Speeding up interval estimation for R^2-based mediation effect of high-dimensional mediators via cross-fitting*

Zhichao Xu[1,+] Chunlin Li[2,+] Sunyi Chi[1] Tianzhong Yang[3] Peng Wei[1]

Abstract

Mediation analysis is a useful tool in biomedical research to investigate how molecular phenotypes, such as gene expression, mediate the effect of an exposure on health outcomes. However, commonly used mean-based total mediation effect measures may suffer from cancellation of component-wise mediation effects of opposite directions in the presence of high-dimensional omics mediators. To overcome this limitation, a variance-based R-squared total mediation effect measure has been recently proposed, which, nevertheless, relies on the computationally intensive nonparametric bootstrap for confidence interval estimation. In this work, we formulate a more efficient two-stage cross-fitted estimation procedure for the R-squared measure. To avoid potential bias, we perform iterative Sure Independence Screening (iSIS) in two subsamples to exclude the non-mediators, followed by ordinary least squares (OLS) regressions for the variance estimation. We then construct confidence intervals based on the newly-derived closed-form asymptotic distribution of the R-squared measure. Extensive simulation studies demonstrate that the proposed procedure is hundreds of times more computationally efficient than the resampling-based method with comparable coverage probability. Furthermore, when applied to the Framingham Heart Study, the proposed method replicated the established finding of gene expression mediating age-related variation in systolic blood pressure and discovered the role of gene expression profiles in the relationship between sex and high-density lipoprotein cholesterol. The proposed cross-fitted interval estimation procedure is implemented in R package `RsqMed`.

Keywords: Confidence interval; Cross-Fitting; gene expression; Iterative sure independence screening; Mediation analysis; R^2 mediation effect measure.

*Correspondence to T. Yang (yang3704@umn.edu) and P. Wei (pwei2@mdanderson.org). +These authors contributed equally to this work. [1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, U.S.A. [2]School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A. [3]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.

1 Introduction

Recent advances in high-throughput technologies have enabled researchers to measure thousands or even millions of molecular variables, such as DNA methylation or gene expression, in a variety of tissues and cells, and to provide unprecedented opportunities to study biological mechanisms. High-dimensional mediation analysis is a critical research topic exploring the role of molecular phenotypes, such as gene expression, in mediating the effect of an exposure on health outcomes. Most existing high-dimensional mediation analysis methods rely on mean-based total mediation effect size measures and may suffer from cancellation of component-wise mediation effects of opposite directions, which are ubiquitous in the presence of high-dimensional genomics mediators (Zhao and Luo, 2022; Huang and Pan, 2016; Dai et al., 2022; Song et al., 2020; Zeng et al., 2021). As a complement, Yang et al. (2021) proposed a variance-based R-squared measure for the total mediation effect under the high-dimensional setting. The R-squared measure, denoted as R^2_{Med}, can be interpreted as the variance of the outcome variable explained by the exposure through the mediators. It can provide useful insights especially when individual molecular mediators may have mediating effects of opposite directions.

The R-squared measure, defined as $R^2_{Med} = R^2_{Y,X} + R^2_{Y,M} - R^2_{Y,MX}$, is essentially an additive function of the variance of the outcome explained by the exposure, mediators, and exposure and mediators together. Estimating variance under the high-dimensional setting is generally challenging and has been less explored than parameter estimation of individual mediation effects (Zhao and Luo, 2022; Gao et al., 2019). As demonstrated in Yang et al. (2021), the R^2_{Med} can be seriously biased when spurious mediators are included. In real data analysis with high-dimensional mediators, the identity of the true mediators is rarely known *a priori* and hard to distinguish from the spurious ones with a finite sample. The earlier work by Yang et al. (2021) used a variable selection method with the oracle property to filter out spurious variables based on half of the sample and estimated R^2_{Med} through mixed-effect models based on the remaining half. Furthermore, nonparametric bootstrap was used to compute confidence intervals, which showed satisfactory coverage probability, but was computationally intensive as each iteration of the bootstrap involved a variable selection step and an estimation step.

We herein propose a new two-stage cross-fitted interval estimation procedure for R^2_{Med} which is hundreds of times faster than the nonparametric bootstrap and can improve mediator selection against spurious correlations (Yang et al., 2021). We derive the asymptotic distribution of the R^2_{Med} estimator and demonstrate that the resulting asymptotic confidence intervals have satisfactory coverage probabilities comparable to the bootstrap-based confidence intervals in extensive simulation settings. Using this newly proposed estimation procedure, we replicated a previously established mediating relationship among age, gene expression, and systolic blood pressure (BP) (Yang et al., 2021) and investigated how gene expression could mediate the well-known relationship between sex and high-density lipoprotein cholesterol (HDL-C) (Lawlor et al., 2001; Weidner et al., 1991; Wilson et al., 1983) in the Framingham Heart Study (FHS). Lastly, we implement our new estimation procedure in the updated `RsqMed` R package on `CRAN`.

2 Methods

2.1 Mediation model and R^2 measure

A mediation model consists of the following equations,

$$\begin{aligned} \boldsymbol{M} &= \boldsymbol{\alpha} X + \boldsymbol{\xi}, \\ Y &= \gamma X + \boldsymbol{\beta}^\top \boldsymbol{M} + \varepsilon, \end{aligned} \tag{1}$$

where X is an exposure variable, Y is a response variable, and $\boldsymbol{M}$ is a vector of p potential mediators, $\boldsymbol{\xi}, \varepsilon$ are errors, and $\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma$ are regression coefficients. Moreover, the potential mediators $\boldsymbol{M} = (\boldsymbol{M}_{\mathcal{T}}, \boldsymbol{M}_{\mathcal{I}_1}, \boldsymbol{M}_{\mathcal{I}_2}, \boldsymbol{M}_{\mathcal{I}_3})$ can be partitioned into true mediators and three types of non-mediators, respectively. As illustrated in Figure 1, true mediators $M_{\mathcal{T}}$ are associated with both exposure and outcome ($\alpha_j \neq 0$ and $\beta_j \neq 0$ for $j \in \mathcal{T}$), non-mediators $\boldsymbol{M}_{\mathcal{I}_1}$ are only associated with the outcome ($\alpha_j = 0$ and $\beta_j \neq 0$ for $j \in \mathcal{I}_1$), non-mediators $\boldsymbol{M}_{\mathcal{I}_2}$ are only associated with the exposure ($\alpha_j \neq 0$ and $\beta_j = 0$ for $j \in \mathcal{I}_2$) and noise variables

$\boldsymbol{M}_{\mathcal{I}_3}$ are not associated with neither exposure nor outcome ($\alpha_j = 0$ and $\beta_j = 0$ for $j \in \mathcal{I}_3$). In high-dimensional mediation analysis, ignoring non-mediators can potentially introduce bias which distorts the mediation effect, leading to untrustworthy discoveries.

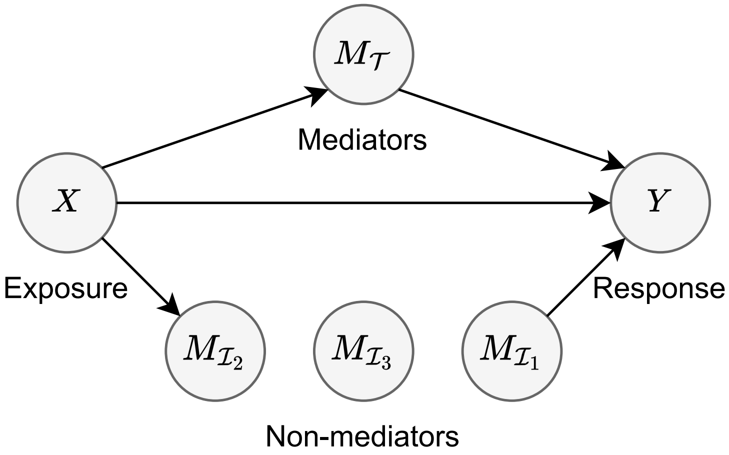

Figure 1: Graph representation of a mediation model.

