Group lasso based selection for high-dimensional mediation analysis

Allan Jérolon, Flora Alarcon, Florence Pittion, Magali Richard, Olivier François, Etienne E. Birmelé, Vittorio Perduca

Group lasso based selection for high-dimensional mediation analysis

Allan Jérolon[1], Flora Alarcon[1], Florence Pittion[2], Magali Richard[2], Olivier François[2], Etienne Birmelé[3†], Vittorio Perduca[1†]

[1]Université Paris Cité, CNRS, MAP5, F-75006 Paris, France.
[2]Univ. Grenoble Alpes, CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, 38000 Grenoble, France.
[3]Institut de Recherche Mathématique Avancée, UMR 7501 Université de Strasbourg et CNRS, 7 rue René-Descartes, 67000 Strasbourg, France.

Contributing authors: allan.jerolon@gmail.com; flora.alarcon@u-paris.fr; florence.pittion@univ-grenoble-alpes.fr; magali.richard@univ-grenoble-alpes.fr; olivier.francois@univ-grenoble-alpes.fr; etienne.birmele@unistra.fr; vittorio.perduca@u-paris.fr;
†These authors contributed equally to this work.

Abstract

Mediation analysis aims to identify and estimate the effect of an exposure on an outcome that is mediated through one or more intermediate variables. In the presence of multiple intermediate variables, two pertinent methodological questions arise: estimating mediated effects when mediators are correlated, and performing high-dimensional mediation analysis when the number of mediators exceeds the sample size. This paper presents a two-step procedure for high-dimensional mediation analysis. The first step selects a reduced number of candidate mediators using an ad-hoc lasso penalty. The second step applies a procedure we previously developed to estimate the mediated and direct effects, accounting for the correlation structure among the retained candidate mediators. We compare the performance of the proposed two-step procedure with state-of-the-art methods using simulated data. Additionally, we demonstrate its practical application by estimating the causal role of DNA methylation in the pathway between smoking and rheumatoid arthritis using real data.

Keywords: mediation analysis, high-dimensional statistics, group lasso, variable selection, methylation data.

1 Introduction

Mediation analyses methods are widely used in biomedical and social sciences to disentangle the causal effect of a treatment on an outcome through intermediate variables called mediators. Modern causal mediation analysis is based on counterfactual variables and aims at decomposing the total effect into a direct effect and the mediated effect(s) carried by the mediator(s) [1, 2].

In many practical problems, for instance in biomedical applications with intermediate variables of genomic nature, the number of potential mediators exceeds the sample size, leading to the high-dimensional mediation problem. Several methods have been proposed in recent years to address this challenging problem, for a review of the literature see [3, 4]. Existing methods can be can be broadly categorized into two main families based on their approach to dimensionality reduction.

Methods in the first family build uncorrelated linear combinations of potentials mediators, using PCA [5], sparse PCA [6] or PLS [7] approaches. In [8] a linear combinations of candidate mediators is chosen by maximising a criterion based on the joint likelihood of the treatment/mediator and mediator/outcome models. This approach is extended in [9] using a generalized version of population value decomposition (PVD). With any of these methods, the mediated effect carried by each linear combination can be evaluated, and the weights of the mediators within these linear combinations reveal their contribution to the mediated effects.

A second family of approaches, to which this paper belongs, involves screening the candidate mediators to select a subset and subsequently estimating their mediation effects. [10] proposes to explore the set of possible mediators by a coordinate descent updating at each step the status of a small number of potential mediators. [11] reduces the dimensionality by introducing a small set of latent variables governing both the potential mediators and the outcome. To introduce further approaches, let us assume linear (or logistic) regression models, and let $\boldsymbol{\alpha}$ be the vector of the coefficients of the exposure in the regression models of the candidate mediators given the exposure (one model per mediator), and $\boldsymbol{\beta}$ the vector of the coefficients of the candidate mediators in the model of the outcome given the mediators and the exposure. With these notations, a third way to select mediators is to suppose that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ follow Gaussian mixture models whose base distributions are centered and with either small or large variance. [12] proposes a Bayesian Sparse Linear Mixed Model for high-dimensional mediation analysis which is a one-step method. In contrast, the HDMAX2 method [13] makes no distributional assumption. For each mediator M_k, the HDMAX2 method tests the nullity of α_k and β_k, and the squared maximum of the two corresponding p-values is considered as a new p-value used as a selection criterion.

Other methods for the selection of mediators rely on penalized likelihood optimization with the selection method varying according to the considered model and penalization. After reducing the pool of mediators from a large number to a moderate number by employing the sure independence screening, [14], and its extension [15], conduct variable selection with the minimax concave penalty, or a de-biased lasso procedure respectively, and finally carry out joint significance testing for mediation effect. Interestingly, [16] considers a different definition of the mediated effect, called interventional indirect effect, that needs less stringent hypothesis on the joint law of

the mediators. The selection strategy relies on two penalized regression, for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively.

In this article, we propose a new two-step approach for the selection of candidate mediators and the estimation of individual indirect effects. The first filtering step reduces the number of candidate mediators by solving a penalized optimization problem with group lasso penalty that takes simultaneously the parameters of interest $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ into account. Moreover, the first step also allows to consider a predefined group structure among the possible mediators. Once the number of candidate mediators is lower than the sample size, the second step consists in running the algorithm developed in [17] to estimate and test the indirect effects of the retained mediators, together with the direct effect.

This article is organized as follows. Section 2 defines the problem of high-dimensional mediation analysis and introduces the notations and underlying hypotheses. Our algorithm is detailed in Section 3. The results of the comparisons with previously published methods on synthetic dataset are reported in Section 4. An illustration on a real dataset is shown in Section 5. Section 6 discusses our results.

2 A high-dimensional mediation analysis model

We consider a mediation model with P binary exposures (or treatments) $(T_1, \ldots, T_P)$, K candidate mediators $(M_1, \ldots, M_K)$ and an outcome Y. An example is shown in Figure 1. Let $(X_1, \ldots, X_L)$ be the vector of pretreatment confounders. If K is large, in particular larger than the sample size n, the problem of identifying and inferring the direct and indirect effects in the model is referred to as *high-dimensional mediation analysis*. In this high-dimensional setting, the aim of our algorithm is to identify which candidate mediators truly have a mediation effect and to estimate the corresponding direct and indirect effects. The inclusion of multiple treatments is furthermore designed to promote the selection of mediators that are common across different treatments. Both the candidate mediators and the outcome are assumed to be either Gaussian or binary, and are therefore modeled using either Gaussian or logistic regression models, respectively. We consider the following data structures:

- $n \times P$ matrix $\mathbf{T}$, where the entry t_{ip} is the i^{th} observation of T_p
- $n \times K$ matrix $\mathbf{M}$, where the entry m_{ik} is the i^{th} observation of M_k
- $n \times 1$ column vector $\mathbf{y}$, where the entry y_i is the i^{th} observation of Y
- $n \times L$ matrix $\mathbf{X}$, where the entry X_{il} is the i^{th} observation of the l^{th} pretreatment variable X_l.

Regression models for the candidate mediators M_k

If the k^{th} potential mediator is continuous, we assume the following Gaussian model:

$$M_k = \alpha_{0k} + \sum_{p=1}^{P} \alpha_{pk} T_p + \sum_{l=1}^{L} \xi_{lk} X_l + \epsilon_k \text{ with } \epsilon_k \sim \mathcal{N}(0, \sigma_k^2)$$

