CoxMKF: A Knockoff Filter for High-Dimensional Mediation Analysis with a Survival Outcome in Epigenetic Studies

Peixin Tian[1]*, Minhao Yao[1], Tao Huang[2,3], Zhonghua Liu[1*]

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong SAR, China.

[2] Department of Epidemiology and Biostatistics, School of Public Health, Peking University, China.

[3] Key Laboratory of Molecular Cardiovascular Sciences (Peking University), Ministry of Education, Beijing, China.

June 29, 2022

Abstract

Motivation: It is of scientific interest to identify DNA methylation CpG sites that might mediate the effect of an environmental exposure on a survival outcome in high-dimensional mediation analysis. However, there is a lack of powerful statistical methods that can provide a guarantee of false discovery rate (FDR) control in finite-sample settings.

Results: In this article, we propose a novel method called CoxMKF, which applies aggregation of multiple knockoffs to a Cox proportional hazards model for a survival outcome with high-dimensional mediators. The proposed CoxMKF can achieve FDR control even in finite-sample settings, which is particularly advantageous when the sample size is not large. Moreover, our proposed CoxMKF can overcome the randomness of the unstable model-X knockoffs. Our simulation results show that CoxMKF controls FDR well in finite samples. We further apply CoxMKF to a lung cancer data set from The Cancer Genome Atlas (TCGA) project with 754 subjects and 365 306 DNA methylation CpG sites, and identify four DNA methylation CpG sites that might mediate the effect of smoking on the overall

[1]Contributed equally.

survival among lung cancer patients.

Availability: The R package **CoxMKF** is publicly available at https://github.com/MinhaoYaooo/CoxMKF.

Contact: zhhliu@hku.hk

1 Introduction

Cigarette smoking is a well-known risk factor for inducing lung cancer (Walser et al., 2008). However, it's unclear how epigenetic changes (e.g., DNA methylation) might mediate the effect of smoking on lung cancer survival. DNA methylation is the covalent bonding of a methyl group (CH_3) onto the C5 position of the cytosine, which can regulate gene expressions by engaging the relevant genes or by blocking transcription factors binding to DNA (Moore et al., 2013). Since DNA methylation is a reversible process (Wu and Zhang, 2014), it's thus of scientific interest to investigate the mediator role of DNA methylation CpG sites in the causal pathway from smoking to lung cancer survival for targeted intervention.

However, with high-dimensional DNA methylation data, it is challenging to select causal DNA methylation CpG sites with a guarantee of finite-sample false discovery rate (FDR) control. Recently, Luo et al. (2020) extended the Cox proportional hazards model (Cox, 1972) to high-dimensional mediators (e.g., CpG sites) settings by applying the minimax concave penalty (MCP) (Zhang, 2010) and Sobel test (Sobel, 1982) to select DNA methylation CpG sites as mediators. Furthermore, Zhang et al. (2021) introduced a de-biased Lasso estimator to handle ultra-dimensional mediators with a survival outcome. However, both methods cannot guarantee FDR control in finite-sample settings.

To address this challenge, we propose a novel mediator selection method named CoxMKF to identify DNA methylation CpG sites in the causal pathway from an environmental exposure to a survival outcome by incorporating aggregation of multiple knockoffs (AKO) with a guarantee of finite-sample FDR control (Nguyen et al., 2020). The key idea of the knockoffs inference is to generate fake copies of the original variables (conditionally on the original variables) and then perform variable selection (Candes et al., 2018). The model-X knockoff filter proposed by Candes et al. (2018) generates knockoffs by mimicking the original design matrix which are

flagged as false positives in a regression model. The finite-sample FDR control is achieved by using a data-dependent threshold. However, the model-X knockoff filter is unstable and might be difficult to reproduce research findings.

Our proposed CoxMKF method brings the idea of statistical aggregation from Meinshausen et al. (2009) to conduct variable selection via AKO (Nguyen et al., 2020) to improve stability. Specifically, our method has three steps. First, we take the p-values from the linear regression model for the exposure-mediator relationship, and then carry out the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) to select top d significant DNA methylation CpG sites. Second, we generate model-X knockoffs according to the correlation structure of the d DNA methylation CpG sites. We then apply the Cox proportional hazards model (Cox, 1972) to the d DNA methylation CpG sites and their knockoff variables to obtain the regression coefficients. Third, we repeat the aforementioned steps for B times and propose two new feature statistics named Cox Coefficient Difference (CCD) and Multiple Cox Statistic (MCS) to determine a data-dependent threshold by applying statistical quantile aggregation with a pre-specified FDR level for variable selection. Our simulations indicate that the proposed CoxMKF may effectively control FDR with better stability in finite-sample settings. We further apply CoxMKF to a lung cancer data set from The Cancer Genome Atlas (TCGA) project and finally identify four DNA methylation CpG sites that might mediate the effect of smoking on lung cancer survival.

The rest of this article is organized as follows. In Section 2, we describe the regression models and introduce our CoxMKF procedure for high-dimensional mediation analysis with a survival outcome. In Section 3, we evaluate the performance of CoxMKF across a wide range of simulation settings. In Section 4, we apply CoxMKF to the lung cancer data from the TCGA project. Section 5 provides some concluding remarks.