In model (1), the exposure variable X influences the outcome Y directly (i.e., direct effect γ) and through the true mediators $\boldsymbol{M}_{\mathcal{T}}$ (i.e., mediation effect). The R^2_{Med} measure is defined as

$$\begin{aligned} R^2_{Med} &= R^2_{Y,X} + R^2_{Y,M} - R^2_{Y,MX} \\ &= 1 - \frac{\operatorname{Var}(Y \mid X) + \operatorname{Var}(Y \mid \boldsymbol{M}_{\mathcal{T}}) - \operatorname{Var}(Y \mid X, \boldsymbol{M}_{\mathcal{T}})}{\operatorname{Var}(Y)}, \end{aligned} \tag{2}$$

where $R^2_{Y,X} = 1 - \operatorname{Var}(Y \mid X)/\operatorname{Var}(Y)$, $R^2_{Y,M} = 1 - \operatorname{Var}(Y \mid \boldsymbol{M}_{\mathcal{T}})/\operatorname{Var}(Y)$, and $R^2_{Y,MX} = 1 - \operatorname{Var}(Y \mid X, \boldsymbol{M}_{\mathcal{T}})/\operatorname{Var}(Y)$ are the coefficients of determination of Y regressing over $\boldsymbol{M}_{\mathcal{T}}$, X, and $(X, \boldsymbol{M}_{\mathcal{T}})$, respectively.

Lemma 1. *In model (1), suppose $(X, \xi_1, \ldots, \xi_p, \varepsilon)$ are independent and normally distributed. Then*

$$R^2_{Med} = 1 - \frac{\operatorname{Var}(Y \mid X) + \operatorname{Var}(Y \mid \boldsymbol{M}_{\mathcal{S}}) - \operatorname{Var}(Y \mid X, \boldsymbol{M}_{\mathcal{S}})}{\operatorname{Var}(Y)}, \tag{3}$$

where $\mathcal{S} = \mathcal{T} \cup \mathcal{I}_1$ is the union of the true mediators and the non-mediators only associated with the outcome.

By Lemma 1 (Proof in Web Appendix A), the estimation of R^2_{Med} can be naturally decomposed to the estimation of $\mathrm{Var}(Y \mid X)$, $\mathrm{Var}(Y \mid \boldsymbol{M}_{\mathcal{S}})$, $\mathrm{Var}(Y \mid X, \boldsymbol{M}_{\mathcal{S}})$ and $\mathrm{Var}(Y)$, where the set $\mathcal{S}$ can be estimated by variable selection in the regression of Y over $(X, \boldsymbol{M})$. To infer R^2_{Med}, Yang et al. (2021) used nonparametric bootstrap to perform interval estimation, which leads to great computational challenges, especially when p is large.

In what follows, we develop a method that is faster and valid for statistical inference of the R^2_{Med} measure. Of note, we treat the effects $\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma$ in model (1) as fixed, whereas $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ were treated as random effects in Yang et al. (2021).

2.2 Cross-fitted estimation of the R^2 measure

In equation (2), R^2_{Med} is constituted by variance $V_Y = \mathrm{Var}(Y)$ and conditional variances $V_{Y|X} = \mathrm{Var}(Y \mid X)$, $V_{Y|M} = \mathrm{Var}(Y \mid \boldsymbol{M})$, and $V_{Y|MX} = \mathrm{Var}(Y \mid X, \boldsymbol{M})$. This observation suggests that the R^2_{Med} estimation can be reduced to variance estimation in regressions. To this end, we propose an estimation procedure for R^2_{Med} based on sample-splitting and cross-fitting. To proceed, suppose an independent and identically distributed sample $\mathcal{D} = \{(X_i, Y_i, \boldsymbol{M}_i) : i = 1, \ldots, n\}$ is given. The procedure is summarized in Figure 2 and detailed as follows.

- For $V_{Y|X}$, the estimate $\widehat{V}_{Y|X}$ is computed with the full sample $\mathcal{D}$ based on ordinary least squares (OLS) regression of Y on X.

- For $V_{Y|M}$ and $V_{Y|MX}$, the estimation requires high-dimensional regression involving mediator selection/screening. Motivated by Fan et al. (2012), we use cross-refitted estimates for $V_{Y|M}$ and $V_{Y|MX}$ to eliminate the potential selection bias. Specifically, the original sample $\mathcal{D}$ is randomly split into two equal subsamples $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$. Then we apply a mediator selection method based on regression of Y over $(X, \boldsymbol{M})$ with two subsamples $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$, respectively. Letting $\boldsymbol{M}_{\widehat{\mathcal{S}}^{(1)}}$ and $\boldsymbol{M}_{\widehat{\mathcal{S}}^{(2)}}$ be the selected mediators based on $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$, we estimate $\widehat{V}^{(1)}_{Y|MX}$ and $\widehat{V}^{(1)}_{Y|M}$ by refitting OLS regressions of Y over $(X, \boldsymbol{M}_{\widehat{\mathcal{S}}^{(2)}})$ and $\boldsymbol{M}_{\widehat{\mathcal{S}}^{(2)}}$ using subsample $\mathcal{D}^{(1)}$. Similarly,

$\widehat{V}^{(2)}_{Y|MX}$ and $\widehat{V}^{(2)}_{Y|M}$ are estimated using $\mathcal{D}^{(2)}$. The final estimates of $V_{Y|M}$ and $V_{Y|MX}$ are $\widehat{V}_{Y|M} = (\widehat{V}^{(1)}_{Y|M} + \widehat{V}^{(2)}_{Y|M})/2$ and $\widehat{V}_{Y|MX} = (\widehat{V}^{(1)}_{Y|MX} + \widehat{V}^{(2)}_{Y|MX})/2$.

- The final estimate of R^2_{Med} measure is $\widehat{R}^2_{Med} = 1 - \widehat{V}_{Y|X} - \widehat{V}_{Y|M} + \widehat{V}_{Y|MX}$.