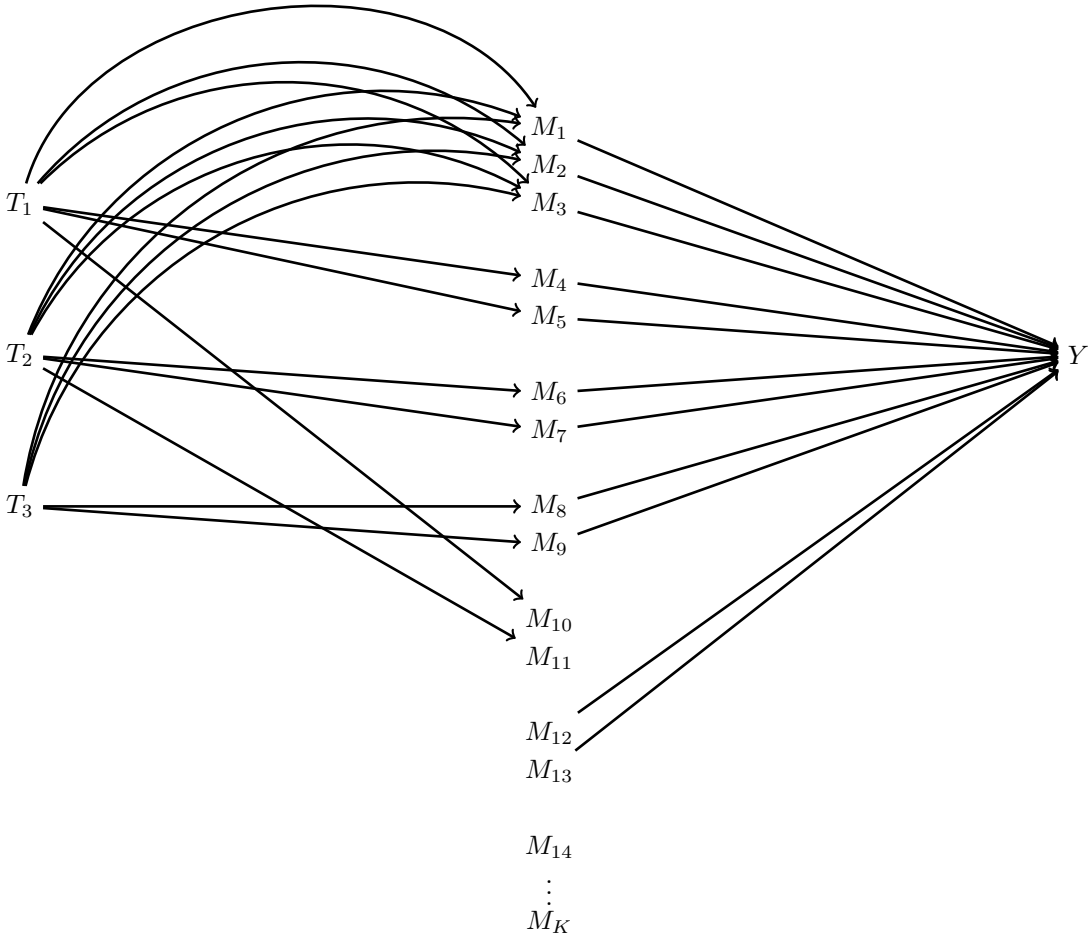

Fig. 1: Example of a high-dimensional mediation model with three treatments. Direct effects from (T_1, T_2, T_3) on the outcome are included in the model but omitted from the figure for readability. Candidate mediators M_1 to M_9 are true mediators, while M_{10} to M_K are not. Pretreatment confounders are not shown for clarity.

We denote by $\hat{m}_{ik}(\boldsymbol{\alpha}, \boldsymbol{\xi})$ the associated prediction for the i^{th} individual seen as a function of the model parameters: $\hat{m}_{ik}(\boldsymbol{\alpha}, \boldsymbol{\xi}) = \alpha_{0k} + \sum_{p=1}^{P} \alpha_{pk} t_{ip} + \sum_{l=1}^{L} \xi_{lk} x_{il}$. If the k^{th} potential mediator is binary, we assume the following logistic regression model:

$$\log\left(\frac{\mathbb{P}(M_k = 1)}{1 - \mathbb{P}(M_k = 1)}\right) = \alpha_{0k} + \sum_{p=1}^{P} \alpha_{pk} T_p + \sum_{l=1}^{L} \xi_{lk} X_l$$

We then denote $\hat{m}_{ik}(\boldsymbol{\alpha}, \boldsymbol{\xi})$ the associated prediction for $\mathbb{P}(m_{ik} = 1)$, that is $\hat{m}_{ik} = e^{\nu_{ik}}/(1 + e^{\nu_{ik}})$ with $\nu_{ik}(\boldsymbol{\alpha}, \boldsymbol{\xi}) = \alpha_{0k} + \sum_{p=1}^{P} \alpha_{pk} t_{ip} + \sum_{l=1}^{L} \xi_{lk} x_{il}$. All predictions $\hat{m}_{ik}$ are compiled into the matrix $\hat{\mathbf{M}}(\boldsymbol{\alpha}, \boldsymbol{\xi})$.

Regression model for the outcome Y

If the outcome is continuous, we assume the following Gaussian model

$$Y = \gamma_0 + \sum_{p=1}^{P} \gamma_p T_p + \sum_{k=1}^{K} \beta_k M_k + \sum_{l=1}^{L} \psi_l X_l + \epsilon \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2).$$

We denote $\hat{y}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\psi})$ the prediction for the i^{th} individual:

$$\hat{y}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\psi}) = \gamma_0 + \sum_{p=1}^{P} \gamma_p t_{ip} + \sum_{k=1}^{K} \beta_k m_{ik} + \sum_{l=1}^{L} \psi_l x_{il}.$$

If the outcome is binary, we consider the following logistic model

$$\log\left(\frac{\mathbb{P}(Y=1)}{1-\mathbb{P}(Y=1)}\right) = \gamma_0 + \sum_{p=1}^{P} \gamma_p T_p + \sum_{k=1}^{K} \beta_k M_k + \sum_{l=1}^{L} \psi_l X_l.$$

In this case, $\hat{y}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\psi}) = e^{z_i}/(1 + e^{z_i})$ with $z_i = \gamma_0 + \sum_{p=1}^{P} \gamma_p t_{ip} + \sum_{k=1}^{K} \beta_k m_{ik} + \sum_{l=1}^{L} \psi_l x_{il}$. In both cases, all predictions $\hat{y}_i$ are compiled into the vector $\hat{\mathbf{y}}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\psi})$.

3 MAHI: a two-step algorithm for Mediation Analysis with HIgh-dimensional data

3.1 Step 1: from high to low dimension

The goal of the first step of our MAHI algorithm is to select a number $K_0 < n$ of candidate mediators to avoid the high-dimensional setting while retaining as many true mediators as possible. This step relies on an ad hoc loss function depending of the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\psi})$ and on a group lasso procedure with stability selection.

Definition of the loss functions

We consider the following loss functions for the regression models of the candidate mediators and the outcome:

$$\ell_{M_k}(\boldsymbol{\alpha}, \boldsymbol{\xi}) = \begin{cases} \frac{1}{2}\sum_{i=1}^{n} (\hat{m}_{ik} - m_{ik})^2 & \text{if } M_k \text{ is Gaussian} \\ \sum_{i=1}^{n} (-m_{ik}\nu_{ik} + \log(1 + e^{\nu_{ik}})) & \text{if } M_k \text{ is binary} \end{cases}$$

and

$$\ell_Y(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\psi}) = \begin{cases} \frac{1}{2}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 & \text{if } Y \text{ is Gaussian} \\ \sum_{i=1}^{n} (-y_i z_i + \log(1 + e^{z_i})) & \text{if } Y \text{ is binary.} \end{cases}$$

The loss function associated to the whole model is then defined as

$$f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\psi}) = \frac{1}{n} \sum_{k=1}^{K} w_k \ell_{M_k}(\boldsymbol{\alpha}, \boldsymbol{\xi}) + \frac{w_Y}{n} \ell_Y(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\psi}),$$

where $\mathbf{w} = (w_1, \ldots, w_K)$ is a vector of weights that allows to tune the relative importance of the treatment-mediator relationships and, the weight w_Y allows to tune the relative importance of the treatment-outcome and mediators-outcome relationships.

The group lasso and the proximal operator

The group lasso [18, 19] is used to select mediators by minimizing a penalized version of f, with a penalty that promotes sparsity by encouraging the nullity of some predefined groups of parameters. More precisely, let $\mathcal{G} = (G_1, \ldots, G_R)$ be a user-specified partition of the candidate mediators. For a group G_r, denote $\boldsymbol{\alpha}_{|G_r}$ and $\boldsymbol{\beta}_{|G_r}$ the subsets of the model coefficients corresponding to the candidate mediators in G_r, namely

$$\boldsymbol{\alpha}_{|G_r} = \{\alpha_{pk} | k \in G_r, p \in 1, \ldots, P\} \text{ and } \boldsymbol{\beta}_{|G_r} = \{\beta_k | k \in G_r\}.$$

Note that if $P = 1$ (i.e., there is only one treatment) and all candidate mediators form a single group, the overall idea is to employ a procedure where the coefficients α_k and β_k of each candidate mediator M_k are jointly selected either out of the model (false mediators) or into the model (promising candidate mediators that deserve further inspection). In the general case, with P treatments and any pre-specified groups of candidate mediators, the method will favor the selection of groups having a common mediation effect across the treatments.

The considered problem can then be written, for a given regularization parameter $\lambda > 0$, as

$$\underset{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\psi}}{\operatorname{argmin}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\psi}) + \lambda \sum_{r=1}^{R} \|(\boldsymbol{\alpha}_{|G_r}, \boldsymbol{\beta}_{|G_r})\|_2. \tag{1}$$

To solve this optimization problem, we employ the proximal method as described in [20]. This method relies on an iterative procedure described in Appendix A.