2 Methods

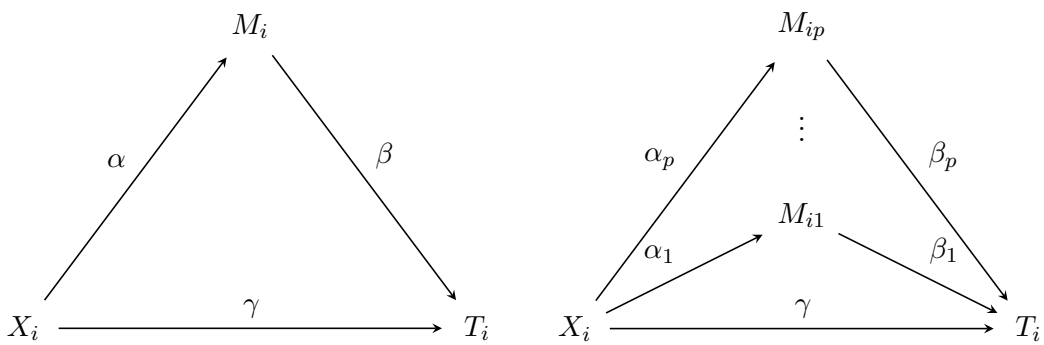
Let D_i be the duration of an event (e.g., lifespan) of individual i, and C_i be the corresponding censoring time. The observed survival outcome is $T_i = \min\{C_i, D_i\}$, and the censoring indicator is $\delta_i = I(D_i \leq C_i)$. We evaluate the causal mechanism from an exposure X_i to a survival outcome T_i through a p-dimensional mediator vector $\boldsymbol{M}_i = (M_{i1}, \cdots, M_{ip})^\top$ (see Figure **??**).

If $p = 1$, this problem reduces to a single mediator setting. In the multiple mediators setting ($p \geq 2$), we perform mediation analysis using the following two regression models:

$$M_{ik} = c_k + \alpha_k X_i + \boldsymbol{\vartheta}^\top \boldsymbol{Z}_i + \varepsilon_{ik}, \quad k \in [p], \tag{1}$$

$$\lambda_i(t|X_i, \boldsymbol{M}_i) = \lambda_0(t) \exp\left\{\gamma X_i + \boldsymbol{\theta}^\top \boldsymbol{Z}_i + \boldsymbol{\beta}^\top \boldsymbol{M}_i\right\}, \tag{2}$$

where $[p]$ denotes the set of $\{1, 2, \cdots, p\}$, $\lambda_i(t|X_i, \boldsymbol{M}_i)$ is the Cox proportional hazards model, $\lambda_0(t)$ is an unspecified baseline hazard function, $\boldsymbol{Z}_i = (Z_{i1}, \cdots, Z_{iq})$ is a q-dimensional vector of the baseline covariates (e.g., gender, age), and ε_{ik} is the noise term. In model (1), $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)^\top$ is a parameter vector relating the exposure to the mediators, and $\boldsymbol{c} = (c_1, \cdots, c_p)^\top$ is the intercept term. In model (2), γ is the direct effect of the exposure on the outcome, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top$ represents the effect of the mediators on the survival outcome adjusting for the effect of the exposure. $\boldsymbol{\vartheta}$ and $\boldsymbol{\theta}$ are the regression coefficients of the covariates on the mediators and on the survival outcome, respectively.



(a) Single mediator

(b) Multiple mediators

Figure 1: Causal diagram. The single mediator(left) and multiple mediators(right) related with exposure variable X_i and outcome T_i.

We use the potential outcomes framework (Rubin, 1974; Splawa-Neyman et al., 1990) to decompose the effect of the exposure on the outcome. For the Cox proportional hazards model, we can decompose the effect using the difference in counterfactual log hazards (VanderWeele, 2011; Huang and Yang, 2017). Specifically, let $T(x, \boldsymbol{m})$ denotes the survival time when the exposure is set to x and the vector of mediators is set to $\boldsymbol{m}$, and $\boldsymbol{M}(x) = (M_1(x), \cdots, M_p(x))^\top$ denotes the vector of mediators when the exposure is set to x. Let $\boldsymbol{Z}$ denotes the baseline

covariates, and x^* denotes the reference level of the exposure. The total effect (TE) of the exposure on the outcome is defined as $\log\lambda(T(x,\boldsymbol{M}(x));t|\boldsymbol{Z}) - \log\lambda(T(x^*,\boldsymbol{M}(x^*));t|\boldsymbol{Z})$; the natural direct effect (NDE) is defined as $\log\lambda(T(x,\boldsymbol{M}(x^*));t|\boldsymbol{Z}) - \log\lambda(T(x^*,\boldsymbol{M}(x^*));t|\boldsymbol{Z})$; and the natural indirect effect (NIE) is defined as $\log\lambda(T(x,\boldsymbol{M}(x));t|\boldsymbol{Z}) - \log\lambda(T(x,\boldsymbol{M}(x^*));t|\boldsymbol{Z})$. The following equation summarizes the relationship among these effects:

$$\begin{aligned}
&\underbrace{\log\lambda(T(x,\boldsymbol{M}(x));t|\boldsymbol{Z}) - \log\lambda(T(x^*,\boldsymbol{M}(x^*));t|\boldsymbol{Z})}_{\text{TE}} \\
=&\underbrace{\log\lambda(T(x,\boldsymbol{M}(x^*));t|\boldsymbol{Z}) - \log\lambda(T(x^*,\boldsymbol{M}(x^*));t|\boldsymbol{Z})}_{\text{NDE}} \\
+&\underbrace{\log\lambda(T(x,\boldsymbol{M}(x));t|\boldsymbol{Z}) - \log\lambda(T(x,\boldsymbol{M}(x^*));t|\boldsymbol{Z})}_{\text{NIE}}.
\end{aligned} \tag{3}$$

According to VanderWeele (2011) and Huang and Yang (2017), the following standard assumptions are required to ensure the identification of NDE and NIE:

(C1) $T(x,\boldsymbol{m}) \perp X|\boldsymbol{Z}$: no unmeasured confounding between the exposure and the outcome.