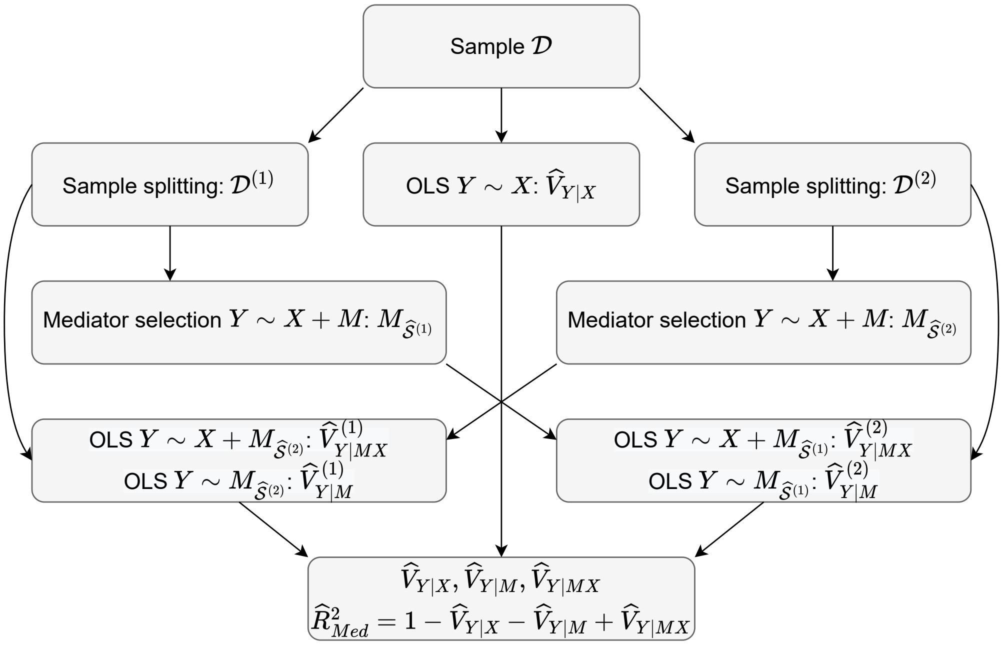
Sample $\mathcal{D}$
Sample splitting: $\mathcal{D}^{(1)}$
OLS $Y \sim X$: $\widehat{V}_{Y|X}$
Sample splitting: $\mathcal{D}^{(2)}$
Mediator selection $Y \sim X + M$: $M_{\widehat{\mathcal{S}}^{(1)}}$
Mediator selection $Y \sim X + M$: $M_{\widehat{\mathcal{S}}^{(2)}}$
OLS $Y \sim X + M_{\widehat{\mathcal{S}}^{(2)}}$: $\widehat{V}^{(1)}_{Y|MX}$
OLS $Y \sim M_{\widehat{\mathcal{S}}^{(2)}}$: $\widehat{V}^{(1)}_{Y|M}$
OLS $Y \sim X + M_{\widehat{\mathcal{S}}^{(1)}}$: $\widehat{V}^{(2)}_{Y|MX}$
OLS $Y \sim M_{\widehat{\mathcal{S}}^{(1)}}$: $\widehat{V}^{(2)}_{Y|M}$
$\widehat{V}_{Y|X}, \widehat{V}_{Y|M}, \widehat{V}_{Y|MX}$
$\widehat{R}^2_{Med} = 1 - \widehat{V}_{Y|X} - \widehat{V}_{Y|M} + \widehat{V}_{Y|MX}$

Figure 2: Cross-fitted estimation of R^2_{Med}. The sample $\mathcal{D}$ is used for estimation of $\widehat{V}_{Y|X}$ and for sample splitting, which yields $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}$. Then $\mathcal{D}^{(k)}$ is used for mediator selection $\boldsymbol{M}_{\widehat{\mathcal{S}}^{(k)}}$; $k = 1, 2$. Next, $\widehat{V}^{(1)}_{Y|MX}, \widehat{V}^{(1)}_{Y|M}$ are estimated based on the subsample $\mathcal{D}^{(1)}$ and the selected mediators $\boldsymbol{M}_{\widehat{\mathcal{S}}^{(2)}}$, and similarly for $\widehat{V}^{(2)}_{Y|MX}, \widehat{V}^{(2)}_{Y|M}$. Finally, $\widehat{V}_{Y|MX}, \widehat{V}_{Y|M}$ are estimated from $\widehat{V}^{(k)}_{Y|MX}, \widehat{V}^{(k)}_{Y|M}$; $k = 1, 2$, and $\widehat{R}^2_{Med} = 1 - \widehat{V}_{Y|X} - \widehat{V}_{Y|M} + \widehat{V}_{Y|MX}$.

2.3 Theoretical properties

In this subsection, we establish the large-sample properties of the proposed cross-fitted estimator. In particular, we derive the asymptotic normality of conditional variance estimators, which enables us to construct confidence intervals for the R^2 measure.

Theorem 1. *In model (1), suppose $(X, \xi_1, \ldots, \xi_p, \varepsilon)$ are independent and normally distributed. Let $\zeta = Y - \mathrm{E}(Y \mid \boldsymbol{M}_{\mathcal{S}})$, $\eta = Y - \mathrm{E}(Y \mid X)$ and let $\boldsymbol{A}$ be the covariance matrix of $(\varepsilon^2, \eta^2, \zeta^2, Y^2)$. Assume $\boldsymbol{A}$ is constant, $|\mathcal{S}| = |\mathcal{T} \cup \mathcal{I}_1| = o(n)$, and the mediator selection procedure satisfies $P(\widehat{\mathcal{S}}^{(k)} = \mathcal{S}) \to 1$ as $n \to \infty$ for $k = 1, 2$. Then*

$$\sqrt{n}\begin{pmatrix} \widehat{V}_{Y|MX} - V_{Y|MX} \\ \widehat{V}_{Y|X} - V_{Y|X} \\ \widehat{V}_{Y|M} - V_{Y|M} \\ \widehat{V}_Y - V_Y \end{pmatrix} \xrightarrow{d} N\left(\boldsymbol{0}, \boldsymbol{A}\right). \tag{4}$$

Consequently,

$$\sqrt{n}\frac{(\widehat{R}^2_{Med} - R^2_{Med})}{\sqrt{\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}}} \xrightarrow{d} N(0,1),$$

where $\boldsymbol{u} = (1/V_Y, -1/V_Y, -1/V_Y, (V_{Y|X} + V_{Y|M} - V_{Y|MX})/V_Y^2)$.

As suggested by Theorem 1, the estimator $\widehat{R}^2_{Med}$ is consistent and achieves the asymptotic variance of the hypothetical oracle estimator, therefore it is optimal. For statistical inference, we estimate the asymptotic covariance matrix $\boldsymbol{A}$ by the residuals of the corresponding least squares regressions, and use the plugin estimator

$$\widehat{\boldsymbol{u}} = (1/\widehat{V}_Y, -1/\widehat{V}_Y, -1/\widehat{V}_Y, (\widehat{V}_{Y|X} + \widehat{V}_{Y|M} - \widehat{V}_{Y|MX})/\widehat{V}_Y^2)$$

for $\boldsymbol{u}$. Detailed technical proofs of Lemma 1 and Theorem 1 are provided in Web Appendix A.

3 Simulation Studies

3.1 Simulation Settings

The existence of non-mediator $\boldsymbol{M}_{\mathcal{I}_1}$ and noise variables does not affect the estimation, whereas non-mediator $\boldsymbol{M}_{\mathcal{I}_2}$ can result in a biased and inconsistent estimation in high-dimensional settings (Yang et al., 2021). Therefore, we used the iterative Sure Independence Screening (iSIS) (Fan and Lv, 2008) along with the Minimax Concave Penalty (MCP) (Zhang, 2010) screening procedure (iSIS-MCP) for variable selection.

We first compared the proposed cross-fitted OLS estimation method (CF-OLS) with the previously established method (B-Mixed) (Yang et al., 2021) which estimates the R^2_{Med} measure in the mixed model framework along with bootstrap-based confidence interval. We computed the coverage probability, width of the confidence interval, bias, mean squared error (MSE), empirical standard deviation of estimator (i.e., standard deviation of the sampling distribution of the estimator based on simulation replications), variable selection accuracy, and computational efficiency in various high-dimensional settings.