Stability selection and parameter choices

We emphasize that the goal of Step 1 is to transform the initial problem into a small-dimensional problem while discharging as few true mediators as possible. The selection of some false mediators is not problematic because Step 2 will test individual indirect effects. As a consequence, the values of the penalty parameter λ do not need to be fine-tuned and can be chosen loosely. Moreover, lasso selection is known to be highly unstable, a problem that can be addressed using the stability selection procedure introduced in [21]. The idea behind stability selection is to select variables based on the number of times they are chosen when running the original selection procedure on multiple bootstrap samples. Based on these two considerations, we propose the following procedure for the first selection step:

- The user chooses the vector $\mathbf{w}$ of the relative weights of the candidate mediators. By default, a constant vector is chosen, meaning that each candidate mediator is given the same importance.
- w_Y is not fixed. The default procedure is to use a grid of values.
- For each value of w_Y, the optimization procedure is run N_{boot} times on data subsamples, with N_{boot} large. For each of these subsamples, a value of λ is chosen by dichotomy such that the number of candidate mediators kept is in a pre-defined interval, by default $[n/2, n]$ where n is the sample size.
- Candidate mediators are then ranked from most to least frequently selected across all obtained lists. The rationale is that a true mediator should be selected more often than a non-mediating variable, which will only be selected occasionally as a false positive. Finally, the $K_{\max}$ best ranked mediators are selected by Step 1.
- The choice of $K_{\max}$ is guided by the fact that Step 2 is based on estimating the parameters of "classic" (i.e., non-penalized) regression models and that the number of the explanatory variables has to be chosen accordingly. For a continuous outcome the default value is $2n/\log(n)$. For a binary outcome $K_{\max}$ must at most be equal to the integer part of $-2 + n/50$ according to [22].

3.2 Step 2: estimation of direct and indirect effects

The second step of MAHI involves estimating and testing, for each treatment, the direct effect and the indirect effects through each of the selected candidate intermediate variables, which we denote $M_1, \ldots, M_{K_{\max}}$ (up to a permutation of the original indices). This is accomplished using the identifiability assumptions and method for low-dimensional multiple mediation analysis described in [17]. For clarity, we recall the corresponding quasi-Bayesian algorithm adapted from [2] and refer the reader to [17] for its theoretical justification.

Algorithm for low-dimensional multiple mediation analysis:

1. Fit parametric models for the outcome and the retained candidate mediators as in the previous section. We denote the vectors of parameter estimates as $\widehat{\Theta}_Y$ and $\widehat{\Theta}_Z = \left(\widehat{\Theta}^1, \ldots, \widehat{\Theta}^{K_{\max}}\right)$, respectively.
2. For each model, sample N times its parameters according to their multivariate sampling distribution, and obtain the vectors or parameters $\widehat{\Theta}_{Y(n)}$ and $\widehat{\Theta}_{Z(n)} = \left(\widehat{\Theta}^1_{(n)}, \ldots, \widehat{\Theta}^{K_{\max}}_{(n)}\right)$, for $n = 1, \ldots, N$. As in [2], the law of the parameters is approximated by a multivariate normal distribution, with mean and variance equal to the estimated parameters and their estimated asymptotic covariance matrix, respectively.
3. For each candidate mediator M_k, with $k = 1, \ldots, K_{\max}$, repeat I times the followings steps:
   - Simulate the counterfactual values of each mediator. In particular, for each pair $t, t' \in \{0, 1\}$, sample the vector of counterfactual variables $Z_k^{(i)}(t, t') = \left(M_k^{(i)}(t), W_k^{(i)}(t')\right)$, where W_k denotes the vector of all mediators but M_k.

- Simulate the counterfactual outcomes given the simulated values of the counterfactual mediators, denoted by $Y^{(i)}\left(t, Z_k^{(i)}(t', t)\right)$ for each k and $t, t' \in \{0, 1\}$.
- Estimate the individual mediation effects:

$$\hat{\delta}_{(r)}^k(t) = \frac{1}{I} \sum_{i=1}^{I} \left\{ Y_{(ri)}\left(t, Z_{(ri)}^k(1, t)\right) - Y_{(ri)}\left(t, Z_{(ri)}^k(0, t)\right) \right\}$$

4. From the empirical distribution of each effect above, obtain point estimates together with p-values and confidence intervals.

The final selection of mediators consists of the set of candidate mediators whose confidence intervals do not contain 0 after correction for the $K_{\max}$ multiple comparisons. Note that, as detailed in [17], this algorithm also allows for the estimation of the direct and joint mediated effects.

4 Simulation study

We ran simulations to validate MAHI and to compare it to methods recently introduced in the literature.

4.1 Models for simulated data

4.1.1 Continuous outcome

We simulated 100 datasets, including $n = 100$ observations and $K = 500$ candidate mediators each, according to the model

$$\begin{aligned} M_{ik} &= \mu_k + \alpha_k T_i + \epsilon_{ik} \\ Y_i &= 20 + 50 T_i + \textstyle\sum_k \beta_k M_{ik} + \epsilon_{i0} \end{aligned} \tag{2}$$

where $1 \leq i \leq n$ and $1 \leq k \leq K$. The only exposure variable T follows a Bernoulli distribution, $T \sim \mathcal{B}(0.3)$, μ_k is drawn uniformly in the interval $[-2, 2]$ for each variable M_k, and $\epsilon_k \sim \mathcal{N}(0, 1)$ for $k \in (0, \ldots, 500)$. Note that variables M_k are causally unrelated one to each other. Table 1 shows the values of α_k and β_k for the first 50 variables M_k. The higher the absolute value of $\alpha_k \beta_k$, the greater the indirect effect through M_k. As such, the first 10 mediators have strong indirect effects (and are, in principle, easier to select), the next 10 have mild indirect effects (less easy to detect) and the next 10 have weak indirect effects (hard to detect). All other 470 variables M_k are not true mediators because either $\alpha_k = 0$ or $\beta_k = 0$.

k	1	2	3	4	5	6	7	8	9	10
α_k	-95	90	95	95	-100	95	-95	85	-95	-100
β_k	185	-195	190	185	-190	185	195	-190	100	185
k	11	12	13	14	15	16	17	18	19	20
α_k	-2.75	3.25	-3.50	2.50	-3.75	3.00	-3.25	2.75	3.00	3.50
β_k	3.25	-2.50	3.75	-3.00	3.50	-2.75	3.75	-3.25	3.25	-2.75
k	21	22	23	24	25	26	27	28	29	30
α_k	-0.875	0.625	-0.375	-0.25	0.50	-0.875	0.125	-1.125	0.375	-0.25
β_k	0.375	-0.625	0.625	-1.125	0.75	0.25	-0.50	0.375	-0.625	0.75
k	31	32	33	34	35	36	37	38	39	40
α_k	25	25	25	25	25	25	25	25	25	25
β_k	0	0	0	0	0	0	0	0	0	0
k	41	42	43	44	45	46	47	48	49	50
α_k	0	0	0	0	0	0	0	0	0	0
β_k	45	45	45	45	45	45	45	45	45	45

Table 1: Values of α_k and β_k for $k = 1, \ldots, 50$. For $k = 51, \ldots, 500$, $\alpha_k = \beta_k = 0$.

4.1.2 Binary outcome

We simulated 100 datasets, including $n = 1350$ observations and $K = 2000$ candidate mediators each, according to the model

$$\begin{aligned} M_{ik} &= 1 + \alpha_k T_i + \epsilon_{ik} \\ Y_i^* &= -65 + T_i + \textstyle\sum_k \beta_k M_{ik} + \epsilon_{i0} \\ Y_i &= \mathbb{1}_{Y_i^* > 0} \end{aligned} \tag{3}$$

where $1 \leq i \leq n$ and $1 \leq k \leq K$. The only exposure variable T follows a Bernoulli distribution, $T \sim \mathcal{B}(0.3)$, the residual ϵ_0 follows a logistic distribution, $\epsilon_0 \sim \mathcal{L}(0, 1)$, and $\epsilon_k \sim \mathcal{N}(0, 1)$ for $k \in (1, \ldots, 2000)$. Note that mediators are causally independent. As shown in Table 2, the 15 true mediators $M_1, \ldots, M_{15}$ are split in three groups of 5 mediators each, with strong, mild and weak mediated effects respectively.

k	1	2	3	4	5	6	7	8	9	10
α_k	2	2	2	2	2	1	1	1	1	1
β_k	2	2	2	2	2	1	1	1	1	1
k	11	12	13	14	15	16	17	18	19	20
α_k	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
β_k	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0
k	21	22	23	24	25	26	27	28	29	30
α_k	0	0	0	0	0	0	0	0	0	0
β_k	5	5	5	5	5	0	0	0	0	0

Table 2: Values of α_k and β_k for $k = 1, \ldots, 30$. For $k = 31, \ldots, 2000$, $\alpha_k = \beta_k = 0$.

4.2 Methods settings

We implemented our method in the `mahi` function of the GitHub R package `AllanJe/mahi`. For our simulation study, we considered $N_{boot} = 30$ and selected λ to retain between $n/2$ and n candidate mediators. This constraint ensures that the second step no longer deals with a high-dimensional setting. For the second step, p-values were adjusted using the Hochberg correction with a threshold of 0.05.

