_{i=1}^n \{Y_i - \mu(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{X}_i)\} \boldsymbol{X}_i = 0$. The (scaled) score statistic for β is

$$U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) = \frac{1}{n^{1/2}} \sum_{i=1}^n \{Y_i - \mu(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{X}_i)\} \{R_i S_i + (1 - R_i) \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}\}.$$

Note that this coincides with the imputation-based score statistic, that is, the score statistic when the missing values of S are imputed using the estimated mean $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}}$.

To obtain an asymptotic size-α test, we need to derive the asymptotic distribution of $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})$ under H_0. This is highly challenging, because the model-selection event $\{\mathcal{K}^* = \mathcal{K}\}$ is random, and the usual arguments based on the Taylor's series expansion of the score statistic do not apply. Nevertheless, as we establish in Section 3, $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})$, properly scaled by a variance term that can be consistently estimated by an empirical sum-of-squares estimator, is asymptotically normal. In particular, the variance term resembles that derived from the usual Taylor's series expansion on the score statistic. Let $\boldsymbol{\alpha}_0$ be the true value of

$\boldsymbol{\alpha}$, and for a given selected model $\mathcal{K}$, define $\boldsymbol{\gamma}_{0\mathcal{K}} \equiv \arg\min_{\boldsymbol{\gamma}} \mathrm{E}\{R(S-\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^2\}$ as the true value of $\boldsymbol{\gamma}_{\mathcal{K}}$. Let $\boldsymbol{I}_{\alpha\alpha} = \mathrm{E}\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}$, $\boldsymbol{I}_{\beta\alpha} = -\mathrm{E}[\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}\{RS+(1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\}]$, $\boldsymbol{I}_{\gamma\gamma} = \mathrm{E}(R\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}})$, and $\boldsymbol{I}_{\beta\gamma} = \mathrm{E}[\{Y-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\}(1-R)\boldsymbol{W}_{\mathcal{K}}]$, where μ' denotes the first derivative of μ. If the model $\mathcal{K}$ is prespecified, then the Taylor's series expansion of $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ at $(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_{0\mathcal{K}})$ yields

$$U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \Big[\{Y_i - \mu(\alpha_0^{\mathrm{T}}\boldsymbol{X}_i)\}\big\{R_iS_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\big\} + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}R_i(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})\Big] + o_p(1), \tag{2.2}$$

under regularity conditions. Based on this expansion, we can estimate the asymptotic variance of $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ by $\widehat{\sigma}^2(\mathcal{K}) = n^{-1}\sum_{i=1}^{n}\{\widehat{\sigma}_i(\mathcal{K}) - \overline{\sigma}(\mathcal{K})\}^2$, where

$$\widehat{\sigma}_i(\mathcal{K}) = \{Y_i - \mu(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{X}_i)\}\big\{R_iS_i + (1-R_i)\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i} + \widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\big\} + \widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}R_i(S_i - \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}),$$

$\overline{\sigma}(\mathcal{K}) = n^{-1}\sum_{i=1}^{n}\widehat{\sigma}_i(\mathcal{K})$, and $\widehat{\boldsymbol{I}}_{\alpha\alpha}$, $\widehat{\boldsymbol{I}}_{\beta\alpha}$, $\widehat{\boldsymbol{I}}_{\beta\gamma}$, and $\widehat{\boldsymbol{I}}_{\gamma\gamma}$, are the empirical counterparts of $\boldsymbol{I}_{\alpha\alpha}$, $\boldsymbol{I}_{\beta\alpha}$, $\boldsymbol{I}_{\beta\gamma}$, and $\boldsymbol{I}_{\gamma\gamma}$, respectively, with the expectations replaced by the empirical means and true parameters replaced by the estimators. We can show that, even though this variance term is derived based on fixed $\mathcal{K}$, $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})/\widehat{\sigma}(\mathcal{K}^*)$ converges to the standard normal distribution under H_0. Therefore, for an asymptotic size-α test, we reject H_0 if $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})^2/\widehat{\sigma}^2(\mathcal{K}^*) > \chi^2_{1,\alpha}$.

The proposed test does not require correct specifications of the models of Y and S. For the outcome model, we require only that $\mathrm{E}[\{Y-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\}(\boldsymbol{X}^{\mathrm{T}}, S)^{\mathrm{T}}] = \boldsymbol{0}$ under the null hypothesis, because an empirical variance estimator is used instead of a model-based estimator. For the covariate model, as discussed in Section 3, we require the association structure between S and $\boldsymbol{X}$ to be correctly specified, but allow an arbitrary association between S and $\boldsymbol{A}$; in general, a correct specification of the association between S and $\boldsymbol{X}$ is needed (Derkach, Lawless and Sun (2015); Lawless (2018)). The association structure between S and $\boldsymbol{A}$ affects the power of the test, but not its validity under the null hypothesis.

3. Asymptotic Properties of the Post-Selection Score Test

For any $\mathcal{K}$, let $\boldsymbol{\gamma}_{0X}$ and $\boldsymbol{\gamma}_{0A,\mathcal{K}}$ be the subvectors of $\boldsymbol{\gamma}_{0\mathcal{K}}$ that correspond to $\boldsymbol{X}$ and the selected components of $\boldsymbol{A}$, respectively. Define

$$\sigma_1^2(\mathcal{K}) = \mathrm{Var}\big[\epsilon\big\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{X}\big\}\big]$$

$$\begin{aligned}\sigma_2^2(\mathcal{K}) &= \mathrm{Var}\Big[(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\{\mathrm{E}(\epsilon \mid R, \boldsymbol{X})\boldsymbol{X} - \mathrm{E}(\epsilon\boldsymbol{X} \mid R)\} \\ &\quad + \{\mathrm{E}(\epsilon \mid R, \boldsymbol{X}) - \mathrm{E}(\epsilon \mid R)\}\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K}} \\ &\quad + \{\mathrm{E}(\epsilon \mid R, \boldsymbol{X}) + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K}}\}R(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})\Big] \\ \sigma_3^2(\mathcal{K}) &= \mathrm{Var}\Big\{(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\mathrm{E}(\epsilon\boldsymbol{X} \mid R) + \mathrm{E}(\epsilon \mid R)\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K}}\Big\},\end{aligned}$$

where $\epsilon = Y - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})$, and let $\sigma^2(\mathcal{K}) = \sum_{k=1}^{3}\sigma_k^2(\mathcal{K})$. Let $\|\cdot\|_{\psi_\xi}$ be an Orlicz norm, such that $\|X\|_{\psi_\xi} = \inf\{\eta > 0 : \mathrm{E}(e^{|X|^\xi/\eta^\xi}) \le 2\}$, and let $\|\cdot\|$ be the Euclidean norm. We assume the following conditions. Some conditions involve a generic positive constant M.