For the B-Mixed method, we performed iSIS variable selection in the first half subsample and obtained point estimation and confidence intervals in the second half subsample. For each replication, the confidence interval for R^2_{Med} was computed from 500 nonparametric bootstrap resamplings. Then we obtained the coverage probability and empirical standard deviation of the estimation from 200 replications. For the CF-OLS method, within each replication iSIS was applied independently to two subsamples as illustrated in Figure 2. The asymptotic standard error, bias, MSE, true positive rate, and false positive rate were the mean of their respective estimates in the subsamples. Then the Wald confidence interval for R^2_{Med} was constructed based on the asymptotic standard error. The coverage probability and empirical standard deviation of the estimation from 200 replications were directly reported. For both methods, the width of the confidence interval, bias, MSE, true positive rate, and false positive rate were averaged across 200 replications.

The performance of the two methods was evaluated in various scenarios (A1)–(A6), where different types or numbers of non-mediators were included. In scenarios (A1)–(A2), a substantial number of noise variables $\boldsymbol{M}_{\mathcal{I}_3}$ were added, and in scenarios (A3)–(A4), numerous non-mediators $\boldsymbol{M}_{\mathcal{I}_1}$ and $\boldsymbol{M}_{\mathcal{I}_2}$ were simulated respectively. Scenarios (A5)–(A6) examined a combination of different types of non-mediators.

In each scenario, the same parameters were simulated across 200 replications so that the true R^2_{Med} remained the same. Data were simulated under model (1) at varied sample size n = 750, 1500, and 3000. The errors in model (1) independently follow the standard normal distribution, $\boldsymbol{\xi} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$ and $\varepsilon \sim N(0, 1)$. Exposure variable X was simulated from the standard normal distribution $N(0,1)$ and coefficient γ in model (1) was set to 3. Let (p_0, p_1, p_2, p_3) denote the number of true mediators, two types of non-mediators, and noise variables $(\boldsymbol{M}_{\mathcal{T}}, \boldsymbol{M}_{\mathcal{I}_1}, \boldsymbol{M}_{\mathcal{I}_2}, \boldsymbol{M}_{\mathcal{I}_3})$, respectively. The total number of variables in $\boldsymbol{M}$ was set to $p = 1500$. The maximum number of iterations for iSIS was set equal to 3. We used Bayesian information criterion (BIC) (Schwarz, 1978) to select the tuning regularization parameters in the penalized likelihood functions.

The details of simulation scenarios (A1)-(A6) can be found as follows.

- (A1) $(p_0, p_1, p_2, p_3) = (15, 0, 0, 1485)$: $\alpha_i \sim N(0, 1.5^2)$, $\beta_i \sim N(0, 1.5^2)$ for $i = 1, ..., 15$; $\alpha_i = \beta_i = 0$ for $i = 16, ..., 1500$.
- (A2) $(p_0, p_1, p_2, p_3) = (150, 0, 0, 1350)$: $\alpha_i \sim N(0, 1.5^2)$, $\beta_i \sim N(0, 1.5^2)$ for $i = 1, ..., 150$; $\alpha_i = \beta_i = 0$ for $i = 151, ..., 1500$.
- (A3) $(p_0, p_1, p_2, p_3) = (150, 1350, 0, 0)$: $\alpha_i \sim N(0, 1.5^2)$, $\beta_i \sim N(0, 1.5^2)$ for $i = 1, ..., 150$; $\alpha_i = 0, \beta_i \sim N(0, 1.5^2)$ for $i = 151, ..., 1500$.
- (A4) $(p_0, p_1, p_2, p_3) = (150, 0, 1350, 0)$: $\alpha_i \sim N(0, 1.5^2)$, $\beta_i \sim N(0, 1.5^2)$ for $i = 1, ..., 150$; $\alpha_i \sim N(0, 1.5^2)$, $\beta_i = 0$ for $i = 151, ..., 1500$.
- (A5) $(p_0, p_1, p_2, p_3) = (150, 150, 0, 1200)$: $\alpha_i \sim N(0, 1.5^2)$, $\beta_i \sim N(0, 1.5^2)$ for $i = 1, ..., 150$; $\alpha_i = 0$, $\beta_i \sim N(0, 1.5^2)$ for $i = 151, ..., 300$; $\alpha_i = \beta_i = 0$ for $i = 301, ..., 1500$.
- (A6) $(p_0, p_1, p_2, p_3) = (150, 150, 150, 1050)$: $\alpha_i \sim N(0, 1.5^2), \beta_i \sim N(0, 1.5^2)$ for $i = 1, ..., 150$; $\alpha_i = 0$, $\beta_i \sim N(0, 1.5^2)$ for $i = 151, ..., 300$; $\alpha_i \sim N(0, 1.5^2)$, $\beta_i = 0$ for $i = 301, ..., 450$; $\alpha_i = \beta_i = 0$ for $i = 451, ..., 1500$.

3.2 Simulation Results

Table 1 presents the comparison of the statistical inference under the high-dimensional setting between the CF-OLS method and the B-Mixed method. In general, CF-OLS performed reasonably well in all scenarios.

For mediator selection, two methods shared a comparable performance when iSIS-MCP was used. Generally, a high average true positive rate was achieved when the sample size was 3000. In particular, a substantial proportion of true mediators $\boldsymbol{M}_{\mathcal{T}}$ were identified in scenario (A1). Besides, iSIS-MCP controlled the average false positive rate at a low level across all scenarios. The average false positive rate increased as the sample size increased in scenarios (A3), (A5), and (A6) for both methods because $\boldsymbol{M}_{\mathcal{I}_1}$ was associated with outcome Y given X, and thus were not filtered out by iSIS. In Web Appendix B, we highlight the trade-off between true positives (i.e., selecting true mediators) and false positives (i.e., falsely selecting non-mediators). Of note, including non-mediator $\boldsymbol{M}_{\mathcal{I}_1}$ will not bias the point estimation of R^2_{Med}, as suggested by Lemma 1.

The empirical coverage probability using CF-OLS was satisfactory across all scenarios, and it yielded narrower confidence intervals compared with B-Mixed. Meanwhile, we found that the empirical standard deviation of replicated estimations of CF-OLS (i.e., from its sampling distribution) was lower than that of B-Mixed. This is because CF-OLS makes full use of the two subsamples as illustrated in Figure 2, in contrast to B-Mixed which conducts inference using only half of the data. In scenarios (A2), (A4), (A5), and (A6), relatively sizeable MSE was observed for both methods when the sample size was 750, due to overselection of $\boldsymbol{M}_{\mathcal{I}_2}$ and underselection of $\boldsymbol{M}_{\mathcal{T}}$ by iSIS. The bias and MSE improved in all scenarios with increasing sample size.

Figure 3 displays asymptotic standard errors and the empirical standard deviation of replicated estimations using the CF-OLS method in the scenarios (A1)–(A6). The asymptotic standard error is the mean value of 200 replications, and error bars represent one standard error for the mean. Generally, the asymptotic standard errors and empirical standard deviation tracked each other closely as the sample size increased from 500 to 3000. As expected, we observed a decreasing trend of the asymptotic standard errors and empirical standard deviation with increasing sample sizes.