4.2.1 Comparison to state-of-the-art methods for continuous outcomes

We compared MAHI to the following six alternative methods on simulated data with continuous outcomes:

1. [10] introduced an approach for high-dimensional mediation analysis, called the *Coordinate-wise Mediation Filter* (CMF). The CMF implementation consists of two components: an internal algorithm which performs the selection of mediators by coordinate descent using a decision function D, and an external algorithm that runs several times the internal algorithm and aggregates the corresponding outputs. The entire procedure is implemented in the GitHub R package `vankesteren/cmfilter`. In our simulations, the decision function is the Sobel test. The external algorithm is run 1000 times. Once the selection rate for each mediator is calculated, a mediator is chosen if its selection rate is greater than 0.079, the value recommended by the authors.
2. [14] introduced the HIMA (*HIgh-dimensional Mediation Analysis*) algorithm, which is based on penalized regressions and uses a lasso-type penalty function called the concave minimax penalty (MCP) [23]. The HIMA implementation consists of three steps: first the set of candidate mediators is reduced by means of the sure independent screening (SIS) method [24], then the estimates $\widehat{\beta_k}$ are calculated using the MCP penalization criterion, and at last indirect effects are tested and p-values are adjusted according to the Bonferroni correction. The entire procedure is implemented and available in the R package `hima`. In our simulations, we chose the first $n/\log(n)$ mediators obtained with the SIS method, as recommended by the authors.
3. [25] proposed a variation of HIMA allowing the selection of correlated candidate mediators, called HDMA (*High-Dimensional Mediation Analysis*). The HDMA method differs from HIMA in the second step, where debiased estimates of $\widehat{\beta_k}$ are calculated. The entire procedure is available in the GitHub R package `YuzhaoGao/High-dimensional-mediation-analysis-R`. In our simulations the settings are the same as for HIMA.
4. [12] introduced the BAMA (*Bayesian Mediation Analysis*) approach. It is a Bayesian inference method using continuous shrinkage priors to extend previous causal mediation analyses techniques to a high-dimensional setting. For each candidate mediator, the posterior inclusion probability (PIP) is estimated measuring the association strength between exposure and mediators and between mediators and outcome. The candidate mediators with the highest PIP are selected as the active mediators. The entire procedure is implemented and available in the R package `bama`. In our simulations we chose a PIP threshold of 0.1 for selection.

5. [26] introduced the SPCMA (*Sparse Principal Component Mediation Analysis*) algorithm. When candidate mediators are potentially causally related to one another, one approach is to perform a principal component analysis (PCA) to obtain orthogonal principal components (PCs), which can be treated as new, conditionally independent mediators. However, these new candidate mediators, which are linear combinations of the original candidate mediators, can be difficult to interpret. The sparse high-dimensional mediation analysis approach proposed in [26] applies PCA with sparse loadings, making the principal components more interpretable as they are linear combinations of a few original candidate mediators. The entire procedure is implemented in the GitHub R package `zhaoyi1026/spcma`. In our simulations, variables M_k are causally independent so we used the function recommended by the authors in this case, which performs marginal causal mediation analysis under the linear structural equation modeling framework.
6. [27, 28] introduced the HDMAX2 procedure (*High Dimensional mediation analysis with* $\max^2$ *test*). The selection procedure of HDMAX2 involves fitting latent factor mixed models (LFMMs, [29]) to estimate the effects of exposure on mediators and the effect of each mediator on the outcome. For each candidate mediator, two p-values (P_x and P_y) are derived from these models, testing the null hypotheses of no effect of exposure on the mediator and no effect of the mediator on the outcome, respectively. Candidate mediators are then selected using the $\max^2$ test, a novel test that uses the p-value $p = \max\{P_x, P_y\}^2$. Similar to the Sobel test, the $\max^2$ test rejects the null hypothesis that either the effect of exposure on the mediator or the effect of the mediator on the outcome is null. The selected candidate mediators are subsequently ranked by significance, and only those below a given threshold proceed to the second step. This step consists of performing simple mediation analyses for each selected candidate mediator using the `mediation` package [30] to estimate and test their indirect effects. The threshold can be determined using data-adaptive approaches, such as false discovery rate (FDR) control, or set manually by the user. In our study, we retained the 50 candidate mediators with the lowest $\max^2$ p-values. HDMAX2 is available in the GitHub R package `bcm-uga/hdmax2`.

4.2.2 Comparison to state-of-the-art methods for binary outcomes

We compared MAHI to HIMA, HDMA and HDMAX2, all of which can also be applied to binary outcomes. After the first step of MAHI, we retained the top $\left\lfloor \frac{n}{50} - 2 \right\rfloor$ candidate mediators to proceed to the second step. For the three other methods we proceeded as follows :

1. For HIMA, we chose the first $\lceil n/(2\log(n)) \rceil$ candidate mediators obtained with the SIS method, as recommended by the authors for a binary outcome.
2. For HDMA, we also chose the first $\lceil n/(2\log(n)) \rceil$ mediators obtained with the SIS method, as recommended by the authors for a binary outcome.
3. For HDMAX2, we retained the top 25 candidate mediators at the end of the first step to proceed to the second step. We then applied the Hochberg correction to the results of the second step at a threshold of 0.05.

Note that the implementations of HIMA and HDMA allow to choose different penalisation methods to obtain sparsity. We run them all, which explains the multiple results for each of the methods in Table 4.

4.3 Results

Table 3 and Table 4 show, for each method, the mean of three performance metrics, namely precision (or positive predictive value), recall and specificity, over 100 replicates, for continuous and binary outcomes respectively. In particular, the metrics are defined with respect to four selection problems:

- the selection of all true mediators,
- the selection of strong mediators,
- the selection of mild mediators,
- the selection of weak mediators.

Figure B1 and Figure B3 show the distribution of the three metrics over 100 replicates, for continuous and binary outcomes respectively. Figure B2 and Figure B4 show the false discovery rate (1-precision), the false negative rate (1-recall) and the false positive rate (1-specificity). We considered three selection problems:

- the selection of false mediators,
- the selection of false mediators with $\alpha_k \neq 0$ and $\beta_k = 0$,
- the selection of false mediators with $\alpha_k = 0$ and $\beta_k \neq 0$.

4.3.1 Results, continuous outcomes

Table 3 shows that our method MAHI had an overall precision, or positive predictive value, close to 100%, meaning the almost all selected candidate mediators were true mediators, and that it was the most precise method among the tested approaches. The mean recall of MAHI was 35%, meaning that 35% of the true mediators were actually selected. In particular, MAHI detected only a few mild mediators and no weak mediators at all. The precision of HDMAX2 was close to 100% and its recall was as low as 16%, as it only selected 50% of the true strong mediators and none of the mild and weak mediators. Similarly, the precision of CMF was greater than 80% but its recall was as low as 17%. BAMA had the fourth best precision (67%), but achieved the best recall (64%). Indeed almost all strong and, notably, true mild mediators were selected by BAMA. Even though MCMA ranked fifth according to precision (53%), it did not performed well on model (2) as its recall was as low as 6%. HIMA was slightly less precise (41%) but selected 31% of the true mediators. HDMA had the lowest precision, as in average only 24% of the selected mediators were true mediators, but its recall (34%) was close to those of the best performing methods. The specificity was close to 100% for all methods, which is expected given the small proportion of true mediators.

4.3.2 Results, binary outcome

Table 4 shows that MAHI and HDMAX2 had the best precision, nearly 100%. MAHI also achieved the best recall, with an average of only 23% of the true mediators not

	Method	Precision	Recall	Specificity
All true mediators	MAHI	**0.998**	0.355	0.999
	CMF	0.820	0.167	0.997
	HIMA	0.412	0.306	0.970
	HDMA	0.236	0.346	0.928
	BAMA	0.673	**0.639**	0.980
	MCMA	0.535	0.058	0.944
	HDMAX2	0.991	0.161	**1.000**
Strong mediators	MAHI	0.933	**0.988**	0.998
	CMF	0.671	0.410	0.996
	HIMA	0.374	0.828	0.969
	HDMA	0.197	0.868	0.928
	BAMA	0.326	0.926	0.961
	MCMA	0.535	0.026	0.981
	HDMAX2	**0.989**	0.482	**1.000**
Medium mediators	MAHI	0.064	0.076	0.980
	CMF	0.139	0.084	**0.989**
	HIMA	0.025	0.061	0.953
	HDMA	0.025	0.111	0.912
	BAMA	0.327	**0.930**	0.961
	MCMA	**0.465**	0.019	0.981
	HDMAX2	0.001	0.001	**0.990**
Weak mediators	MAHI	0.000	0.000	0.978
	CMF	0.011	0.008	0.987
	HIMA	0.013	0.029	0.953
	HDMA	0.014	**0.060**	0.911
	BAMA	0.020	**0.060**	0.943
	MCMA	**0.465**	0.019	0.981
	HDMAX2	0.000	0.000	**0.990**

Table 3: Comparison of high-dimensional mediation analysis methods with regards to the ability to select the true mediators $M_1, \ldots, M_{15}$: mean precision, recall and specificity over the 100 data sets simulated with **continuous** outcomes according to model (2).

being selected. More specifically, the selection of weak true mediators was particularly challenging, as more than 63% of them were not selected. In comparison, the recalls of HDMA_lasso, HIMA_MCP, and HDMAX2 were 62%, 30%, and 30%, respectively.