(C2) $T(x,\boldsymbol{m}) \perp M_k|X,\boldsymbol{Z}$ for $k \in [p]$: no unmeasured confounding between mediators and the outcome.

(C3) $X \perp M_k|\boldsymbol{Z}$ for $k \in [p]$: no unmeasured confounding between the exposure and mediators.

(C4) $T(x,\boldsymbol{m}) \perp M_k(x^*)|\boldsymbol{Z}$ for $k \in [p]$: no exposure-induced confounding between the mediators and the survival outcome.

If assumptions (C1) - (C4) hold, then the counterfactual log hazard can be approximated as:

$$\begin{aligned}
\log\lambda(T(x,\boldsymbol{M}(x^*));t|Z) \approx& \log\lambda_0(t) + \boldsymbol{\theta}^\top\boldsymbol{Z} + \frac{1}{2}\sigma^2_{W_\beta} \\
&+ \sum_{k=1}^{p}\beta_k(c_k + \boldsymbol{\vartheta}^\top\boldsymbol{Z}) + \gamma x + \sum_{k=1}^{p}\alpha_k\beta_k x^*,
\end{aligned} \tag{4}$$

where $\sigma^2_{W_\beta} = \sum_{k=1}^{p}\beta_k^2\sigma^2_{M_k}$ and $\sigma^2_{M_k}$ is the variance of the kth mediator (Huang and Yang, 2017; Luo et al., 2020). Therefore, both NDE and NIE can be estimated on the log hazard scale. In

our models (1) and (2), we further specify the expressions of NDE and NIE in equation (3):

$$\text{NDE} = \gamma(x - x^*); \quad \text{NIE}_k = \alpha_k \beta_k (x - x^*).$$

One main challenge here is that when the dimension p is much larger than the sample size n, the Cox proportional hazards model (2) cannot be fitted directly. To address this challenge, we propose a novel framework for variable selection of the mediators that can provide finite-sample FDR control, which contains three major steps (as shown in Figure 2).

First, we apply variable screening to the mediators using the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995). Specifically, we obtain the p-value in the $X_i \to M_{ik}$ relationship and select a subset of significant mediators of size d by controlling the FDR at level q_1.

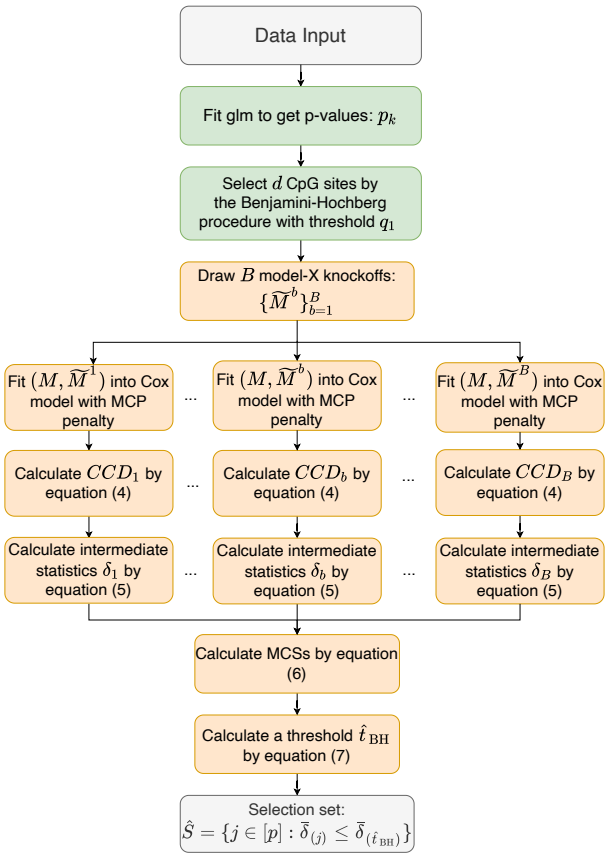


Figure 2: Flowchart of CoxMKF. First, we perform variable screening using the BH procedure in the $X_i \to M_{ik}$ relationship to obtain d mediators. Second, we sample B model-X knockoffs to fit in the Cox model. Third, we compute CCDs and MCSs to select the candidate mediators with a pre-specified FDR level.

Second, we introduce aggregation of multiple knockoffs (AKO) (Nguyen et al., 2020) to control the finite-sample FDR in the selection of mediators with non-zero NIEs on the survival outcome. Suppose that we have n individuals. We denote $\boldsymbol{M} = (\boldsymbol{M}_1, \cdots, \boldsymbol{M}_n)^\top$ as the $n \times d$ design matrix of mediators after variable screening, and $\boldsymbol{T} = (T_1, \cdots, T_n)^\top$ as the vector of the observed survival outcomes. We generate B model-X knockoffs $\{\widetilde{\boldsymbol{M}}^b\}_{b=1}^B$ using the R package 'knockoff' (Candes et al., 2018). Moreover, our procedure maintains two key properties of model-X knockoffs (also shown in Figure 3) (Candes et al., 2018; Nguyen et al., 2020):

(1) For any subset $S \subseteq [d]$ and $b \in [B]$, $(\boldsymbol{M}, \widetilde{\boldsymbol{M}}^b)_{\text{swap}(S)} \stackrel{d}{=} (\boldsymbol{M}, \widetilde{\boldsymbol{M}}^b)$, where $(\boldsymbol{M}, \widetilde{\boldsymbol{M}}^b)_{\text{swap}(S)}$ is obtained by swapping the kth columns of $\boldsymbol{M}$ and $\widetilde{\boldsymbol{M}}^b$ for each $k \in S$;

(2) $\widetilde{\boldsymbol{M}}^b \perp \boldsymbol{T} | \boldsymbol{M}$ for any $b \in [B]$.