(C1) For some $\xi \in (0, 2]$, $\|Y\|_{\psi_\xi} + \|S\|_{\psi_\xi} + \max_j \|A_j\|_{\psi_\xi} < M$. The covariate $\boldsymbol{X}$ is bounded, such that $P(\|\boldsymbol{X}\| < M) = 1$. Furthermore, the estimator $\widehat{\boldsymbol{\alpha}}$ is strongly consistent under $\beta = 0$, $\mu(\cdot)$ is twice continuously differentiable, and $\lambda_{\min}[\mathrm{E}\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}] > M^{-1}$, where $\lambda_{\min}(C)$ denotes the minimum eigenvalue of the matrix C.

(C2) There exists a sequence of collections of models Ω_n, such that $P(\mathcal{K}^* \in \Omega_n) \to 1$, $\sup_{\mathcal{K}\in\Omega_n}|\mathcal{K}| = O(n^\tau)$, and $\log|\Omega_n| = O(n^\kappa)$, where τ and κ are constants that satisfy $\tau < 4\xi/(5\xi + 12)$, $5\tau/4 + 3\kappa/\xi < 1$, and $\tau + 4\kappa/\xi < 1$, and $|\mathcal{C}|$ denotes the cardinality of the set $\mathcal{C}$. In addition, $\inf_{\mathcal{K}\in\Omega_n}\lambda_{\min}\{\mathrm{E}(R\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}})\} > M^{-1}$, $\sup_{\mathcal{K}\in\Omega_n}\mathrm{E}\{(\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^4\} < M$, and $\inf_{\mathcal{K}\in\Omega_n}\sigma^2(\mathcal{K}) > M^{-1}$.

(C3) The probability $P(R = 1 \mid Y, \boldsymbol{X}) > M^{-1}$, almost surely.

(C4) Under $\beta = 0$, the residual $(S - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X})$ and $\boldsymbol{X}$ are independent, and $\boldsymbol{A}$ is independent of $(Y, \boldsymbol{X})$.

(C5) The models selected based on the estimated residuals $(S_i - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}}\boldsymbol{X}_i)_{i:R_i=1}$ and the actual residuals $(S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)_{i:R_i=1}$ are such that

$$P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, \mathcal{A}) \neq \mathcal{K}^*(\mathcal{S} - \mathcal{X}\boldsymbol{\gamma}_{0X}, \mathcal{A})\Big\} = o(1)$$

and

$$\sup_{\mathcal{K}\in\Omega_n}\frac{P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, \mathcal{A}) = \mathcal{K}\Big\}}{P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\boldsymbol{\gamma}_{0X}, \mathcal{A}) = \mathcal{K}\Big\}} < M.$$

(C6) For a random sample of size m, let $\widetilde{\mathcal{S}} = (S_1, \ldots, S_m)^{\mathrm{T}}$, $\widetilde{\mathcal{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)^{\mathrm{T}}$, and $\widetilde{\mathcal{A}} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m)^{\mathrm{T}}$. The random variable

$$\sup_{\mathcal{K}\in\Omega_m}\left|\frac{P\{\mathcal{K}^*(\widetilde{\mathcal{S}}-\widetilde{\mathcal{X}}\boldsymbol{\gamma}_{0X},\widetilde{\mathcal{A}})=\mathcal{K}\mid\widetilde{\mathcal{A}}\}}{P\{\mathcal{K}^*(\widetilde{\mathcal{S}}-\widetilde{\mathcal{X}}\boldsymbol{\gamma}_{0X},\widetilde{\mathcal{A}})=\mathcal{K}\}}-1\right|$$

converges in mean to zero as $m\to\infty$.

Remark 1. Condition (C1) imposes constraints on the tail probabilities of the observed variables. With $\xi=1$ or $\xi=2$, we assume each component of $(Y,S,\boldsymbol{A})$ to be sub-exponential or sub-Gaussian, respectively. To maintain a flexible model for Y, we assume that $\boldsymbol{X}$ is bounded. Desired theoretical results can be obtained by requiring only $\max_j\|X_j\|_{\psi_\xi}<M$, but additional conditions on μ would then be required. Condition (C2) allows the set of "possibly-selected models" Ω_n to grow exponentially with n, and the size of the selected model to increase at a polynomial rate of n. For example, for $\xi=2$, we allow $\sup_{\mathcal{K}\in\Omega_n}|\mathcal{K}|=O(n^{1/4})$ and $|\Omega_n|=O\{\exp(n^{1/4})\}$. Note that if the model-selection procedure yields consistent selection, then Ω_n can be chosen as a singleton set, consisting only of the true model. In our setting, we allow the model-selection event to be genuinely random, even when n increases to infinity. Condition (C3) ensures that a nonvanishing portion of subjects have observed S.

Remark 2. Condition (C4) requires that S exhibits a linear association structure with $\boldsymbol{X}$ and that $(Y,\boldsymbol{X})$ are independent of the auxiliary variables. This guarantees that $(Y,\boldsymbol{X})$ are independent of the model-selection event, which is based on the residuals in the model of S and the auxiliary variables. In cancer genomic studies, where $\boldsymbol{X}$ represents demographic variables and $\boldsymbol{A}$ represents genomic variables (e.g., gene expressions in a tumor), $\boldsymbol{X}$ and $\boldsymbol{A}$ are plausibly independent. In general, because $\boldsymbol{X}$ is low dimensional, the independence between $\boldsymbol{A}$ and $\boldsymbol{X}$ can be (approximately) achieved by projecting the components of $\boldsymbol{A}$ onto the orthogonal complement of the span of $\boldsymbol{X}$ or functions of $\boldsymbol{X}$. The independence between $\boldsymbol{A}$ and Y can be relaxed to allow some auxiliary variables not associated with S to depend on Y; the technical formulation of the relaxed condition is deferred to Appendix A. For marginal screening, the relaxed condition allows the auxiliary variables not in any models in Ω_n to depend on Y (and $\boldsymbol{X}$). Requiring the (potentially) selected auxiliary variables to be independent of Y is quite reasonable under the null hypothesis, because these variables are, in general, associated with S. If they are also associated with Y, then except at some specific parameter values, S and Y are marginally associated, and the null hypothesis does not hold.