	Method	Precision	Recall	Specificity
All true mediators	MAHI	**0.992**	**0.767**	**1.000**
	HIMA_lasso	0.760	0.153	0.999
	HIMA_MCP	0.742	0.304	0.999
	HIMA_SCAD	0.687	0.225	0.998
	HDMA_lasso	0.717	0.619	0.998
	HDMA_ridge	0.709	0.522	0.998
	HDMAX2	0.991	0.295	**1.000**
Strong mediators	MAHI	0.438	**1.000**	0.997
	HIMA_lasso	0.583	0.324	0.999
	HIMA_MCP	0.547	0.622	0.998
	HIMA_SCAD	0.500	0.430	0.998
	HDMA_lasso	0.390	0.994	0.996
	HDMA_ridge	0.448	0.962	0.997
	HDMAX2	**0.837**	0.732	**1.000**
Medium mediators	MAHI	**0.406**	**0.936**	0.997
	HIMA_lasso	0.180	0.102	0.999
	HIMA_MCP	0.173	0.224	0.997
	HIMA_SCAD	0.206	0.180	0.998
	HDMA_lasso	0.241	0.632	0.995
	HDMA_ridge	0.198	0.456	0.995
	HDMAX2	0.126	0.128	0.998
Weak mediators	MAHI	**0.148**	**0.366**	0.995
	HIMA_lasso	0.097	0.030	0.998
	HIMA_MCP	0.072	0.066	0.997
	HIMA_SCAD	0.131	0.062	0.997
	HDMA_lasso	0.085	0.230	0.994
	HDMA_ridge	0.064	0.148	0.995
	HDMAX2	0.028	0.026	0.998

Table 4: Comparison of high-dimensional mediation analysis methods with regards to the ability to select the true mediators $M_1, \ldots, M_{15}$: mean precision, recall and specificity over the 100 data sets simulated with **binary** outcomes according to model (3).

5 Illustration on real data : mediation of smoking on rheumatoid arthritis outcomes

5.1 Biological context

Rheumatoid Arthritis (RA) is a chronic inflammatory disease influenced by both genetic and environmental factors. Smoking has been identified as one of the most important extrinsic risk factor for its development and severity [31]. DNA methylation (DNAm), an epigenetic mechanism that involves the methylation of specific bases in

the DNA strand, can regulate gene transcription, thereby affecting disease development. The relationship between DNAm levels and RA occurence was first investigated in [32]. In addition, several association studies have already established the impact of tobacco consumption on DNAm [33]. As a case study, we explored to which extent DNAm mediates the effect of tobacco consumption on the occurrence of RA. The dataset was collected from the Gene Expression Omnibus (GEO) database using the accession number GSE42861 [32]. It consists of Illumina HumanMethylation450 BeadChip array in peripheral blood leukocytes (PBLs) from RA patients (n = 354) and normal controls (n = 333). Clinical data including age, gender, smoking status and residential area were provided for each sample. Two patients were excluded from the analysis because their smoking status was unknown.

5.2 Mediation analysis

To proceed with the mediation analysis, the categorial smoking status variable was transformed into a binary variable: 0 for never and non-regular cigarette smokers and 1 for former and current cigarette smokers. Additionally, age and gender were included as adjustment variables in the model. Due to the very large initial number of probes and the resulting computational issues, a preliminary selection was done using the HDMAX2 method. This method has shown its effectiveness in identifying DNAm markers in a high-dimensional mediation analysis [27]. First, we used the *hdmax2_step1* approach to run association studies for all potential mediators and to test the significance of the estimated indirect effects. Then we applied a filter to select the top 1000 probes with the most significant P-values (Figure 2A). The resulting subset of DNAm probes is still high-dimensional but computationally less expensive. Subsequently, the MAHI method was applied to this refined subset of DNAm probes. Mediated ORs, corresponding to the indirect effect mediated by DNAm probes, were estimated for the selected subset of CpGs along with their CI. The top 50 CpGs mediators are depicted in Figure 2B.

5.3 Biological interpretation

Table 5 summarizes the results and relevant biological information for the selected CpG mediators that show ORs greater than 1.10 and lower than 0.9. When OR values are lower than 0.9, occurrence of RA is significantly reduced. In this context, our method identified two CpG mediators (cg04332373 and cg16854986) for which methylation appears to decrease in RA patients. When OR values are greater than 1.1, the occurence of RA is significantly increased. Interestingly, we observe varying scenarios in terms of indirect effects for ORs ≥ 1.1. In some instances, the methylation of CpG mediators decreases in RA patients compared to controls (e.g. cg23314866, cg15702277 and cg22446264), while in other cases, it increases (e.g. cg07119168, cg12916723 and cg15956469). This illustrates complex mediation pathways, suggesting that different biological processes are likely at play. We also examined whether some genes associated with the selected CpG mediators were previously known in the literature to be linked to RA (Table 5, "Pubmed hits" column). Interestingly, our approach not only identified known candidates (i.e. *CD38*) but also discovered new

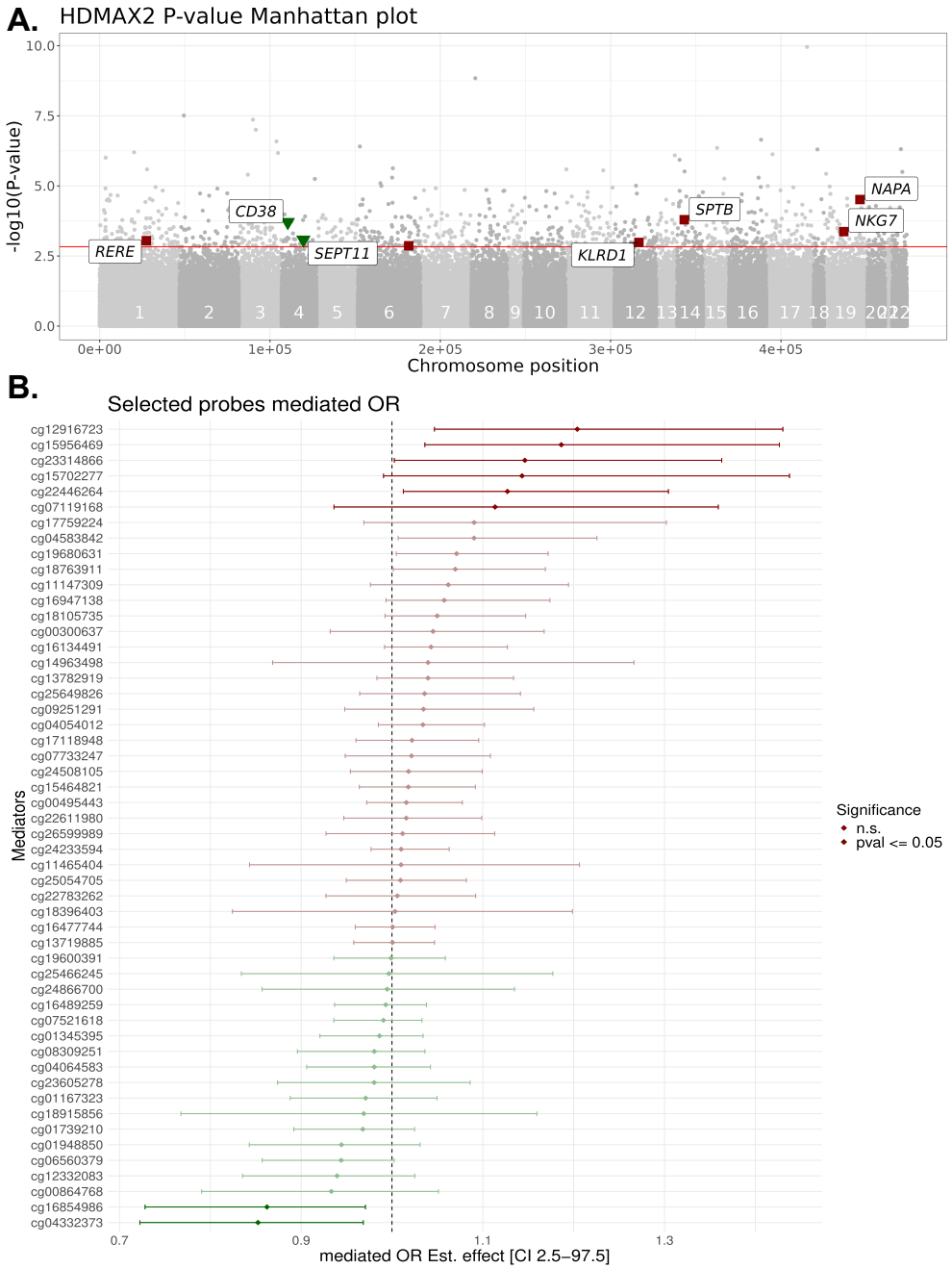

Fig. 2: Summary of mediation analysis of smoking on RA occurrence through DNA methylation. **A** Manhattan plot displaying the –log10 transformed P values estimated using the max-squared method (HDMAX2) for each CpG site. Each dot represents an individual CpG, ordered on the x-axis according to their genomic position. The red line indicates the threshold for the top $1,000$ CpGs selected for further analysis, on which MAHI was applied. Red squares represent probes with MAHI ORs greater than 1.10, while green triangles represent probes with MAHI ORs lower than 0.9. Labels correspond to genes associated with the selected probes, if any. Chromosome numbers are labeled in white. **B** Mediated ORs for the top 50 mediators. The estimate effect is represented by a dot and its 95% CI by the bar. Symbols correspond to the significance cut off of 5% (square for P-value ≥ 0.05, circle P-value < 0.05). Colors correspond to the sign and importance of the effect (dark green for estimated OR under 0.9, light green for estimated OR between 0.9 and 1, pink for estimated OR between 1 and 1.1 and dark red for estimated OR over 1.1).

probes that had not previously been associated with RA, opening the way to new research perspectives and experimental validation.