Furthermore, as expected, CF-OLS significantly outperformed bootstrap-based B-Mixed in terms of computation. Table 1 provided the mean and standard error of the computational

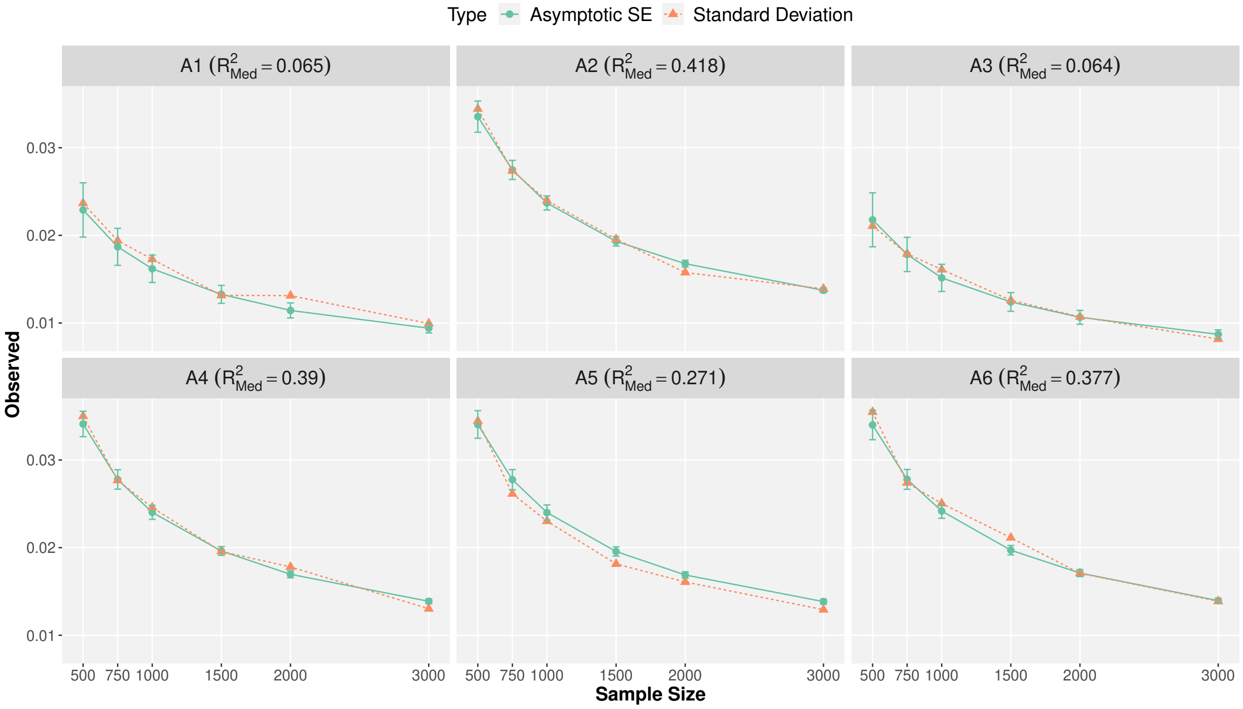

Figure 3: Plots of asymptotic standard error and empirical standard deviation of replicated estimations across 200 simulation replications using the CF-OLS method for scenarios (A1)–(A6). SE refers to standard errors. The sample size increased from 500 to 3,000. The true value of R^2_{Med} is listed within the parentheses. Error bars represent one standard error of the mean of asymptotic standard error across 200 replications in each scenario.

time measured in minutes based on 200 replications using the CF-OLS and the B-Mixed methods. For example, in scenario (A6) with a sample size of 750, CF-OLS spent about 2.42 minutes constructing one confidence interval using a single CPU core. By comparison, the B-Mixed took about 36.6 minutes to achieve the same goal using 20 cores in parallel. For all the scenarios with a sample size of 3000, the proposed method shortened the time to compute the coverage probability based on 200 replications from over 380 hours to less than 30 hours with R version 4.1.0. In practice, we found that the computational time of B-Mixed fluctuated highly, while that of CF-OLS was quite stable. The difference in computational time is very important in real data applications, advocating the use of CF-OLS. For both methods, the most time-consuming part was the variable selection step instead of the estimation step.

In Web Appendix B, we further evaluated the proposed method in additional scenarios (B1)–(B6) and (C1)–(C6). In scenarios (B1)–(B6), the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ followed the uniform distribution $U(-2, 2)$, and in scenarios (C1)–(C6), $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ followed the standard normal distribution $N(0, 1^2)$ when they were not set to 0. Overall, the coverage probability was satisfactory. When the sample size was at 3000, the variable selection procedure captured an extensive number of true mediators $\boldsymbol{M}_{\mathcal{T}}$, which gave a reasonable average true positive rate. Furthermore, the average false positive rate was controlled at a low level by eliminating most of the non-mediators $\boldsymbol{M}_{\mathcal{I}_2}$. We also found that the increased average false positive rate was due to the presence of the selected non-mediators $\boldsymbol{M}_{\mathcal{I}_1}$ in scenarios (B3), (B5), (C3), and (C5). However, it is promising that the number of selected non-mediators $\boldsymbol{M}_{\mathcal{I}_2}$ was still reasonably low, and the number of selected noise variables was nearly 0. As expected, a smaller MSE was observed with a larger sample size. Asymptotic standard errors approximated the empirical standard deviation of replicated estimations well for scenarios (B1)–(B6) and (C1)–(C6) (See Web Appendix B). To summarize, the performance of CF-OLS under various settings is satisfactory in terms of mediator selection, coverage probability, and computational efficiency.

Moreover, in Web Appendix C, we explored some alternative options for the iSIS procedure along with CF-OLS that may reduce the computational time and/or increase the

Table 1: Simulation results using the CF-OLS method and B-Mixed method for scenarios (A1)–(A6). N refers to the sample size. CP refers to coverage probability based on 200 replications. Width refers to half the width of the 95% confidence interval. SE refers to the average asymptotic standard error. SD refers to the empirical standard deviation of replicated estimations. MSE refers to mean squared error. TP refers to the average true positive rate. FP refers to the average false positive rate. True value of R^2_{Med} is listed within the parentheses. Time refers to the mean computational time in minutes for each replication and its standard error is listed within the parentheses. The computational time for CF-OLS was observed using a single CPU core. The computational time for B-Mixed was observed using 20 cores in parallel.