Remark 3. Conditions (C5) and (C6) impose mild conditions on the model-selection operator. Condition (C5) requires that the model selected based on the

estimated residuals and that selected based on the actual residuals are asymptotically equal. This is easily satisfied, because the least-squares estimator $\widehat{\boldsymbol{\gamma}}_X$ is consistent. Condition (C6) requires that the marginal probability of selecting a model is asymptotically equal to the conditional probability of the same event, given the auxiliary variables. This is true of common model-selection operators, which select a model based on the association between the outcome and the covariates, and the covariates alone do not contain information about the model-selection event. We discuss the verification of these conditions under a marginal screening procedure in the Supplementary Material.

We impose conditions on the number of possibly selected models rather than on the total number of auxiliary variables, because the former is directly relevant to the asymptotic distribution of the score statistic. Nevertheless, for a given maximal selected model size $q_n \equiv \sup_{\mathcal{K}\in\Omega_n} |\mathcal{K}|$, we have

$$r_n \equiv |\Omega_n| \leq \sum_{s=1}^{q_n} \binom{p_n}{s} \leq \left(\frac{ep_n}{q_n}\right)^{q_n},$$

where p_n is the total number of auxiliary variables. The condition on r_n is satisfied if $\log p_n = O(n^{\kappa-\tau})$, with κ and τ satisfying the inequalities in condition (C2). In fact, if most auxiliary variables are only weakly associated with S, then r_n could be much smaller than the above upper bound.

We have the following results.

Theorem 1. *Under conditions* (C1)–(C6) *and* H_0, $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})/\sigma(\mathcal{K}^*)$ *converges weakly to the standard normal distribution.*

Theorem 2. *Under conditions* (C1)–(C6) *and* H_0,

$$\mathrm{E}\left\{ \sup_{\mathcal{K}\in\Omega_n} |\widehat{\sigma}^2(\mathcal{K}) - \sigma^2(\mathcal{K})| \right\} = o(1).$$

Remark 4. Theorem 1 states that the scaled score statistic, which is derived from a randomly selected model, converges in distribution to a standard normal distribution marginally. A key step in the proof is to show that the score statistic can be (asymptotically) written as a sum of independent variables that are mean zero, conditional on the model-selection event and possibly other components of the observed data. Then, we can employ the Lindeberg approach to the proof of the central limit theorem to establish the desired result. Theorem 2 states that the scaling term of the score statistic in Theorem 1 can be uniformly consistently estimated by the proposed sum-of-squares estimator over the set of

possibly selected models Ω_n.

An outline of the proof of Theorem 1 is given in Appendix B; complete proofs of Theorems 1 and 2 are given in the Supplementary Material. Combining the above results, we have the following corollary.

Corollary 1. *Under conditions* (C1)–(C6) *and* H_0, $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})/\widehat{\sigma}(\mathcal{K}^*)$ *converges weakly to the standard normal distribution.*

4. Simulation Studies

Let $\boldsymbol{X} = (X_1, \ldots, X_5)^{\mathrm{T}}$, where (X_1, X_2, X_3) are mean-zero multivariate normal variables, with $\mathrm{Cov}(X_j, X_k) = 0.5^{|j-k|}$ $(j, k = 1, 2, 3)$, $X_4 \sim$ Bernoulli(0.1), and $X_5 \sim$ Bernoulli(0.2), and X_4 and X_5 are independent of each other and (X_1, X_2, X_3). Let $\boldsymbol{A}$ be a p-vector of independent standard normal variables. We set $S = \boldsymbol{\gamma}_X^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\gamma}_A^{\mathrm{T}}\boldsymbol{A} + \boldsymbol{\gamma}_{A,2}^{\mathrm{T}}\boldsymbol{A}^2 + \delta$, where $\boldsymbol{A}^2$ is a p-vector of the squared components of $\boldsymbol{A}$, δ is standard normal, $\boldsymbol{\gamma}_X = (0.1, \ldots, 0.1)^{\mathrm{T}}$, and $\boldsymbol{\gamma}_{A,2}$ is 0.1 at the first five components, and zero elsewhere. We consider two values of $\boldsymbol{\gamma}_A$. In Setting 1, we set $\boldsymbol{\gamma}_A$ to be 0.25 at the first 20 components, and zero at the remaining components, and in Setting 2, we set $\boldsymbol{\gamma}_A$ to be 0.25 at the first 20 components, 0.02 at the subsequent 80 components, and zero at the remaining components. In Setting 1, the model is sparse, and a small number of auxiliary variables have strong effects on S. In Setting 2, the model contains a mixture of strong and weak signals from the auxiliary variables.

We consider a quantitative and a binary outcome variable Y. For the quantitative outcome, we set $Y = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S + \epsilon$, where ϵ is standard normal, and $\boldsymbol{\alpha} = (1, -1, 1, -1, 1)^{\mathrm{T}}$. For the binary outcome, we set $\mathrm{logit}\{P(Y = 1 \mid \boldsymbol{X}, S)\} = -2.2 + \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S$, where $\boldsymbol{\alpha}$ is the same as that under the linear model; the proportion of subjects with $Y = 1$ is about 15–20%. We consider two missing-data mechanisms. The first mechanism is missing completely at random (MCAR), where the missing-data status is independent of the other variables. The second mechanism is missing at random (MAR). For the quantitative outcome, an equal number of subjects at the two extreme tails of the distribution of Y are selected to have observations on S. For the binary outcome, all subjects with $Y = 1$ are selected, and a fraction of subjects with $Y = 0$ are selected to attain the desired missing proportion. We consider sample sizes of $n = 500$ and 1000 and numbers of auxiliary variables of $p = 200$, 500, 1000, 1500, and 2000. For the alternative hypothesis, we set $\beta = 2n^{-1/2}$ and $6n^{-1/2}$ for the quantitative and binary outcome variables, respectively. For each setting, we simulated 100,000 and 10,000 replicates for $\beta = 0$ and $\beta \neq 0$, respectively.