It is worth noting that the second property is guaranteed if $\widetilde{\boldsymbol{M}}^b$ is constructed without looking at the response variable $\boldsymbol{T}$.

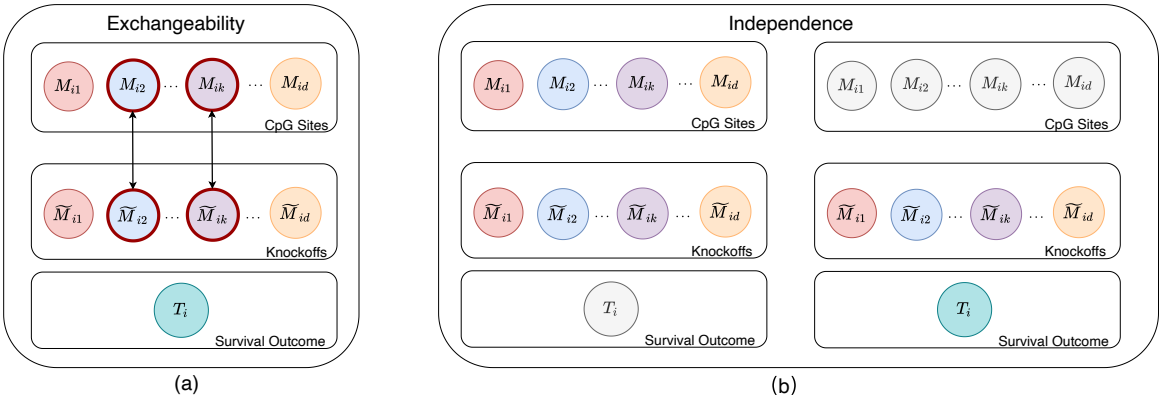


Figure 3: Illustration of knockoffs variables. (a) The knockoff copy keeps exchangeable property between $\boldsymbol{M}$ and $\widetilde{\boldsymbol{M}}$. As the left penal shows, the knockoff variable $\widetilde{M}_k$ can exchange the position with M_j. In this case, the model result will be unchanged if the positions of M_k and $\widetilde{M}_k$ are swapped. (b) The knockoff copy shows independence with the response T (uncolored) when conditional on the original features (colored), and shows dependence with T (colored) when not conditional on the original features (uncolored).

Third, we introduce a new aggregated feature statistic, called Multiple Cox Statistic (MCS) to select mediators with non-zero NIEs on the outcome. To calculate the MCS, we first obtain a B-by-d Cox Coefficient Difference (CCD) matrix $\boldsymbol{\tau} = (\tau_j^b)_{B \times d}$ whose entries are:

$$\tau_j^b = |\widehat{\beta}_j^b| - |\widetilde{\beta}_j^b|, \ b \in [B], \ j \in [d], \tag{5}$$

where $\widehat{\boldsymbol{\beta}}^b = (\widehat{\beta}_1^b, \cdots, \widehat{\beta}_d^b)^\top$ and $\widetilde{\boldsymbol{\beta}}^b = (\widetilde{\beta}_1^b, \cdots, \widetilde{\beta}_d^b)^\top$ are the coefficients obtained by fitting the Cox proportional hazards model with the MCP using both the mediators $\boldsymbol{M}$ and the bth knockoffs $\widetilde{\boldsymbol{M}}^b$. Based on the CCD matrix $\boldsymbol{\tau}$, we then calculate another B-by-d matrix of intermediate statistics $\boldsymbol{\Delta} = (\delta_j^b)_{B\times d}$ whose entries are:

$$\delta_j^b = \begin{cases} \frac{1+\#\{k\in[d]:\tau_k^b\leq-\tau_j^b\}}{d}, & \tau_j > 0; \\ 1, & \tau_j \leq 0. \end{cases} \tag{6}$$

Generally, a small δ_j^b value suggests a strong non-null signal for the jth mediator in the bth replication. Following Meinshausen et al. (2009), we aggregate the intermediate statistics on B rows to generate the MCS for each candidate mediator respectively:

$$\bar{\delta}_j = \min\left\{1, \frac{Q_\eta(\{\delta_j^b : b \in [B])}{\eta}\right\}, j \in [d], \tag{7}$$

where η is the pre-specified quantile point, and $Q_\eta(\cdot)$ denotes the η-quantile function. Specifically, a small value of $\bar{\delta}_j$ suggests that the jth mediator is considered to have a non-zero mediation effect over replications, and thus is more likely to be the true mediator. Following Nguyen et al. (2020). We then apply the BH procedure to determine a data-dependent threshold $\widehat{k}_{\text{BH}}$ to control the FDR:

$$\widehat{k}_{\text{BH}} = \max\left\{k \in [d] : \bar{\delta}_{(k)} \leq \frac{k}{q_2 d}\right\}, \tag{8}$$

where q_2 is a pre-specified target FDR level, and $\bar{\delta}_{(k)}$ is the kth smallest value of $\{\bar{\delta}_1, \cdots, \bar{\delta}_d\}$. Then we define the selection set of mediators with non-zero mediation effects as:

$$\hat{\mathcal{S}} = \{j \in [d] : \bar{\delta}_j \leq \bar{\delta}_{(\widehat{k}_{\text{BH}})}\}.$$