		CF-OLS									B-Mixed							
Scenario	**N**	**CP**	**Width**	**SE**	**Bias**	**SD**	**MSE**	**TP**	**FP**	**Time**	**CP**	**Width**	**Bias**	**SD**	**MSE**	**TP**	**FP**	**Time**
(R^2_{Med})		%	($\times 10^{-2}$)	(10^{-2})	(10^{-2})	(10^{-2})	(10^{-4})				%	($\times 10^{-2}$)	(10^{-2})	(10^{-2})	(10^{-4})			
A1	750	92.0	3.664	1.870	0.739	1.940	4.292	0.945	0.021	0.12 (0.00)	98.5	5.159	0.149	2.646	6.990	0.940	0.020	44.96 (2.27)
(0.065)	1500	93.5	2.601	1.327	0.658	1.316	2.155	0.929	0.018	3.44 (0.04)	95.0	3.615	0.236	2.084	4.377	0.923	0.015	85.09 (4.44)
	3000	93.5	1.844	0.941	0.133	0.994	1.001	0.967	0.008	4.80 (0.07)	93.0	2.591	0.138	1.491	2.230	0.968	0.008	153.49 (8.12)
A2	750	94.5	5.383	2.747	-0.032	2.736	7.450	0.403	0.001	1.98 (0.04)	95.0	7.702	-0.263	3.908	15.266	0.402	0.001	51.23 (2.83)
(0.418)	1500	92.0	3.787	1.932	0.334	1.956	3.920	0.694	0.003	5.30 (0.11)	94.0	5.353	0.355	2.647	7.097	0.696	0.003	88.22 (4.54)
	3000	94.5	2.691	1.373	-0.131	1.390	1.940	0.943	0.003	6.78 (0.04)	94.0	3.777	-0.103	1.953	3.807	0.943	0.002	149.68 (6.28)
A3	750	93.5	3.494	1.782	0.269	1.790	3.259	0.310	0.011	2.13 (0.04)	92.5	5.054	0.365	2.762	7.725	0.311	0.011	38.51 (1.56)
(0.064)	1500	95.0	2.431	1.240	0.198	1.259	1.617	0.505	0.026	5.10 (0.05)	94.0	3.390	-0.008	1.820	3.297	0.506	0.026	74.06 (2.69)
	3000	95.0	1.707	0.871	0.168	0.817	0.692	0.762	0.065	8.62 (0.10)	96.0	2.391	0.015	1.118	1.245	0.763	0.065	147.08 (4.46)
A4	750	96.0	5.445	2.778	0.029	2.769	7.630	0.130	0.025	1.47 (0.03)	93.5	7.781	-0.227	4.088	16.680	0.131	0.026	41.79 (1.54)
(0.390)	1500	95.0	3.845	1.962	-0.255	1.956	3.873	0.386	0.022	4.95 (0.08)	96.5	5.430	-0.456	2.479	6.321	0.382	0.022	72.28 (2.57)
	3000	97.0	2.720	1.388	0.113	1.303	1.702	0.724	0.001	6.78 (0.12)	95.0	3.831	-0.011	1.839	3.367	0.723	0.002	125.16 (3.89)
A5	750	96.0	5.440	2.776	0.025	2.615	6.802	0.352	0.006	1.39 (0.02)	94.5	7.758	-0.215	4.096	16.738	0.354	0.006	40.09 (1.32)
(0.271)	1500	97.0	3.834	1.956	0.183	1.814	3.309	0.578	0.018	3.10 (0.08)	95.0	5.376	0.148	2.617	6.834	0.579	0.017	73.34 (2.43)
	3000	97.0	2.714	1.385	0.046	1.292	1.664	0.879	0.051	8.88 (0.12)	95.0	3.812	-0.016	1.899	3.587	0.878	0.051	139.04 (4.40)
A6	750	96.5	5.447	2.779	0.041	2.740	7.471	0.238	0.019	2.42 (0.04)	93.5	7.765	-0.313	4.165	17.359	0.237	0.019	36.60 (1.48)
(0.377)	1500	92.5	3.863	1.971	0.052	2.113	4.447	0.400	0.034	4.14 (0.10)	95.5	5.466	-0.208	2.830	8.011	0.401	0.034	64.18 (2.49)
	3000	95.5	2.735	1.396	-0.024	1.388	1.918	0.622	0.072	8.34 (0.12)	94.5	3.837	-0.013	1.959	3.817	0.624	0.072	114.23 (3.68)

accuracy of variable selection. First, we considered increasing the maximum number of iterations for iSIS from 3 to 10 to evaluate its influence on the selection accuracy. We discovered that increasing the total number of iSIS iterations made a negligible difference in statistical inference, despite increasing computational burden in most scenarios. Then, we considered Lasso (Tibshirani, 1996), a popular replacement to MCP for sparse regression. Based on the same scenarios (A1)–(A6) in Table 1, we examined how our method performs with Lasso using the Akaike information criterion (AIC) (Akaike, 1998) for tuning the regularization parameter. Under this setting, iSIS-Lasso kept the non-mediators $\boldsymbol{M}_{\mathcal{I}_1}$ and noise variables $\boldsymbol{M}_{\mathcal{I}_3}$ at a similar level to iSIS-MCP but failed to exclude the non-mediators $\boldsymbol{M}_{\mathcal{I}_2}$. Unlike iSIS-MCP, the model selection with iSIS-Lasso suffered from a higher average false positive rate as the sample size increased. A possible reason is that the Lasso regression tends to include an extensive number of false positives (Martinez et al., 2010). Despite this, the coverage probability and bias had a minor discrepancy from those of iSIS-MCP using CF-OLS, which performed well across all scenarios.

4 Application to the Framingham Heart Study

Hypertension is a leading cause for cardiovascular disease (CVD) and mortality worldwide (Roth et al., 2018). Of the adult population worldwide in year 2010, around 1.39 billion had hypertension, whose symptom is persistently high blood pressure (BP), expressed as high systolic BP and diastolic BP (Mills et al., 2016). Its prevalence increases with chronological age, contributing to the current pandemic of cardiovascular disease (CVD) (Kearney et al., 2005). On the other hand, a higher plasma level of high-density lipoprotein cholesterol (HDL-C) is associated with a lower risk for coronary heart disease in several epidemiological studies (Castelli, 1988). A previous prospective cohort study found that the incidence and mortality of coronary heart disease among men were around 3-fold and 5-fold greater than those among women, respectively, where the difference in HDL-C was the major determinant (Jousilahti et al., 1999). Our motivation is to investigate the effect of chronological age on systolic BP and the effect of sex on HDL-C mediated by genome-wide gene expression.

We applied our method to the individuals who attended the 8^{th} and 9^{th} examinations of the FHS Offspring Cohort or the 2^{nd} and 3^{rd} examinations of the FHS Third-Generation Cohort. BP was measured as the average value of two physician BP readings (to the nearest 2 mm Hg). Then BP was adjusted according to the intake of antihypertensive medication by adding 15 mm Hg to the measurement for treated individuals (Tobin et al., 2005). HDL-C was measured from the EDTA plasma (mg/dL) and age was measured at which the subject attended the examination. The covariates were body mass index (kg/m^2); smoking status (current smoker vs. current non-smoker); drinking status (never vs. ever), and the cohort the subject belongs to (Offspring Cohort vs. Third-Generation Cohort). Age and sex were adjusted in the model while the other one was considered as the exposure variable of interest. High-throughput gene expression profiling of 17,873 genes was measured from whole blood mRNA using Affymetrix Human Exon 1.0 ST GeneChip (Joehanes et al., 2012). We extracted age, sex, covariates, and gene expression levels from the Offspring Cohort 8^{th} examination and Third-Generation Cohort 2^{nd} examination. Phenotypes were from the Offspring Cohort 9^{th} examination and Third-Generation Cohort 3^{rd} examination, which follows the establishment in Kraemer et al. (2002) that the exposure affects the mediators which in turn precedes the outcome. A total of 4,542 subjects with complete data were included in the analysis for systolic BP and 4,481 for HDL-C. For comparison, we followed Yang et al. (2021) by regressing covariates out from exposure, phenotypes, and gene expression levels to obtain the residuals for the following analyses to control for confounding effects. The descriptive statistics for the FHS samples were summarized in Web Appendix D.

Table 2 compares the results of data analysis using CF-OLS and B-Mixed. We discovered that both methods gave comparable point estimation and confidence intervals, suggesting that the new method is able to give reliable inferences. For the CF-OLS method, 8.01% of systolic BP variation could be explained by age, and 201 and 238 genes were selected in each of the two subsamples. Note that 3.22% (95% CI = (2.19%, 4.26%)) of the variance in systolic BP was attributable to the indirect effect of age through the mediation by gene expression. Similarly, 16.57% of the variance in HDL-C was explained by sex, and 8.91%

Table 2: Mediation effect size and 95% confidence interval estimated using the CF-OLS method and B-Mixed method in the Framingham Heart Study (FHS) data. N refers to the sample size. $\hat{p}_1$ and $\hat{p}_2$ refer to the number of genes selected in the first and second subsample, respectively. $\hat{p}$ refers to the number of genes selected in the B-Mixed method. 95% confidence interval listed within the parentheses for the B-Mixed method is computed over 500 bootstrap samples. Time refers to the computational time in hours. The computational time for CF-OLS is observed using a single core. The computational time for systolic BP using B-Mixed was observed using 25 cores. The computational time for HDL-C using B-Mixed was observed using 10 cores.