CpG Probes	mediated OR	mean DNAm cases	mean DNAm controls	Chr	Associated genes	Pubmed hits
OR less than 0.90						
cg04332373	0.85[0.72, 0.97]**	0.20 ± 0.03	0.22 ± 0.03	chr4	*CD38*	147
cg16854986	0.86[0.73, 0.97]**	0.12 ± 0.03	0.13 ± 0.04	chr4	*SEPT11*	0
OR more than 1.10						
cg23314866	1.15[1.00, 1.36]**	0.24 ± 0.05	0.29 ± 0.04	chr19	*NAPA*	1
cg07119168	1.11[0.94, 1.35]	0.81 ± 0.03	0.78 ± 0.04	chr14	*SPTB*	3
cg12916723	1.20[1.04, 1.43]***	0.63 ± 0.03	0.61 ± 0.04	chr19	*NKG7*	2
cg15702277	1.14[0.99, 1.44]*	0.25 ± 0.05	0.31 ± 0.05	chr1	*RERE*	0
cg15956469	1.18[1.04, 1.43]**	0.89 ± 0.04	0.85 ± 0.05	chr12	*KLRD1*	8
cg22446264	1.13[1.01, 1.30]**	0.47 ± 0.07	0.54 ± 0.07	chr6	-	-

Table 5: For each selected probes: mediated OR (with CI, *, **, *** , res. significant OR at 5%, 1% and 0.1% type I error), DNAm mean ± standard deviation for cases group and controls group, chromosome in which probe is located, nearest gene (identified using Illumina annotations), and the number of Pubmed matching hits with gene symbol and RA.

6 Discussion and conclusion

In this article we introduced MAHI, a two step-procedure for high-dimensional mediation analysis where the candidate intermediate variables outnumbers the available observations. MAHI first performs variable selection in the pool of candidate mediators through a group lasso penalty that we adapted specifically to the mediation problem. Then, MAHI estimates and tests the direct and indirect causal effects in the resulting lower-dimensional mediation model using the multiple mediation analysis method we developed in [17].

On simulated data, MAHI achieved better results compared to competing methods. More precisely, it outperformed existing methods in precision, recall and specificity when applied to binary outcomes. On simulated data with continuous outcomes, MAHI had the best precision and specificity but a lower recall than BAMA. More specifically, MAHI missed the true mediators with a mild effect, which were still selected by BAMA. However, BAMA also selected false mediators, particularly those causally linked to the exposure but not to the outcome. Indeed, the posterior inclusion probabilities of false

mediators not linked to the outcome were similar to those of true mediators (data not shown). On the contrary, MAHI almost never selected such false mediators. Moreover, it is important to stress that the performance of BAMA depends on a user-specified PIP threshold, for which the choice criterion is not straightforward.

The principal methodological novelty of this work is the mediator selection step of MAHI. Our simulation results suggest that integrating this initial step with our previously developed inferential algorithm yields highly satisfactory performance. However, it is important to note that our mediation selection procedure can, in principle, be implemented prior to any method designed for low-dimensional analysis. Nevertheless, when handling correlated candidate mediators, we suggest following through with the second step of MAHI, as detailed in this article. When dealing with an extremely large number of candidate mediators, such as hundreds of thousands, the current R implementation of MAHI may become computationally ineffective. In these cases, we recommend running mediator pre-selection with the fast first step of the HDMAX2 approach. We employed this strategy combining the first step of HDMAX2 and MAHI to detect and assess the role of DNA CpG site methylation in mediating the impact of smoking on the occurrence of rheumatoid arthritis and identified 8 significant probes. Remarkably, one of the 8 selected probes was associated with the *CD38* gene, which shows a strong association with RA in PubMed research, with 147 hits. CD38 is important in the regulation of innate immunity [34] and has already been identified as a potential therapeutical target for autoimmune diseases such as RA, but also systemic lupus or multiple sclerosis [35].

In the first step, MAHI can take into account user-defined groups of candidate mediators. This ability is especially valuable for genomic applications, where the focus is frequently on evaluating the mediated effects of specific genomic regions. Note that [36] had already proposed a multiple testing procedure to determine which groups had a significant mediating effect. However, MAHI is to our knowledge the first screening method capable of taking group structure into account, as well as considering several treatments simultaneously and promoting the selection of common mediators. This interesting features allows to select candidate mediators with indirect effects with respect to all exposures and to discharge intermediate variables that act as mediators only with respect to some of the exposures.

Several methodological questions remain open and constitute challenging tasks for the future. Notably, it would be interesting to adapt MAHI to other types of data, in particular to longitudinal data and/or survival models. A second major question is the sensitivity of the method to violations of the conditional independence conditions upon which the identification of mediated effects relies (see, for instance, [17]). To the best of our knowledge, this challenge has not yet been addressed in the setting of high-dimensional mediation analysis.

Method availability:

The MAHI method is available as an R package at `https://github.com/AllanJe/mahi`.

References

[1] Pearl, J.: Direct and Indirect Effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. UAI'01, pp. 411–420. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)

[2] Imaï, K., Keele, L., Tingley, D.: A general approach to causal mediation analysis. Psychological Methods **15**(4), 309–334 (2010)

[3] Blum, M.G., Valeri, L., François, O., Cadiou, S., Siroux, V., Lepeule, J., Slama, R.: Challenges raised by mediation analysis in a high-dimension setting. Environmental health perspectives **128**(5), 055001 (2020)

[4] Han, Q., Wang, Y., Sun, N., Chu, J., Hu, W., Shen, Y.: Mediation analysis method review of high throughput data. Statistical Applications in Genetics and Molecular Biology **22**(1), 20230031 (2023)

[5] Huang, Y.-T., Pan, W.-C.: Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators: Hypothesis Test of Mediation Effect in Causal Mediation Model with High-Dimensional Continuous Mediators. Biometrics **72**(2), 402–413 (2016) https://doi.org/10.1111/biom.12421 . Accessed 2020-06-03

[6] Han, X., Peng, J., Cui, A., Zhao, F.: Sparse Principal Component Analysis via Fractional Function Regularity. Mathematical Problems in Engineering **2020**, 1–10 (2020) https://doi.org/10.1155/2020/7874140 . Accessed 2021-06-03

[7] Assi, N., Fages, A., Vineis, P., Chadeau-Hyam, M., Stepien, M., Duarte-Salles, T., Byrnes, G., Boumaza, H., Knüppel, S., Kühn, T., Palli, D., Bamia, C., Boshuizen, H., Bonet, C., Overvad, K., Johansson, M., Travis, R., Gunter, M.J., Lund, E., Dossus, L., Elena-Herrmann, B., Riboli, E., Jenab, M., Viallon, V., Ferrari, P.: A statistical framework to model the meeting-in-the-middle principle using metabolomic data: application to hepatocellular carcinoma in the EPIC study. Mutagenesis, 045 (2015) https://doi.org/10.1093/mutage/gev045 . Accessed 2021-06-03

[8] Chén, O.Y., Crainiceanu, C., Ogburn, E.L., Caffo, B.S., Wager, T.D., Lindquist, M.A.: High-dimensional multivariate mediation with application to neuroimaging data. Biostatistics **19**(2), 121–136 (2018) https://doi.org/10.1093/biostatistics/kxx027 . Accessed 2020-06-02

[9] Geuter, S., Reynolds Losin, E.A., Roy, M., Atlas, L.Y., Schmidt, L., Krishnan, A., Koban, L., Wager, T.D., Lindquist, M.A.: Multiple brain networks mediating stimulus–pain relationships in humans. Cerebral Cortex **30**(7), 4204–4219 (2020)

[10] Kesteren, E.-J., Oberski, D.L.: Exploratory Mediation Analysis with Many Potential Mediators. Structural Equation Modeling: A Multidisciplinary Journal

26(5), 710–723 (2019) https://doi.org/10.1080/10705511.2019.1588124 . Accessed 2020-06-02

[11] Derkach, A., Pfeiffer, R.M., Chen, T., Sampson, J.N.: High dimensional mediation analysis with latent variables. Biometrics **75**(3), 745–756 (2019) https://doi.org/10.1111/biom.13053 . Accessed 2021-06-03