We compare the performance of five tests: (1) the standard score test using complete data only; (2) the standard score test with missing data imputed under a working linear model of S on $\boldsymbol{X}$ and components of $\boldsymbol{A}$ selected using marginal screening, where a component of $\boldsymbol{A}$ is selected if its absolute empirical correlation with $S - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}} \boldsymbol{X}$ among the subjects with complete data is larger than a certain threshold; (3) the score test based on the full likelihood with a working linear model of S against $\boldsymbol{X}$ alone; (4) the proposed test, where the working model of S is selected in the same way as in (2); and (5) the score test based on the full likelihood with a linear model of S against $\boldsymbol{X}$ and the components of $\boldsymbol{A}$ that are associated with S. We refer to methods (1)–(5) as the complete-case analysis, simple imputation method, covariate-only method, proposed method, and true model method, respectively. In the simple imputation, proposed, and true model methods, only first-order terms of $\boldsymbol{A}$ are in the working models, so none of the models is "correct." Nevertheless, according to our theory, the proposed method is still valid under such a misspecification. For the simple imputation and proposed methods, the threshold for screening is selected using the BIC. For the covariate-only and true model methods, the variance of the score statistic is estimated using the proposed empirical sum-of-squares estimator instead of the usual estimator based on the second derivative of the log-likelihood. This is for ease of comparison with the proposed method, and the two variance estimators are asymptotically equivalent. The true model method is a gold standard, but is not practical, because it requires knowledge of the relevant predictors of S.

The results under a missing proportion of 60% are plotted in Figures 2 and 3, and the results under a missing proportion of 30% are plotted in Figures S1 and S2 of the Supplementary Material; we do not present the power of methods that inflate the type-I error. The significance level is set to 0.05. Under missing at random and the linear outcome model, both the complete-case analysis and simple imputation method inflate the type-I error, because they underestimate the variance of the score statistic. The covariate-only method and the true model method preserve the type-I error; they do not involve model selection, and their validity follows from a conventional argument. The proposed method, despite involving model-selection variability, preserves the type-I error; in fact, under Setting 2, any given model is selected at most 0.006% and 3.804% of the time over all simulation replicates with sample sizes 500 and 1000, respectively. The pattern of results under missing at random and the binary outcome model are similar, but the complete-case analysis preserves the type-I error, owing to the validity of the inference based on the prospective likelihood under a case-control study and the logistic regression model (Prentice and Pyke (1979)). Under missing completely

at random, all methods preserve the type-I error.

Under the alternative hypothesis, the simple imputation method under missing at random has relatively high power, owing to its underestimation of the variance of the score statistic; this is similar for the complete-case analysis under missing at random and the logistic outcome model. When the complete-case analysis preserves the type-I error, the complete-case analysis and the covariate-only method have similar power, because neither method includes information on the auxiliary variables. As expected, the proposed method uses information about the missing data contained in the auxiliary variables, and tends to yield higher power than that of the covariate-only method. The power gain from incorporating the auxiliary variables can be small or even negative when the number of auxiliary variables p is much larger than the (effective) sample size $\sum_{i=1}^{n} R_i$. In this case, the variable selection procedure cannot effectively identify the relevant auxiliary variables. This results in many noise variables being included in the working model of S, which in turn results in a worse fit than that of the covariate-only model, which has no noise variables.

The true model method tends to have high power, because it uses the true model of S. Nevertheless, it is less powerful than the proposed method in some scenarios under Setting 2. This is because the true model contains many auxiliary variables with weak signals, and the extra information contained in these variables does not compensate for the variability in the estimation of their effects. Thus, even when the true model is known, it may be desirable to perform variable selection and retain only those variables with strong signals.

5. A Real Study

Here, we analyze a data set of patients with colorectal adenocarcinoma from TCGA (The Cancer Genome Atlas Network (2012)), available at `http://gdac.broadinstitute.org/`. The study recorded demographic and clinical data, including age at diagnosis, sex, and tumor stage, as well as genomic data, including the expressions of RNA and protein. After removing subjects with missing clinical data, the sample size is 600. The expressions of 18,068 genes, measured by RNA sequencing, are available for most subjects. The expressions of 204 proteins or phospho-proteins are available for only 78.2% of the subjects.

We focus on the association between individual protein expressions and tumor stage. We set the outcome variable to be tumor stage, dichotomized into stage I/II and stage III/IV, with respective proportions of 0.56 and 0.44. In a single analysis, we set the covariate of interest S to be the expression of a protein