In this article, we evaluate the performance of CoxMKF by FDR and true positive proportion (TPP):

$$\text{FDR} = \mathbb{E}\Big[\frac{\#\{j \in \hat{S} \cap H_0\}}{\#\hat{S}}\Big], \quad \text{TPP} = \mathbb{E}\Big[\frac{\#\{j \in \hat{S} \cap H_1\}}{\#H_1}\Big],$$

where $H_0 = \{k : \text{NIE}_k = 0\}$ includes the candidate mediators with no signal, and $H_1 = \{k : \text{NIE}_k \neq 0\}$ includes the true mediators.

3 Simulation Studies

In this section, we conduct a wide range of simulation studies to evaluate the finite-sample performance of our proposed method. First, we generate the survival outcome $\boldsymbol{T} = (T_1, T_2, \cdots, T_n)^\top$ from the exponential model $\lambda(t_i|X_i, Z_{i1}, Z_{i2}, \boldsymbol{M}_i) = 0.5 \exp\left\{\gamma X_i + \boldsymbol{\theta}^\top \boldsymbol{Z}_i + \boldsymbol{\beta}^\top \boldsymbol{M}_i\right\}$, where the exposure X_i is generated from a Bernoulli distribution with success probability 0.6 as Ber(0.6), the binary covariate $\boldsymbol{Z}_1 = (Z_{1,1}, \cdots, Z_{n,1})^\top$ and the continuous covariate $\boldsymbol{Z}_2 = (Z_{1,2}, \cdots, Z_{n,2})^\top$ are generated from a Bernoulli distribution with success probability 0.3 as Ber(0.3) and a uniform distribution as $U(0, 1)$, respectively, and the coefficients of the covariates on the outcome are set to be $\theta_1 = 0.3$ and $\theta_2 = -0.2$. Then we generate the mediators $\boldsymbol{M} = (M_{ik})_{n\times p}$ according to model (1), where the coefficient c_k is generated from a uniform distribution as $U(0, 1)$, the noise ε_{ik} is generated from a standard normal distribution as $N(0, 1)$, and $\boldsymbol{\theta}_k$ is fixed to be $(0.3, 0.2)^\top$. As for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we fix the first ten elements of $\boldsymbol{\alpha}$ to be $(0.55, 0.45, -0.40, -0.45, 0.50, 0.60, -0.40, -0.46, -0.40, 0.50)^\top$, and the first twelve elements of $\boldsymbol{\beta}$ to be $(0.52, 0.45, 0.40, 0.40, -0.54, -0.60, -0.40, -0.50, 0, 0, 0.40, -0.80)^\top$. The rest elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are set to be 0. Thus, $(M_{i1}, M_{i2}, \cdots M_{i8})$ have non-zero causal mediation effects, and $H_1 = \{1, 2, \cdots, 8\}$ is the set including these eight true mediators. The censoring time is generated from an exponential distribution with parameter c_0. By denoting proper c_0, the average censoring rates are controlled at 20%, 40% and 60%, respectively. All the scenarios are based on 500 replications, where the sample sizes $n = (300, 500)$ and the numbers of candidate mediators $p = (500, 1\,000, 10\,000)$, respectively. We set $q_1 = 0.2$ in the screening step and the targeted FDR level $q_2 = 0.05$ in equation (8). In addition, based on the recommendation of Nguyen et al. (2020), we set the number of bootstrap replications B to 25, which could empirically offer stable performance.

Table 1: The FDR and TPP of mediator selection of CoxMKF and HIMAsurvival across 500 simulations with $q_1 = 0.2$, $q_2 = 0.05$, $\eta = 0.1$ and the number of AKO's bootstrap replications $B = 25$. The number of true mediators is 8. CR stands for the censoring rate.

CR	**n**	**p**	**FDR ($<$ 0.05)**		**TPP**	
			CoxMKF	**HIMAsurvival**	**CoxMKF**	**HIMAsurvival**
20%	300	500	0.030	0.013	0.829	0.899
		1000	0.032	0.016	0.767	0.893
		10000	0.032	0.095	0.552	0.849
	500	500	0.018	0.016	0.978	0.992
		1000	0.017	0.016	0.965	0.992
		10000	0.014	0.075	0.883	0.986
40%	300	500	0.031	0.011	0.824	0.870
		1000	0.030	0.015	0.758	0.855
		10000	0.030	0.098	0.540	0.822
	500	500	0.021	0.011	0.979	0.987
		1000	0.020	0.017	0.962	0.988
		10000	0.019	0.073	0.881	0.981
60%	300	500	0.038	0.014	0.784	0.766
		1000	0.033	0.012	0.729	0.726
		10000	0.027	0.086	0.522	0.735
	500	500	0.030	0.011	0.969	0.970
		1000	0.023	0.013	0.959	0.962
		10000	0.028	0.070	0.890	0.954

We compare CoxMKF to Luo et al. (2020)'s method (denoted as HIMAsurvival). We compute the FDR and TPP to assess the accuracy of variable selection, and the results are reported in Table 1. When the number of candidate mediators p is small, CoxMKF and HIMAsurvival can both control the FDR under the pre-specified level. However, the FDR of HIMAsurvival is higher than 0.05 when $p = 10\ 000$. In addition, the median of FDR values across 500 replications is higher than 0.05 when $n = 300$, $p = 10\ 000$ and the censoring rate is controlled at 20% or 40% (as shown in Figure 4). Overall, HIMAsurvival fails to control the finite-sample

FDR when p increases to 10 000. By comparison, CoxMKF can guarantee the FDR under the pre-specified level in all scenarios, even if the sample size is small and the number of candidate mediators is large (for example, $n = 300$ and $p = 10\ 000$).