[12] Song, Y., Zhou, X., Zhang, M., Zhao, W., Liu, Y., Kardia, S.L.R., Diez Roux, A.V., Needham, B.L., Smith, J.A., Mukherjee, B.: Bayesian Shrinkage Estimation of High Dimensional Causal Mediation Effects in Omics Studies. preprint, Epidemiology (November 2018). https://doi.org/10.1101/467399 . http://biorxiv.org/lookup/doi/10.1101/467399 Accessed 2020-06-02

[13] Jumentier, B., Barrot, C.-C., Estavoyer, M., Tost, J., Heude, B., François, O., Lepeule, J.: High-dimensional mediation analysis: a new method applied to maternal smoking, placental dna methylation, and birth outcomes. Environmental Health Perspectives **131**(4), 047011 (2023)

[14] Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., Vokonas, P., Zhao, L., Lv, J., Baccarelli, A., Hou, L., Liu, L.: Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics **32**(20), 3150–3154 (2016) https://doi.org/10.1093/bioinformatics/btw351 . Accessed 2020-06-02

[15] Perera, C., Zhang, H., Zheng, Y., Hou, L., Qu, A., Zheng, C., Xie, K., Liu, L.: Hima2: high-dimensional mediation analysis and its application in epigenome-wide dna methylation data. BMC bioinformatics **23**(1), 296 (2022)

[16] Loh, W.W., Moerkerke, B., Loeys, T., Vansteelandt, S.: Non-linear Mediation Analysis with High-dimensional Mediators whose Causal Structure is Unknown. arXiv:2001.07147 [stat] (2020). arXiv: 2001.07147. Accessed 2020-06-02

[17] Jérolon, A., Baglietto, L., Birmelé, E., Alarcon, F., Perduca, V.: Causal mediation analysis in presence of multiple mediators uncausally related. The International Journal of Biostatistics **0**(0) (2020) https://doi.org/10.1515/ijb-2019-0088 . Accessed 2020-10-22

[18] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**(1), 49–67 (2006) https://doi.org/10.1111/j.1467-9868.2005.00532.x . Accessed 2021-06-03

[19] Meier, L., Van De Geer, S., Bühlmann, P.: The group lasso for logistic regression: Group Lasso for Logistic Regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(1), 53–71 (2008) https://doi.org/10.1111/j.1467-9868.2007.00627.x . Accessed 2021-06-03

[20] Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. arXiv preprint arXiv:1108.0775 (2011)

[21] Meinshausen, N., Bühlmann, P.: Stability selection: Stability Selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(4), 417–473 (2010) https://doi.org/10.1111/j.1467-9868.2010.00740.x . Accessed 2020-06-01

[22] Bujang, M.A., Sa'at, N., Biostatistics Unit, National Clinical Research Centre, Ministry of Health, Kuala Lumpur, Malaysia, Tg Abu Bakar Sidik, T.M.I., Biostatistics Unit, National Clinical Research Centre, Ministry of Health, Kuala Lumpur, Malaysia, Chien Joo, L., Clinical Research Centre, Sarawak General Hospital, Ministry of Health, Kuching, Malaysia: Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. Malaysian Journal of Medical Sciences **25**(4), 122–130 (2018) https://doi.org/10.21315/mjms2018.25.4.12 . Accessed 2022-10-19

[23] Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics **38**(2), 894–942 (2010) https://doi.org/10.1214/09-AOS729 . Accessed 2020-05-15

[24] Fan, J., Lv, J.: Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(5), 849–911 (2008) https://doi.org/10.1111/j.1467-9868.2008.00674.x . Accessed 2020-05-15

[25] Gao, Y., Yang, H., Fang, R., Zhang, Y., Goode, E.L., Cui, Y.: Testing Mediation Effects in High-Dimensional Epigenetic Studies. Frontiers in Genetics **10**, 1195 (2019) https://doi.org/10.3389/fgene.2019.01195 . Accessed 2020-06-02

[26] Zhao, Y., Lindquist, M.A., Caffo, B.S.: Sparse principal component based high-dimensional mediation analysis. Computational Statistics & Data Analysis **142**, 106835 (2020) https://doi.org/10.1016/j.csda.2019.106835 . Accessed 2021-07-08

[27] Jumentier, B., Barrot, C.-C., Estavoyer, M., Tost, J., Heude, B., François, O., Lepeule, J.: High-Dimensional Mediation Analysis: A New Method Applied to Maternal Smoking, Placental DNA Methylation, and Birth Outcomes. Environmental Health Perspectives **131**(4), 047011 https://doi.org/10.1289/EHP11559 . Publisher: Environmental Health Perspectives. Accessed 2023-09-19

[28] Pittion, F., Jumentier, B., Nakamura, A., Lepeule, J., François, O., Richard, M.: hdmax2, an R package to perform high dimension mediation analysis. working paper or preprint (2024). https://hal.science/hal-04658960

[29] Caye, K., Jumentier, B., Lepeule, J., François, O.: Lfmm 2: fast and accurate inference of gene-environment associations in genome-wide studies. Molecular biology and evolution **36**(4), 852–860 (2019)

[30] Tingley, D., Yamamoto, T., Hirose, K., Keele, L., Imai, K.: mediation: R package for causal mediation analysis. Journal of Statistical Software **59**(5), 1–38 (2014) https://doi.org/10.18637/jss.v059.i05

[31] Chang, K., Yang, S.M., Kim, S.H., Han, K.H., Park, S.J., Shin, J.I.: Smoking and rheumatoid arthritis. Int J Mol Sci **15**(12), 22279–22295 (2014)

[32] Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., Shchetynsky, K., Scheynius, A., Kere, J., Alfredsson, L., Klareskog, L., Ekström, T.J., Feinberg, A.P.: Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol **31**(2), 142–147 (2013)

[33] Kaur, G., Begum, R., Thota, S., Batra, S.: A systematic review of smoking-related epigenetic alterations. Arch Toxicol **93**(10), 2715–2740 (2019)

[34] Ye, X., Zhao, Y., Ma, W., Ares, I., Martínez, M., Lopez-Torres, B., Martínez-Larrañaga, M.-R., Wang, X., Anadón, A., Martínez, M.-A.: The potential of CD38 protein as a target for autoimmune diseases. Autoimmunity Reviews **22**(4), 103289 (2023) https://doi.org/10.1016/j.autrev.2023.103289 . Accessed 2024-07-04

[35] Peclat, T.R., Shi, B., Varga, J., Chini, E.N.: The NADase enzyme CD38: an emerging pharmacological target for systemic sclerosis, systemic lupus erythematosus and rheumatoid arthritis. Curr Opin Rheumatol **32**(6), 488–496 (2020)

[36] Djordjilović, V., Page, C.M., Gran, J.M., Nøst, T.H., Sandanger, T.M., Veierød, M.B., Thoresen, M.: Global test for high-dimensional mediation: Testing groups of potential mediators. Statistics in Medicine, 8199 (2019) https://doi.org/10.1002/sim.8199 . Accessed 2021-06-03

Appendix A Theoretical details

We describe how we solve the optimization problem 1 with the proximal method. This method can be written, with $v = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\psi})$ and $\Omega(v) = \sum_{r=1}^{R} \|(\boldsymbol{\alpha}_{|G_r}, \boldsymbol{\beta}_{|G_r})\|_2$, as

$$v^{t+1} = \underset{v}{\operatorname{argmin}} \quad f(v^t) + < \nabla f(v^t), v - v^t > + \lambda\Omega(v) + \frac{L}{2}\|v - v^t\|_2^2$$

for a well-chosen L. It can also be rewritten as

$$\begin{aligned} v^{t+1} &= \underset{v}{\operatorname{argmin}} \quad \frac{1}{2}\left\| v - \left(v^t - \frac{1}{L}\nabla f(v^t) \right) \right\|_2^2 + \frac{\lambda}{L}\Omega(v) \\ &= \operatorname{Prox}_{\frac{\lambda}{L}\Omega}\left(v^t - \frac{1}{L}\nabla f(v^t) \right) \end{aligned}$$

where the proximal operator is defined as

$$\operatorname{Prox}_{\mu\Omega}(u) = \underset{w}{\operatorname{argmin}} \quad \frac{1}{2}\|w - u\|_2^2 + \mu\Omega(w).$$

When Ω is a group lasso penalty, the proximal operator is known. In the present case, denoting by u_{G_r} the subvector of u whose coordinates correspond to those of $\boldsymbol{\alpha}_{|G_r}$ and $\boldsymbol{\beta}_{|G_r}$, it is computed by replacing for each G_r the vector u_{G_r} by

$$[\operatorname{Prox}_{\mu\Omega}(u)]_{G_r} = \max\left\{ 0, \left(1 - \frac{\mu}{\|u_{G_r}\|_2} \right) u_{G_r} \right\}.$$

The choice of L is again made according to [20] by increasing it until the former proximal solution verifies

$$f(v^{t+1}) \leq f(v^t) + < \nabla f(v^t), v^{t+1} - v^t > + \frac{L}{2}\|v^{t+1} - v^t\|_2^2.$$

Computing the gradient

In order to run the proximal method to select a subset of candidate mediators, the only step still needed is to compute the gradient of the loss function, which is easily done by the following result.