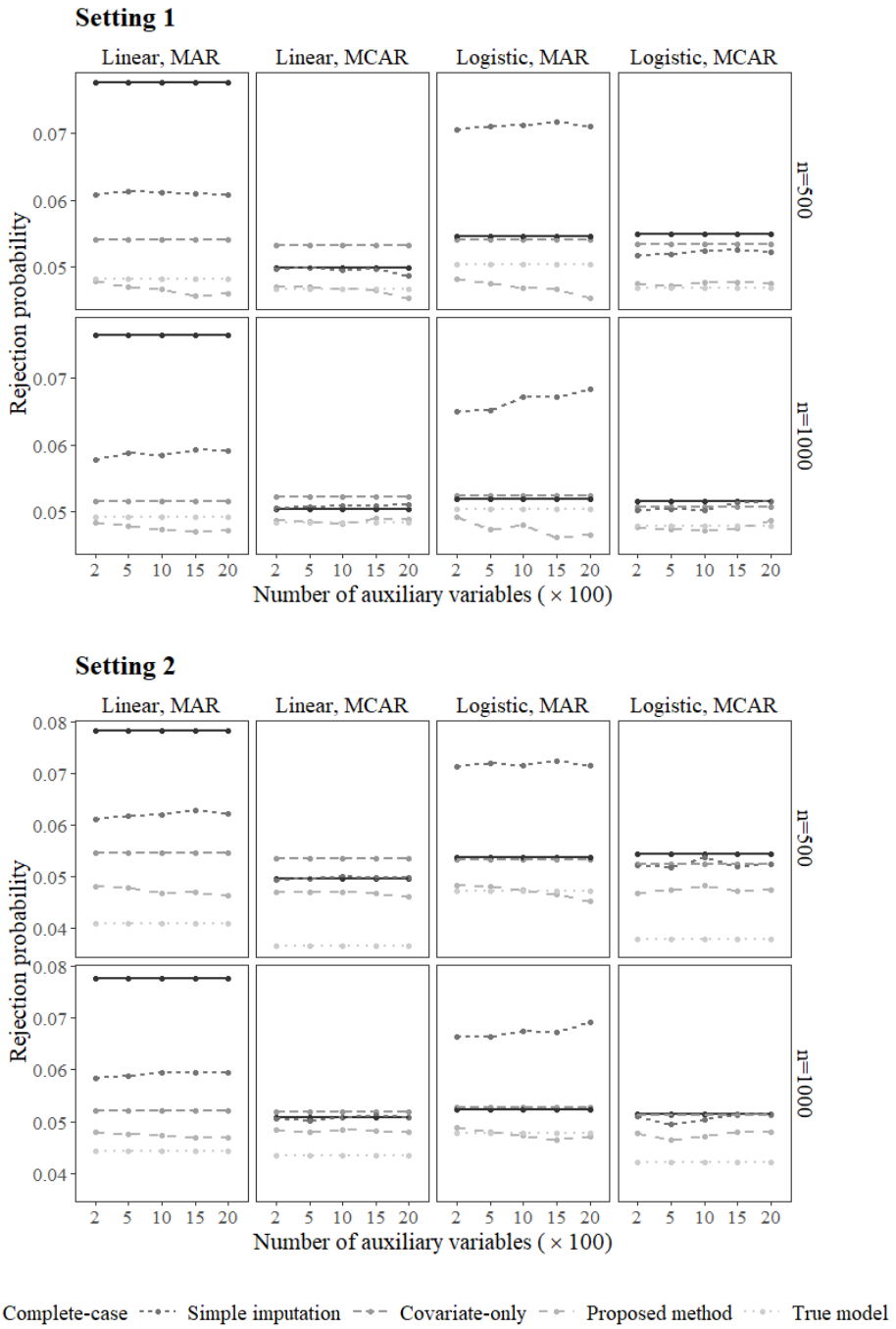

Figure 2. Rejection probabilities under a missing proportion of 60% and the null hypothesis.

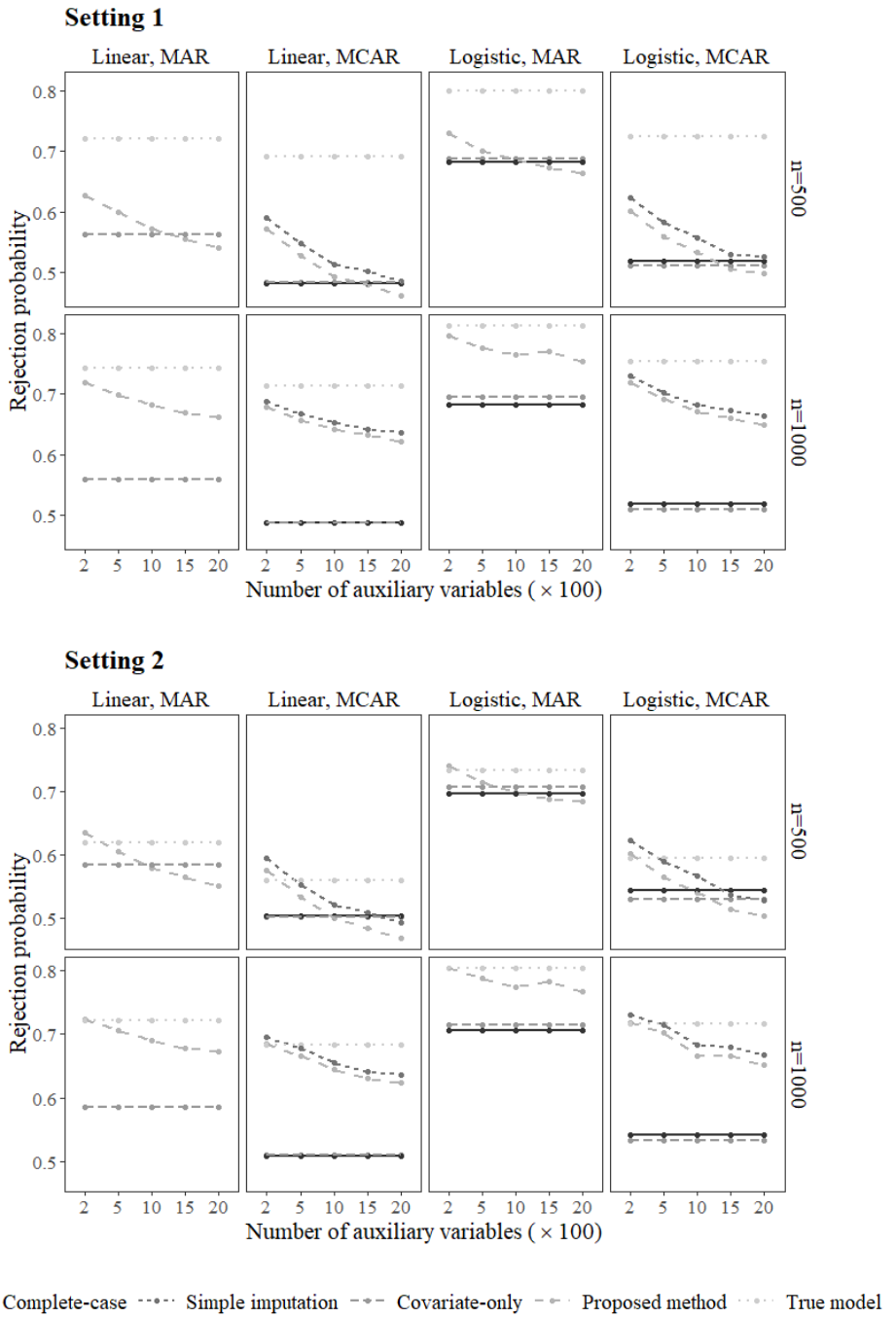

Figure 3. Rejection probabilities under a missing proportion of 60% and the alternative hypothesis.

or phospho-protein. We set sex and age at diagnosis as the covariates in $\boldsymbol{X}$, and set the gene expressions as auxiliary variables. In the resulting model, β represents the association between a protein and tumor stage for subjects with a given age and sex. Note that the auxiliary variables are plausibly independent of $\boldsymbol{X}$, as required by condition (C4). The gene expression data are incomplete, and we impute the missing values using k-nearest neighbor imputation, with $k = 10$. We set the auxiliary variables $\boldsymbol{A}$ to be the top 200 principal components of the gene expressions; they appear to be more predictive than the individual gene expressions. We perform the proposed test with the working model of S selected using the correlation-based marginal screening procedure in the simulation studies, and the screening threshold selected using the BIC. For comparison, we performed the score test using complete data only and the covariate-only method described in the simulation studies.