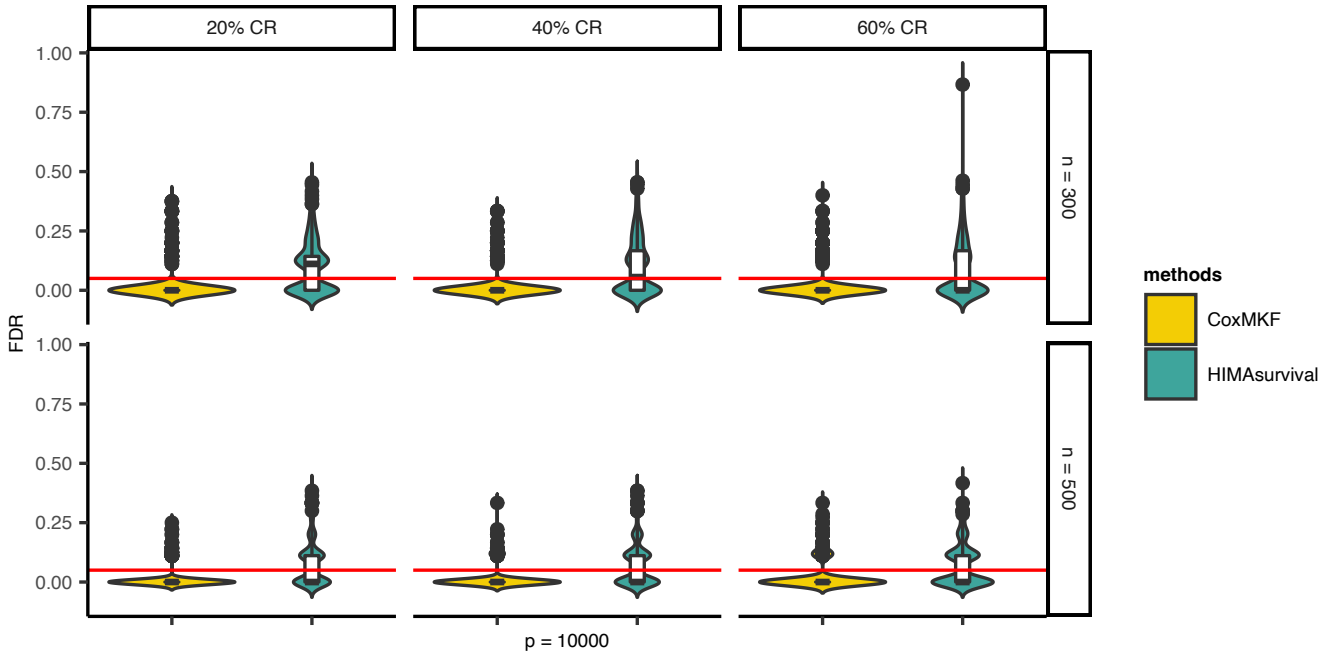


Figure 4: Violin plot and box plot of the FDR of CoxMKF and HIMAsurvival with $q_1 = 0.2$, $q_2 = 0.05$, $\eta = 0.1$, $p = 10\ 000$ and the number of AKO's bootstrap replications $B = 25$ across 500 replications in simulations. CR stands for the censoring rate, and the red line represents the FDR control level $q_2 = 0.05$. The number of true mediators is 8. We report the FDR when $p = 10\ 000$, and differentiate the methods by color, with yellow representing our proposed method CoxMKF and green representing the competing method HIMAsurvival. The width of the plot represents the value density. The greater the width of the shape, the denser the distribution of points.

Due to the inflated FDR of HIMAsurvival, the TPP of HIMAsurvival is generally higher than CoxMKF. The standard errors of FDR and TPP calculated by Monte Carlo method can be found in Supplementary Table S1.

4 An Application to the Lung Cancer Data

Lung cancer is the most common cancer among men and the third most common cancer among women, and remains the leading cause of mortality (Sung et al., 2021). Among all the lung cancer diagnoses, approximately 85% of all diagnoses are non-small cell lung cancer (NSCLC) and 15% of them are small-cell lung cancer (Cao et al., 2018). There are many urgent challenges

for researchers, such as finding an effective therapy and diagnosing the disease at early stage, which can significantly improve the treatment success rate and survival rate (Moore et al., 2013). Tobacco smoking is an important risk factor for lung cancer, and smoking status might affect the tumor genomic landscape (Govindan et al., 2012). Additionally, tobacco smoking is also associated with DNA methylation (Breitling et al., 2011). Therefore, it is of great interest to find DNA methylation CpG sites that mediate the effect of tobacco smoking on the lung cancer survival.