		CF-OLS				**B-Mixed**			
Outcome	**Exposure**	$R^2_{Mediated}$	$R^2_{Y,X}$	$\hat{p}_1/\hat{p}_2$	**Time**	$R^2_{Mediated}$	$R^2_{Y,X}$	$\hat{p}$	**Time**
Systolic BP (N=4542)	Age	0.032 (0.022, 0.043)	0.080	201/238	6.45	0.034 (0.016, 0.058)	0.107 (0.084, 0.132)	265 (196, 256)	135.54
HDL-C (N=4481)	Sex	0.089 (0.073, 0.105)	0.166	175/198	5.19	0.078 (0.060, 0.136)	0.169 (0.141, 0.194)	207 (191, 245)	232.46

(95% CI = (7.33%, 10.49%)) of the HDL-C variation could be explained by sex through gene expression, with 175 and 198 genes selected in each of the two subsamples. We performed the canonical correlation analysis (Harold, 1936) to identify and test the association between two selected gene lists for each trait. Over 90% of the variance in canonical variates for systolic BP can be explained by the top 13 canonical correlations. Similarly, over 90% of the variance in canonical variates for HDL-C can be captured by the top 12 canonical correlations. In conclusion, the selected genes in the two subsamples largely captured similar biological information, likely at the pathway level, even though they did not exactly overlap.

To further gain insights into the mediating biological pathways, we performed pathway enrichment analysis of the selected mediating genes in all subsamples for systolic BP and HDL-C. Four and one nominally significant pathways were identified for systolic BP and

HDL-C, respectively (See Web Appendix D). For instance, proteoglycans have an effect on the development and signalling of extracellular matrix, which plays a crucial role in regulating vascular function and blood pressure (Mouw et al., 2014; Wight, 2018). The plasma levels of proteoglycans and inflammatory proteins have recently been shown to be potential biomarkers for pulmonary hypertension (Arvidsson et al., 2021). See Web Appendix D for discussion on the other identified pathways.

Lastly, the computation time for CF-OLS to construct confidence intervals was substantially reduced compared with B-Mixed; indeed, the CF-OLS method can be 400 times faster with the same computational resource. Specifically, it took about 6.45 hours to finish the analysis for systolic BP with CF-OLS using a single core, while it took around 135.54 hours with nonparametric bootstrap using 25 cores in parallel. For the HDL-C outcome analysis, it took about 5.19 hours for the CF-OLS method using a single core, and around 232.46 hours were required for the B-Mixed method using 10 cores in parallel.

5 Discussion

We have proposed a novel two-stage interval estimation procedure for R^2_{Med}, based on cross-fitting and sample-splitting, to estimate the total mediation effect for high-dimensional mediators. Unlike the estimation method using nonparametric bootstrap in a mixed model framework, our proposed method relies on the asymptotic distribution of $\hat{R}^2_{Med}$ to construct confidence intervals. After splitting the data into two subsamples, we estimated R^2_{Med} using OLS regressions and conducted the inference based on the asymptotic standard error. We excluded the non-mediators by iSIS-MCP in two subsamples independently and fitted OLS regression in the other subsample. In addition, the point estimation improved over the original point estimation method by Yang et al. (2021) in terms of the MSE because the new method used the full data for variable selection and estimation, as shown in our extensive simulation studies in Table 1. The CF-OLS method had comparable coverage probability and variable selection accuracy across various scenarios with the B-Mixed method. Meanwhile, significantly reducing computational time facilitates the exploration of different settings in the variable selection procedure. If instead we used iSIS-Lasso for mediator selection, the coverage probability was reasonable, but the false positive rate in some specific scenarios increased due to failure in excluding the non-mediators $\boldsymbol{M}_{\mathcal{I}_2}$. For both iSIS-MCP and iSIS-Lasso, increasing the maximum number of iterations of iSIS made little difference while sacrificing computational efficiency.

In the FHS data analysis, treating systolic BP and HDL-C as outcomes, we applied the CF-OLS and B-Mixed methods to examine the mediating role of gene expression between exposure and phenotype. As in previously established findings (Yang et al., 2021), a large amount of systolic BP variation can be explained by age through gene expression. In addition, we discovered that the effect of sex on HDL-C was mediated by gene expression. Similar conclusions can be drawn after comparing the R^2_{Med} and its confidence intervals from the two methods, which corroborates the validity of the CP-OLS method. More importantly and as expected, the CF-OLS method is very computationally efficient, because CF-OLS only performs the iSIS variable selection procedure twice to construct confidence intervals instead of 500 times in the resampling-based B-Mixed method. To compute the confidence interval for systolic BP in the FHS dataset, the B-Mixed method needed around 135.5 hours even with multi-core parallel computing, while the CF-OLS method could achieve it efficiently in about 6.5 hours using a single core. This advantage makes OF-OLS more practical to estimate the total mediation effect with confidence intervals under the high-dimensional setting with a relatively massive dataset.

A critical research area in public health is to study how an exposure influences phenotypic variation. It has been well established that exposures, including environmental (Bind et al., 2014; Timms et al., 2016), socioeconomic (Cerutti et al., 2021), and behavioral factors (Zong et al., 2019; Hardy and Tollefsbol, 2011; Tiffon, 2018; Maas et al., 2020), are associated with changes at the molecular level (Bind et al., 2014; Timms et al., 2016; Maas et al., 2020; Huang et al., 2018; Tobi et al., 2018). Mediation analysis provides a useful tool to decompose the relationship between an exposure and an outcome into direct and mediation (indirect) effects. Over the past three decades, mediation analyses extensively studied settings where a single or a few mediators are present (Zeng et al., 2021). These methods are not in general applicable to high-dimensional molecular mediators. Here we have focused on an important but less explored quantity, the total mediation effect that captures the variance of the outcome explained by an exposure through high-dimensional mediators. Accurate estimation of the total mediation effect allows us to better understand the mediating roles of genomic factors in various ways, including exploring the impact of a certain molecular phenotype in the exposure-outcome pathway, identifying the relevant tissues or cell types, and improving the understanding of the time-varying mediating role of a molecular phenotype. In addition to deepening our understanding of the biological mechanism at the molecular level, estimating the total mediation effect has the potential to guide outcome prediction and intervention. For example, incorporating mediators has been shown to benefit the prediction of survival outcomes (Zhou et al., 2022). Also, Tingley et al. (2014) suggests that refining interventions targeting the mechanism that explains a large proportion of the intervention effect on the outcome may be more desirable than the ones that do not.

The proposed method is available in the updated `RsqMed` package on R/`CRAN`, which includes the new CF-OLS method. Lastly, we have focused on continuous outcomes and will extend our proposed approach to accommodate time-to-event or binary outcomes in the future.

Acknowledgements

This research was supported by National Institutes of Health (NIH) grant R01HL116720 (P.W.). T.Y. was partially supported by Children's Cancer Research fund and a Scholar Award from the Saint Baldrick's Foundation. The FHS is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (contract numbers N01-HC-25195, HHSN268201500001I, and 75N92019D00031). This manuscript was not prepared in collaboration with investigators in the FHS and does not necessarily reflect the opinions or views of the FHS, Boston University, or the NHLBI. The data set used for the analyses described in this manuscript was obtained from dbGaP at

https://www.ncbi.nlm.nih.gov/gap/ through accession number phs000007.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer.