A total of 46 proteins were identified to be significantly associated with tumor stage at $\alpha = 0.05$ under at least one of the three tests. Of the significant proteins, 76% have smaller p-values under the proposed method than they do under the complete-case analysis, and 78% have smaller p-values under the proposed method than they do under the covariate-only method. Many of the proteins that are more significant under the proposed method have been identified as being related to the progression of colorectal adenocarcinoma; the significant proteins and relevant references are given in Table S1 of the Supplementary Material. This suggests that the proposed method is more powerful than the other two methods.

To investigate whether the power gain stems from the auxiliary variables, we inspect the relationship between the significance level and the variation explained by the gene expressions in the protein models. For a given protein, we let Z_1 and Z_2 be indicators of whether the proposed method yields a smaller p-value than that of the complete-case analysis and the covariate-only method, respectively. Let R^2 be the coefficient of partial determination of the gene expressions, that is, the percentage of variation explained by the gene expressions, given that sex and age are included in the model. Among the significant proteins, the sample correlations between Z_1 and R^2 and between Z_2 and R^2 are 0.32 and 0.22, respectively. In addition, we classify each protein into one of two groups based on whether it is more significant under the proposed method than it is under the complete-case analysis. Then, we test the difference in the mean of R^2 between the two groups using the two-sample Wilcoxon test, and the p-value is 0.0381. A similar analysis comparing the proposed method and the covariate-only method yields a p-value of 0.1271. The results suggest that proteins with a better fit of the imputation model tend to have a higher power gain, especially when compared

with the complete-case analysis.

6. Conclusion

We have considered the association test between an outcome variable and an incomplete covariate, where the missing covariate values can be imputed using high-dimensional auxiliary variables. We propose a simple two-step procedure that does not require accounting for the variability of the model selection in the first step, and prove that such a procedure is asymptotically valid. This contrasts with the conventional statistical intuition that standard inferential procedures on selected models are invalid and proper adjustments are needed (Fithian, Sun and Taylor (2014); Lee et al. (2016)). In the current setting, the model that involves variable selection is only of secondary interest. Although the fit of this model affects the power of the test, the variability of the model selection does not affect the asymptotic distribution of the score statistic.

We assume a linear working model for the incomplete covariate S, but the validity of the score test does not depend on the correctness of this model. In fact, as demonstrated in the simulation studies, a simple working model may yield higher power than the true model when the latter is complex and involves many unknown parameters. Nevertheless, we require S to exhibit a linear association with the low-dimensional covariates $\boldsymbol{X}$ in the outcome model. To relax this assumption, one may instead assume a nonparametric association between S and $\boldsymbol{X}$ (Derkach, Lawless and Sun (2015)).

We focus on the asymptotic property of the score test under the null hypothesis. Evaluating the asymptotic power of the test under contiguous alternatives is highly challenging, because the power depends on specifics of the model-selection operator. The evaluation is even more complicated when R depends on Y, in which case the missing mechanism for the data on $(S, \boldsymbol{X}, \boldsymbol{A})$ is not at random. To provide some insight into the power gain from the auxiliary variables, we evaluate the power under prespecified, fixed-dimensional sets of auxiliary variables in the Supplementary Material. Under missing completely at random, including additional auxiliary variables always increases the (asymptotic) power. In general, the power does not have a simple form under missing at random. However, our numerical evaluations suggest that the power tends to increase with the number of auxiliary variables. Note that these results are asymptotic and may not apply when the number of auxiliary variables is large compared to the sample size.

Our work can be extended in several directions. First, one may consider more general outcome models. In cancer genomic studies such as TCGA, some

outcomes of interest are (possibly censored) times to events, such as the time to cancer progression or death. It is of interest to consider semiparametric survival models for univariate or recurrent event times.

Second, in the current work, we use only a low-dimensional subset of auxiliary variables to impute the missing data, and the imputation model is fitted using least-squares estimation. It is of interest to consider a general imputation procedure that involves many auxiliary variables based on some regularized estimators, such as the lasso, elastic net, and boosting. These imputation procedures may be more accurate when many auxiliary variables are weakly associated with the incomplete covariate. Here, a theoretical development is highly challenging, because the regularized estimators may not have closed-formed expressions, and the dimension of the working model can be high.

Third, we have focused on hypothesis testing, and developed our theoretical results under the null hypothesis. One may consider estimation and inference of the outcome model. In this case, the two-step procedure is invalid, because the missing mechanism would depend on S through its dependence on Y, and estimation of the model of S using only subjects with observed data would be inconsistent. In addition, one generally needs to account for the selection variability of the model of S using the methods of, for example, Taylor and Tibshirani (2018).

Supplementary Material

The online Supplementary Material provides a discussion of the model selection under marginal screening, an evaluation of the power, proofs of all technical results, and a table and two figures presenting additional simulation and data analysis results.

Acknowledgments

This research was supported by the Hong Kong Research Grants Council grant PolyU 253042/18P.

Appendix

A. Relaxation of Condition (C4)

Let $\mathcal{M}_n \equiv \{j : j \in \mathcal{K} \text{ for some } \mathcal{K} \in \Omega_n\}$ be the collection of all "possibly selected" auxiliary variables and $\mathcal{M}_n^{\mathrm{C}}$ be its complement. Let $\mathcal{S}$ and $\mathcal{X}$ be the vector or matrix of the values of S_i and $\boldsymbol{X}_i$ for subjects with $R_i = 1$ as defined in

Section 2, and $\mathcal{A}_{\mathcal{M}_n}$ be the matrix that consists of rows of $\{\boldsymbol{A}_{\mathcal{M}_n,i} : R_i = 1\}$. For any given $(\mathcal{S}, \mathcal{X}, \mathcal{A}_{\mathcal{M}_n})$, let $\mathcal{K}(\mathcal{S}, \mathcal{X}, \mathcal{A}_{\mathcal{M}_n})$ be the collection of models that could be selected under the given data values, that is,

$$\mathcal{K}(\mathcal{S}, \mathcal{X}, \mathcal{A}_{\mathcal{M}_n}) = \Big\{\mathcal{K} : \mathcal{K}^*\{\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, (\mathcal{A}_{\mathcal{M}_n}, \widetilde{\mathcal{A}}_{\mathcal{M}_n^{\mathrm{C}}})\} = \mathcal{K} \text{ for some } \widetilde{\mathcal{A}}_{\mathcal{M}_n^{\mathrm{C}}} \in \mathbb{R}^{(\sum_i R_i)\times(p_n - |\mathcal{M}_n|)}\Big\}.$$