We apply our proposed CoxMKF to the lung cancer data from the TCGA project including lung squamous cell carcinoma and lung adenocarcinoma (https://xenabrowser.net/datapages/) to identify potential DNA methylation CpG sites between smoking and survival of lung cancer patients. In total, there are 754 patients aged 33-90 years with 365 306 DNA methylation CpG cites after excluding the patients with no survival time, smoking status and other covariates. The patients had DNA methylation profiles measured by the Illumina Infinium HumanMethylation 450 platform, and the DNA methylation values were recorded via BeadStudio software.

We are interested in detecting the DNA methylation CpG sites between smoking status (smoker/non-smoker) and lung cancer patients' survival. The survival of lung patients is encoded as the number of days from the initial diagnosis to the death or the censoring time. The median survival time is 1 632 days, and we observe 305 deaths over the follow-up period with a censoring rate of 60%. We also include the following covariates: sex, age, cancer stage and radiotherapy indicator.

For CoxMKF, we set $q_1 = 5 \times 10^{-4}$ in the screening step, $q_2 = 0.1$ as the targeted FDR level, $\eta = 0.05$ as the quantile point and the number of AKO's bootstrap replications $B = 25$. CoxMKF identifies four DNA methylation CpG sites (cg21926276, cg24129177, cg24200525 and cg07690349) in total, and the results of the effect estimates and the corresponding p-values are reported in Table 2. Among the three DNA methylation CpG sites, cg21926276 and cg24129177 have positive NIEs, through which smoking increases the possibility of mortality. The first methylation biomarker, cg21926276, is located in gene H19 on chromosome 11, and the estimated NIE is 0.149. Cg24129177 is located on chromosome 12, and the estimated NIE is 0.069. Previous studies also support that these two CpG cites play important roles in tumor or

lung cancer patients' survival, which validates our findings. For example, the gene H19 where cg21926276 is located in is found to be related to the growth of lung cancer, and the methylation level of gene H19 is negatively correlated with tobacco smoking (Matouk et al., 2015; Bouwland-Both et al., 2015). Although cg24129177 is not located in any gene, Hair (2014) has found that it is associated with body mass index (BMI), which has been confirmed to be related to the survival of lung cancer (Shepshelovich et al., 2019). Besides, those CpG sites with negative estimated effects may be potential DNA methylation CpG sites to extend the survival time of lung cancer patients. Take cg07690349 as an example, the DNA methylation CpG sites is located in gene MUC5B on chromosome 11. Wakata et al. (2015) indicated that MUC5B is treated as a favourable prognostic biomarker for lung cancer patients with NSCLC carrying EGFR mutations. Furthermore, the combination of MUC5B and TTF-1 expression plays an important role in identifying adenocarcinomas from squamous cell carcinomas with prognostic significance in lung adenocarcinoma patients (Nagashio et al., 2015). Additionally, cg24200525 is located in gene SBF1 on chromosome 12 relating with SERCA whose expression is relevant with the the susceptibility and progression of lung cancer (Li et al., 2014).

For fair comparison, we also apply HIMAsurvival to this data. HIMAsurvival identifies five DNA methylation CpG sites: cg02178957, cg21926276, cg26387335, cg24200525 and cg07690349. Cg21926276, cg24200525, and cg07690349 are the common DNA methylation CpG sites selected by both CoxMKF and HIMAsurvival. Among these five DNA methylation CpG sites, the last four CpG sites were also identified in Luo et al. (2020). The result is consistent with our simulation studies that HIMAsurvival might have higher (inflated) power to detect DNA methylation CpG sites but at the cost of inflated FDR in finite-sample settings.

As for the computational efficiency, the average running time of CoxMKF on this data set is 27.77 minutes with a server of Intel Xeon Silver 4116 CPU and 64 GB RAM memory, which is close to that of HIMAsurvival (26.63 minutes). Actually, the most time-consuming step is the screening process (25.89 minutes), which is a common step shared by both CoxMKF and HIMAsurvival. Notably, the multiple knockoffs generation and inference can be performed in parallel to induce AKO, which is no more costly than the single knockoff filter. Therefore, CoxMKF remains computationally efficient even when dealing with high-dimensional data.

In summary, we apply CoxMKF to the lung cancer data from the TCGA project and discover

four DNA methylation CpG sites from the smoking status to the survival of lung cancer patients, of which most DNA methylation CpG sites are supported by previous studies. These four DNA methylation CpG sites might be helpful for researchers to design interventions to improve the treatment for lung cancer patients with different smoking status.

Table 2: The identified DNA methylation CpG sites and estimated effects using CoxMKF with pre-specified $q_1 = 5 \times 10^{-4}$, $q_2 = 0.1$, $\eta = 0.05$ and $B = 25$.

CpGs	Gene Name	CHR	$\hat{\alpha}$	P_α	$\hat{\beta}$	P_β
cg21926276	H19	11	-0.058	5.32×10^{-8}	-2.554	9.61×10^{-5}
cg24129177	-	12	-0.056	1.70×10^{-7}	-1.225	2.43×10^{-2}
cg24200525	SBF1	12	-0.024	6.91×10^{-8}	5.136	1.52×10^{-4}
cg07690349	MUC5B	11	-0.075	7.17×10^{-8}	1.789	5.02×10^{-5}

Note: CHR stands for the chromosome.

5 Discussion

In this paper, we propose a novel method CoxMKF for high-dimensional mediation analysis with a survival outcome. To achieve finite-sample FDR control, we incorporate a powerful procedure, AKO for variable selection with a survival outcome. Simulation studies show that the proposed CoxMKF can achieve finite-sample FDR control compared to the competing method, HIMAsurvival. We further identify four DNA methylation CpG sites in the pathway from smoking to lung cancer patients' survival using the TCGA lung cancer data.