Arvidsson, M., Ahmed, A., Bouzina, H., and Rådegran, G. (2021). Plasma proteoglycan prolargin in diagnosis and differentiation of pulmonary arterial hypertension. *ESC heart failure*, 8(2):1230–1243.

Bind, M.-A., Lepeule, J., Zanobetti, A., Gasparrini, A., Baccarelli, A. A., Coull, B. A., Tarantini, L., Vokonas, P. S., Koutrakis, P., and Schwartz, J. (2014). Air pollution and gene-specific methylation in the normative aging study: association, effect modification, and mediation analysis. *Epigenetics*, 9(3):448–458.

Castelli, W. (1988). Cholesterol and lipids in the risk of coronary artery disease–the framingham heart study. *The Canadian journal of cardiology*, 4:5A–10A.

Cerutti, J., Lussier, A. A., Zhu, Y., Liu, J., and Dunn, E. C. (2021). Associations between indicators of socioeconomic position and dna methylation: a scoping review. *Clinical Epigenetics*, 13(1):1–20.

Dai, J. Y., Stanford, J. L., and LeBlanc, M. (2022). A multiple-testing procedure for high-dimensional mediation hypotheses. *Journal of the American Statistical Association*, 117(537):198–213.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65.

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(5):849–911.

Gao, Y., Yang, H., Fang, R., Zhang, Y., Goode, E. L., and Cui, Y. (2019). Testing mediation effects in high-dimensional epigenetic studies. *Frontiers in Genetics*, 10:1195.

Hardy, T. M. and Tollefsbol, T. O. (2011). Epigenetic diet: impact on the epigenome and cancer. *Epigenomics*, 3(4):503–518.

Harold, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321.

Huang, J. V., Cardenas, A., Colicino, E., Schooling, C. M., Rifas-Shiman, S. L., Agha, G., Zheng, Y., Hou, L., Just, A. C., Litonjua, A. A., et al. (2018). Dna methylation in blood as a mediator of the association of mid-childhood body mass index with cardio-metabolic risk score in early adolescence. *Epigenetics*, 13(10-11):1072–1087.

Huang, Y.-T. and Pan, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, 72(2):402–413.

Joehanes, R., Johnson, A. D., Barb, J. J., Raghavachari, N., Liu, P., Woodhouse, K. A., O'Donnell, C. J., Munson, P. J., and Levy, D. (2012). Gene expression analysis of whole blood, peripheral blood mononuclear cells, and lymphoblastoid cell lines from the framingham heart study. *Physiological genomics*, 44(1):59–75.

Jousilahti, P., Vartiainen, E., Tuomilehto, J., and Puska, P. (1999). Sex, age, cardiovascular risk factors, and coronary heart disease: a prospective follow-up study of 14 786 middle-aged men and women in finland. *Circulation*, 99(9):1165–1172.

Kearney, P. M., Whelton, M., Reynolds, K., Muntner, P., Whelton, P. K., and He, J. (2005). Global burden of hypertension: analysis of worldwide data. *The lancet*, 365(9455):217–223.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., and Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of general psychiatry*, 59(10):877–883.

Lawlor, D. A., Ebrahim, S., and Smith, G. D. (2001). Sex matters: secular and geographical trends in sex differences in coronary heart disease mortality. *Bmj*, 323(7312):541–545.

Maas, S. C., Mens, M. M., Kühnel, B., van Meurs, J. B., Uitterlinden, A. G., Peters, A., Prokisch, H., Herder, C., Grallert, H., Kunze, S., et al. (2020). Smoking-related changes in dna methylation and gene expression are associated with cardio-metabolic traits. *Clinical epigenetics*, 12(1):1–16.

Martinez, J. G., Carroll, R. J., Muller, S., Sampson, J. N., and Chatterjee, N. (2010). A note on the effect on power of score tests via dimension reduction by penalized regression under the null. *The International Journal of Biostatistics*, 6(1):1–14.

Mills, K. T., Bundy, J. D., Kelly, T. N., Reed, J. E., Kearney, P. M., Reynolds, K., Chen, J., and He, J. (2016). Global disparities of hypertension prevalence and control: a systematic analysis of population-based studies from 90 countries. *Circulation*, 134(6):441–450.

Mouw, J. K., Ou, G., and Weaver, V. M. (2014). Extracellular matrix assembly: a multiscale deconstruction. *Nature reviews Molecular cell biology*, 15(12):771–785.

Roth, G. A., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., et al. (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1736–1788.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Song, Y., Zhou, X., Zhang, M., Zhao, W., Liu, Y., Kardia, S. L., Roux, A. V. D., Needham, B. L., Smith, J. A., and Mukherjee, B. (2020). Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics*, 76(3):700–710.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tiffon, C. (2018). The impact of nutrition and environmental epigenetics on human health and disease. *International journal of molecular sciences*, 19(11):3425.

Timms, J. A., Relton, C. L., Rankin, J., Strathdee, G., and McKay, J. A. (2016). Dna methylation as a potential mediator of environmental risks in the development of childhood acute lymphoblastic leukemia. *Epigenomics*, 8(4):519–536.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59:1–38.

Tobi, E. W., Slieker, R. C., Luijk, R., Dekkers, K. F., Stein, A. D., Xu, K. M., based Integrative Omics Studies Consortium, B., Slagboom, P. E., van Zwet, E. W., Lumey, L., et al. (2018). Dna methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Science advances*, 4(1):eaao4364.

Tobin, M. D., Sheehan, N. A., Scurrah, K. J., and Burton, P. R. (2005). Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in medicine*, 24(19):2911–2935.

Weidner, G., Connor, S. L., Chesney, M. A., Burns, J. W., Connor, W. E., Matarazzo, J. D., and Mendell, N. R. (1991). Sex differences in high density lipoprotein cholesterol among low-level alcohol consumers. *Circulation*, 83(1):176–180.

Wight, T. N. (2018). A role for proteoglycans in vascular disease. *Matrix Biology*, 71:396–420.

Wilson, P. W., Savage, D. D., Castelli, W. P., Garrison, R. J., Donahue, R. P., and Feinleib, M. (1983). Hdl-cholesterol in a sample of black adults: the framingham minority study. *Metabolism*, 32(4):328–332.

Yang, T., Niu, J., Chen, H., and Wei, P. (2021). Estimation of total mediation effect for high-dimensional omics mediators. *BMC bioinformatics*, 22(1):1–17.

Zeng, P., Shao, Z., and Zhou, X. (2021). Statistical methods for mediation analysis in the era of high-throughput genomics: current successes and future challenges. *Computational and structural biotechnology journal*, 19:3209–3224.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhao, Y. and Luo, X. (2022). Pathway lasso: pathway estimation and selection with high-dimensional mediators. *Statistics and Its Interface*, 15(1):39–50.

Zhou, J., Jiang, X., Xia, H. A., Wei, P., and Hobbs, B. P. (2022). Predicting outcomes of phase iii oncology trials with bayesian mediation modeling of tumor response. *Statistics in Medicine*, 41(4):751–768.

Zong, D., Liu, X., Li, J., Ouyang, R., and Chen, P. (2019). The role of cigarette smoke-induced epigenetic alterations in inflammation. *Epigenetics & Chromatin*, 12(1):1–25.