For any $\mathcal{K} \in \Omega_n$, define

$$\overline{\mathcal{K}} = \Big\{\mathcal{M} : \mathcal{M} \in \mathcal{K}(\widetilde{\mathcal{S}}, \widetilde{\mathcal{X}}, \widetilde{\mathcal{A}}_{\mathcal{M}_n}) \text{ for some } (\widetilde{\mathcal{S}}, \widetilde{\mathcal{X}}, \widetilde{\mathcal{A}}_{\mathcal{M}_n}) \text{ such that } \mathcal{K} \in \mathcal{K}^*(\widetilde{\mathcal{S}}, \widetilde{\mathcal{X}}, \widetilde{\mathcal{A}}_{\mathcal{M}_n})\Big\}.$$

We can understand $\overline{\mathcal{K}}$ as the collection of models that are "close" to $\mathcal{K}$: there exist auxiliary variable values $\widetilde{\mathcal{A}}_{\mathcal{M}_n}$ that are compatible with the selection of $\mathcal{K}$ as well as the selection of other elements of $\overline{\mathcal{K}}$. For marginal screening, because the selection of components of $\boldsymbol{A}_{\mathcal{M}_n}$ depends only on $(\mathcal{S}, \mathcal{X}, \mathcal{A}_{\mathcal{M}_n})$ but not $\mathcal{A}_{\mathcal{M}_n^{\mathrm{C}}}$, $\overline{\mathcal{K}}$ consists of models that include variables in $\mathcal{K}$ along with a subset of variables in $\boldsymbol{A}_{\mathcal{M}_n^{\mathrm{C}}}$.

We can replace condition (C4) in Theorems 1 and 2 and Corollary 1 by

(C4') Under $\beta = 0$, the residual $(S - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X})$ and the covariate $\boldsymbol{X}$ are independent, and $\boldsymbol{A}_{\mathcal{M}_n}$ is independent of $(Y, \boldsymbol{X})$. Also, $\sum_{\mathcal{K}\in\Omega_n} P(\mathcal{K}^* \neq \mathcal{K}, \mathcal{K}^* \in \overline{\mathcal{K}}) \to 0$.

For marginal screening, elements of $\{\overline{\mathcal{K}} : \mathcal{K} \in \Omega_n\}$ are mutually exclusive. The second part of condition (C4') is automatically satisfied, because

$$\sum_{\mathcal{K}\in\Omega_n} P(\mathcal{K}^* \neq \mathcal{K}, \mathcal{K}^* \in \overline{\mathcal{K}}) = P\left(\bigcup_{\mathcal{K}\in\Omega_n} \{\mathcal{K}^* \neq \mathcal{K}, \mathcal{K}^* \in \overline{\mathcal{K}}\}\right) \leq P(\mathcal{K}^* \notin \Omega_n) \to 0.$$

B. Outline of the Proof of Theorem 1

We outline the proof of Theorem 1 in this Appendix and relegate the complete proof to the Supplementary Material. By a version of the portmanteau theorem (Pollard (2002, p.177)), it suffices to prove that for any function g with bounded derivatives up to the third order,

$$\mathrm{E}\left[g\left\{\frac{U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})}{\sigma(\mathcal{K}^*)}\right\}\right] \to \mathrm{E}\{g(Z)\}, \tag{B.1}$$

where Z is a standard normal variable. The first step of the proof is to expand $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ as

$$\begin{aligned}
&\frac{1}{n^{1/2}} \sum_{i=1}^{n} \big\{\epsilon_i - \mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i)\big\}\big\{R_i S_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}} \boldsymbol{I}_{\alpha\alpha}^{-1} \boldsymbol{X}_i\big\} \\
&\quad + \frac{1}{n^{1/2}} \sum_{i=1}^{n} \Big[(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}} \boldsymbol{I}_{\alpha\alpha}^{-1})\big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i)\boldsymbol{X}_i - \mathrm{E}(\epsilon \boldsymbol{X} \mid R_i)\big\} \\
&\quad + \big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i) - \mathrm{E}(\epsilon \mid R_i)\big\} \boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}} \boldsymbol{A}_{\mathcal{K},i} \\
&\quad + \big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i) + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}} \boldsymbol{I}_{\gamma\gamma}^{-1} \boldsymbol{W}_{\mathcal{K},i}\big\} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}) \Big] \\
&\quad + \frac{1}{n^{1/2}} \sum_{i=1}^{n} \Big\{ (\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}} \boldsymbol{I}_{\alpha\alpha}^{-1}) \mathrm{E}(\epsilon \boldsymbol{X} \mid R_i) + \mathrm{E}(\epsilon \mid R_i) \boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}} \boldsymbol{A}_{\mathcal{K},i} \Big\} + o_p(1) \\
&\equiv \frac{1}{n^{1/2}} \sum_{i=1}^{n} U_{1i}(\mathcal{K}) + \frac{1}{n^{1/2}} \sum_{i=1}^{n} U_{2i}(\mathcal{K}) + \frac{1}{n^{1/2}} \sum_{i=1}^{n} U_{3i}(\mathcal{K}) + o_p(1),
\end{aligned}$$

where the $o_p(1)$ terms converge in mean to zero uniformly over $\mathcal{K} \in \Omega_n$. As a result, the left-hand side of (B.1) can be written as

$$\int_{\mathcal{K} \in \Omega_n} \mathrm{E}\left[g\left\{ n^{-1/2} \sum_{i=1}^{n} \frac{U_{1i}(\mathcal{K}) + U_{2i}(\mathcal{K}) + U_{3i}(\mathcal{K})}{\sigma(\mathcal{K})} \right\} \Big| \mathcal{K}^* = \mathcal{K} \right] \mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) + o(1), \tag{B.2}$$

where $\mathcal{P}_{\mathcal{K}^*}$ is the probability measure of $\mathcal{K}^*$.