We highlight some advantages of CoxMKF here. First, the proposed CoxMKF applies AKO to conduct variable selection with finite-sample FDR control, which is more stable than the traditional model-X knockoffs method. Second, we propose two new feature statistics CCD and MSC for a survival outcome to extend the application of knockoff procedure to survival outcomes. Third, the proposed CoxMKF is computationally efficient even when dealing with high-dimensional data.

There are several interesting future research topics. First, we did not consider the correlation structure among multiple mediators, which is actually the common practice in this multiple-mediator field. But it will help increase power if the correlation structure is taken into account. Second, it is also interesting to explore other types of survival outcomes such as interval-censored

outcomes.

Acknowledgement

Dr. Zhonghua Liu's work is funded by Hong Kong Research Grants Council—Early Career Scheme (27307920) and the Seed Fund for Basic Research for New Staff (201902159001) of University of Hong Kong.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bouwland-Both, M. I., van Mil, N. H., Tolhoek, C. P., Stolk, L., Eilers, P. H., Verbiest, M. M., Heijmans, B. T., Uitterlinden, A. G., Hofman, A., van Ijzendoorn, M. H., et al. (2015). Prenatal parental tobacco smoking, gene specific DNA methylation, and newborns size: the Generation R study. *Clinical Epigenetics*, 7(1):1–11.

Breitling, L. P., Yang, R., Korn, B., Burwinkel, B., and Brenner, H. (2011). Tobacco-smoking-related differential dna methylation: 27k discovery and replication. *The American Journal of Human Genetics*, 88(4):450–457.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Cao, J., Yuan, P., Wang, Y., Xu, J., Yuan, X., Wang, Z., Lv, W., and Hu, J. (2018). Survival rates after lobectomy, segmentectomy, and wedge resection for non-small cell lung cancer. *The Annals of Thoracic Surgery*, 105(5):1483–1491.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, N. D., Kanchi, K. L., Maher, C. A., Fulton, R., Fulton, L., Wallis, J., et al. (2012). Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*, 150(6):1121–1134.

Hair, B. Y. (2014). *Body mass index, breast tissue, and the epigenome.* PhD thesis, The University of North Carolina at Chapel Hill.

Huang, Y.-T. and Yang, H.-I. (2017). Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology (Cambridge, Mass.)*, 28(3):370.

Li, W., Ouyang, Z., Zhang, Q., Wang, L., Shen, Y., Wu, X., Gu, Y., Shu, Y., Yu, B., Sun, Y., et al. (2014). SBF-1 exerts strong anticervical cancer effect through inducing endoplasmic reticulum stress-associated cell death via targeting sarco/endoplasmic reticulum Ca2+-ATPase 2. *Cell Death & Disease*, 5(12):e1581–e1581.

Luo, C., Fa, B., Yan, Y., Wang, Y., Zhou, Y., Zhang, Y., and Yu, Z. (2020). High-dimensional mediation analysis in survival models. *PLOS Computational Biology*, 16(4):1–15.

Matouk, I. J., Halle, D., Gilon, M., and Hochberg, A. (2015). The non-coding RNAs of the H19-IGF2 imprinted loci: a focus on biological roles and therapeutic potential in Lung Cancer. *Journal of Translational Medicine*, 13(1):1–12.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Moore, L. D., Le, T., and Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38.

Nagashio, R., Ueda, J., Ryuge, S., Nakashima, H., Jiang, S.-X., Kobayashi, M., Yanagita, K., Katono, K., Satoh, Y., Masuda, N., et al. (2015). Diagnostic and prognostic significances of MUC5B and TTF-1 expressions in resected non-small cell lung cancer. *Scientific Reports*, 5(1):1–7.

Nguyen, T.-B., Chevalier, J.-A., Thirion, B., and Arlot, S. (2020). Aggregation of multiple knockoffs. In *International Conference on Machine Learning*, pages 7283–7293. PMLR.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Shepshelovich, D., Xu, W., Lu, L., Fares, A., Yang, P., Christiani, D., Zhang, J., Shiraishi, K., Ryan, B. M., Chen, C., et al. (2019). Body mass index (BMI), BMI change, and overall survival in patients with SCLC and NSCLC: a pooled analysis of the international lung cancer consortium. *Journal of Thoracic Oncology*, 14(9):1594–1607.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13:290–312.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.

VanderWeele, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology (Cambridge, Mass.)*, 22(4):582—585.

Wakata, K., Tsuchiya, T., Tomoshige, K., Takagi, K., Yamasaki, N., Matsumoto, K., Miyazaki, T., Nanashima, A., Whitsett, J. A., Maeda, Y., et al. (2015). A favourable prognostic marker for EGFR mutant non-small cell lung cancer: immunohistochemical analysis of MUC5B. *BMJ Open*, 5(7):e008366.

Walser, T., Cui, X., Yanagawa, J., Lee, J. M., Heinrich, E., Lee, G., Sharma, S., and Dubinett, S. M. (2008). Smoking and lung cancer: the role of inflammation. *Proceedings of the American Thoracic Society*, 5(8):811–815.

Wu, H. and Zhang, Y. (2014). Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell*, 156(1-2):45–68.

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

Zhang, H., Zheng, Y., Hou, L., Zheng, C., and Liu, L. (2021). Mediation analysis for survival data with high-dimensional mediators. *Bioinformatics*.