The main argument of the proof is to show that $n^{-1/2} \sum_{i=1}^{n} U_{ki}(\mathcal{K})$ for $k = 1, 2, 3$ in (B.2) can in turn be replaced by normal variables. Note that conditional on $\mathcal{O}_1 \equiv (R_i, S_i, \boldsymbol{W}_{\mathcal{K},i})_{i=1,\ldots,n}$, $U_{11}(\mathcal{K}), \ldots, U_{1n}(\mathcal{K})$ are mean zero and independent. For $i = 1, \ldots, n$, let

$$\widetilde{U}_{1i}(\mathcal{K}) = \mathrm{Var}(\epsilon \mid R_i, \boldsymbol{X}_i)^{1/2} \big\{ R_i S_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}} \boldsymbol{I}_{\alpha\alpha}^{-1} \boldsymbol{X}_i \big\} Z_{1i},$$

where $Z_{11}, \ldots, Z_{1n}$ are i.i.d. standard normal random variables that are independent of the observed data. Because the first and second moments of U_{1i} and $\widetilde{U}_{1i}$ given $\mathcal{O}_1$ match and $\{\mathcal{K}^* = \mathcal{K}\}$ is implied by $\mathcal{O}_1$, the moments given $\{\mathcal{K}^* = \mathcal{K}\}$ also match. We then use Lindeberg's telescoping argument for the central limit theorem (Chung (2001, p.211)) to show that $n^{-1/2} \sum_{i=1}^{n} U_{1i}(\mathcal{K})$ in (B.2) can be replaced by $n^{-1/2} \sum_{i=1}^{n} \widetilde{U}_{1i}(\mathcal{K})$. We further show that the term can be replaced by a normal variable with mean zero and variance $\sigma_1^2(\mathcal{K})$.

Next, we show that under condition (C5), the event $\{\mathcal{K}^* = \mathcal{K}\}$ in the conditional expectation in (B.2) can be replaced by $\{\mathcal{K}_0^* = \mathcal{K}\}$, where $\mathcal{K}_0^* \equiv \mathcal{K}_0^*(\mathcal{S} - \mathcal{X}\boldsymbol{\gamma}_{0X}, \mathcal{A})$ is the selected model based on the actual residual $(S - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X})$. Then, we note that $\{\mathcal{K}_0^* = \mathcal{K}\}$ is implied by $\mathcal{O}_2 \equiv (R_i, \boldsymbol{A}_i, S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)_{i=1,\ldots,n}$, and conditional on $\mathcal{O}_2$, $U_{21}(\mathcal{K}), \ldots, U_{2n}(\mathcal{K})$ are mean zero and independent; under this conditional probability space, the random element in $U_{2i}(\mathcal{K})$ is $\boldsymbol{X}_i$. We can similarly show that $n^{-1/2}\sum_{i=1}^n U_{2i}(\mathcal{K})$ in (B.2) can be replaced by a normal variable with mean zero and variance $\sigma_2^2(\mathcal{K})$.

Finally, we show that after $n^{-1/2}\sum_{i=1}^n U_{1i}(\mathcal{K})$ and $n^{-1/2}\sum_{i=1}^n U_{2i}(\mathcal{K})$ are replaced by normal variables, the conditional expectation in (B.2) can be replaced by a marginal expectation under condition (C6). It is easy to see that $U_{31}(\mathcal{K}), \ldots, U_{3n}(\mathcal{K})$ are mean zero and independent, and thus $n^{-1/2}\sum_{i=1}^n U_{3i}(\mathcal{K})$ can be replaced by a normal variable with mean zero and variance $\sigma_3^2(\mathcal{K})$. Combining the above results, we conclude that the variable in the function g in (B.2) can be replaced by a standard normal variable, and the desired result follows.

In the conventional argument for the asymptotic distribution of the score statistic, we expand $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ as in (2.2), and the asymptotic normality of the score statistic (given $\boldsymbol{X}$ and $\boldsymbol{A}$) follows from the central limit theorem. However, conditional on the model selection event $\{\mathcal{K}^* = \mathcal{K}\}$, $(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})_{i=1,\ldots,n}$ are dependent, and the central limit theorem does not apply. In our proof, instead of relying on the independence of $(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})$'s, we establish the asymptotic normality based on the (conditional) independence and mean-zero property of functions of $\boldsymbol{X}_i$'s given the model selection event.

References

Bachoc, F., Leeb, H. and Pötscher, B. M. (2019). Valid confidence intervals for post-model-selection predictors. *The Annals of Statistics* **47**, 1475–1504.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics* **41**, 802–837.

Bjørnland, T., Bye, A., Ryeng, E., Wisløff, U. and Langaas, M. (2018). Powerful extreme phenotype sampling designs and score tests for genetic association studies. *Statistics in Medicine* **37**, 4234–4251.

Chung, K. L. (2001). *A Course in Probability Theory*. 3rd Edition. Academic Press, Massachusetts.

Derkach, A., Lawless, J. F. and Sun, L. (2015). Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika* **102**, 988–994.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567–3604.

Fithian, W., Sun, D. and Taylor, J. (2014). Optimal inference after model selection. *Preprint.* Available at *arXiv:1410.2597*.

Heller, R., Chatterjee, N., Krieger, A. and Shi, J. (2018). Post-selection inference following aggregate level hypothesis testing in large-scale genomic data. *Journal of the American Statistical Association* **113**, 1770–1783.

Hu, Y.-J., Li, Y., Auer, P. L. and Lin, D. Y. (2015). Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations. *Proceedings of the National Academy of Sciences* **112**, 1019–1024.

Kuchibhotla, A. K., Brown, L. D., Buja, A., Cai, J., George, E. I. and Zhao, L. H. (2020). Valid post-selection inference in model-free linear regression. *The Annals of Statistics* **48**, 2953–2981.

Lawless, J. (2018). Two-phase outcome-dependent studies for failure times and testing for effects of expensive covariates. *Lifetime Data Analysis* **24**, 28–44.

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

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis With Missing Data.* 3rd Edition. John Wiley & Sons, Hoboken.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45**, 158–195.

Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability.* Cambridge University Press, Cambridge.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for ℓ_1-penalized likelihood models. *Canadian Journal of Statistics* **46**, 41–61.

The Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337.

Tian, X., Loftus, J. R. and Taylor, J. E. (2018). Selective inference with unknown variance via the square-root lasso. *Biometrika* **105**, 755–768.

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

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association* **111**, 600–620.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42**, 1166–1202.

Wong, K. Y., Zeng, D. and Lin, D. Y. (2019). Robust score tests with missing data in genomics studies. *Journal of the American Statistical Association* **114**, 1778–1786.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 217–242.

Kin Yau Wong

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.

E-mail: kin-yau.wong@polyu.edu.hk

Jiahui Feng

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.

E-mail: jia-hui.feng@connect.polyu.hk

(Received July 2021; accepted February 2022)