Theorem 1. *Let $\nabla_{\boldsymbol{\alpha}} f$ (respectively $\nabla_{\boldsymbol{\xi}} f$) be the matrix regrouping all the partial derivatives $\frac{\partial f}{\partial \alpha_{pk}}$ (respectively $\frac{\partial f}{\partial \xi_{lk}}$). Similarly, denote by $\nabla_{\boldsymbol{\beta}} f$, $\nabla_{\boldsymbol{\gamma}} f$ and $\nabla_{\boldsymbol{\psi}} f$ the partial gradients relative to the β_k , the γ_p and the ψ_l coefficients. Finally, let $\mathbf{W}$ be the diagonal matrix with the weight vector $\mathbf{w}$ on the diagonal and $\tilde{\mathbf{T}}$ the matrix obtained by adding a column of 1's on the left of $\mathbf{T}$ (i.e., with a slight abuse of notation, we introduce $\tilde{t}_{ip}$ such that, for all $1 \leq i \leq n$, $\tilde{t}_{i0} = 1$ and $\tilde{t}_{ip} = t_{ip}$ for $1 \leq p \leq P$). Then*

$$\nabla_{\boldsymbol{\alpha}} f = \frac{1}{n}\tilde{\mathbf{T}}'(\hat{\mathbf{M}}(\boldsymbol{\alpha}, \boldsymbol{\xi}) - \mathbf{M})\mathbf{W}$$

$$
\begin{aligned}
\nabla_{\boldsymbol{\xi}} f &= \frac{1}{n}\mathbf{X}'(\hat{\mathbf{M}}(\boldsymbol{\alpha},\boldsymbol{\xi}) - \mathbf{M})\mathbf{W} \\
\nabla_{\boldsymbol{\beta}} f &= \frac{w_Y}{n}\mathbf{M}'(\hat{\mathbf{y}}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\psi}) - \mathbf{y}) \\
\nabla_{\boldsymbol{\gamma}} f &= \frac{w_Y}{n}\tilde{\mathbf{T}}'(\hat{\mathbf{y}}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\psi}) - \mathbf{y}) \\
\nabla_{\boldsymbol{\psi}} f &= \frac{w_Y}{n}\mathbf{X}'(\hat{\mathbf{y}}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\psi}) - \mathbf{y}).
\end{aligned}
$$

Proof of Theorem 1. $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$ play symmetric roles in the mediator models, whether the Gaussian or logistic model is chosen. It is therefore sufficient to prove the equalities for $\boldsymbol{\alpha}$ and the same result holds for $\boldsymbol{\xi}$ by changing $\tilde{\mathbf{T}}$ into $\mathbf{X}$. The same holds for $\boldsymbol{\gamma}$ on one hand and $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ on the other hand, by changing $\tilde{\mathbf{T}}$ into $\mathbf{M}$ and $\mathbf{X}$ respectively. Only the proofs for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are therefore fully developed. Their adaptation to $\boldsymbol{\beta}$, $\boldsymbol{\xi}$ and $\boldsymbol{\psi}$ are straightforward.

Consider k such that M_k is gaussian. Then, for every $0 \leq p \leq P$,

$$
\begin{aligned}
\frac{\partial f}{\partial \alpha_{pk}} &= \frac{1}{n}\frac{\partial w_k \ell_{M_k}}{\partial \alpha_{pk}} \\
&= \frac{1}{2n} w_k \frac{\partial}{\partial \alpha_{pk}} \left(\sum_{i=1}^{n} \left(\sum_{q=0}^{P} \alpha_{qk}\tilde{t}_{iq} + \sum_{l=1}^{L} \xi_{lk} x_{il} - m_{ik} \right)^2 \right) \\
&= \frac{1}{n} w_k \sum_{i=1}^{n} \tilde{t}_{ip} \left(\sum_{q=0}^{P} \alpha_{qk}\tilde{t}_{iq} + \sum_{l=1}^{L} \xi_{lk} x_{il} - m_{ik} \right) \\
&= \frac{1}{n} w_k \sum_{i=1}^{n} \tilde{t}_{ip} (\hat{m}_{ik} - m_{ik}) \\
&= \frac{1}{n} \big(\tilde{\mathbf{T}}'(\hat{\mathbf{M}} - \mathbf{M})\mathbf{W}\big)_{pk}.
\end{aligned}
$$

The same reasoning applies when k is such that M_k is binary:

$$
\begin{aligned}
\frac{\partial f}{\partial \alpha_{pk}} &= \frac{1}{n}\frac{\partial w_k \ell_{M_k}}{\partial \alpha_{pk}} \\
&= \frac{1}{n} w_k \frac{\partial}{\partial \alpha_{pk}} \left(\sum_{i=1}^{n} -m_{ik} \left(\sum_{q=0}^{P} \alpha_{qk}\tilde{t}_{iq} + \sum_{l=1}^{L} \xi_{lk} x_{il} \right) + \log\left(1 + e^{\sum_{q=0}^{P} \alpha_{qk}\tilde{t}_{iq} + \sum_{l=1}^{L} \xi_{lk} x_{il}}\right) \right) \\
&= \frac{1}{n} w_k \sum_{i=1}^{n} \left(-\tilde{t}_{ip} m_{ik} + \frac{\tilde{t}_{ip} e^{\sum_{q=0}^{P} \alpha_{qk}\tilde{t}_{iq} + \sum_{l=1}^{L} \xi_{lk} x_{il}}}{1 + e^{\sum_{q=0}^{P} \alpha_{qk}\tilde{t}_{iq} + \sum_{l=1}^{L} \xi_{lk} x_{il}}} \right) \\
&= \frac{1}{n} w_k \sum_{i=1}^{n} \tilde{t}_{ip} (\hat{m}_{ik} - m_{ik}) \\
&= \frac{1}{n} \big(\tilde{\mathbf{T}}'(\hat{\mathbf{M}} - \mathbf{M})\mathbf{W}\big)_{pk}.
\end{aligned}
$$

The claim concerning $\nabla_{\boldsymbol{\alpha}} f$ is therefore true.

Let us now consider Y to be Gaussian. For every $0 \leq p \leq P$,

$$\begin{aligned}
\frac{\partial f}{\partial \gamma_p} &= \frac{w_Y}{n} \frac{\partial \ell_Y}{\partial \gamma_p} \\
&= \frac{w_Y}{2n} \frac{\partial}{\partial \gamma_p} \left(\sum_{i=1}^{n} \left(\sum_{q=0}^{P} \gamma_q \tilde{t}_{iq} + \sum_{k=1}^{K} \beta_k m_{ik} + \sum_{l=1}^{L} \psi_l x_{il} - y_i \right)^2 \right) \\
&= \frac{w_Y}{n} \sum_{i=1}^{n} \tilde{t}_{ip} \left(\sum_{q=0}^{P} \gamma_q \tilde{t}_{iq} + \sum_{k=1}^{K} \beta_k m_{ik} + \sum_{l=1}^{L} \psi_l x_{il} - y_i \right) \\
&= \frac{w_Y}{n} \sum_{i=1}^{n} \tilde{t}_{ip} (\hat{y}_i - y_i) \\
&= \frac{w_Y}{n} \left(\tilde{\mathbf{T}}' (\hat{\mathbf{y}} - \mathbf{y}) \right)_k .
\end{aligned}$$

In the case of a binary outcome,

$$\begin{aligned}
\frac{\partial f}{\partial \gamma_p} &= \frac{w_Y}{n} \frac{\partial \ell_Y}{\partial \gamma_p} \\
&= \frac{w_Y}{n} \left(\frac{\partial}{\partial \gamma_p} \sum_{i=1}^{n} -y_i \left(\sum_{q=0}^{P} \gamma_q \tilde{t}_{iq} + \sum_{l=1}^{K} \beta_l m_{il} + \sum_{l=1}^{L} \psi_l x_{il} \right) + \right. \\
&\qquad \left. + \log \left(1 + e^{\sum_{q=0}^{P} \gamma_q \tilde{t}_{iq} + \sum_{l=1}^{K} \beta_l m_{il} + \sum_{l=1}^{L} \psi_l x_{il}} \right) \right) \\
&= \frac{w_Y}{n} \sum_{i=1}^{n} \left(-y_i \tilde{t}_{ip} + \frac{\tilde{t}_{ip} e^{\sum_{q=0}^{P} \gamma_q \tilde{t}_{iq} + \sum_{l=1}^{K} \beta_l m_{il} + \sum_{l=1}^{L} \psi_l x_{il}}}{1 + e^{\sum_{q=0}^{P} \gamma_q \tilde{t}_{iq} + \sum_{l=1}^{K} \beta_l m_{il} + \sum_{l=1}^{L} \psi_l x_{il}}} \right) \\
&= \frac{w_Y}{n} \sum_{i=1}^{n} \tilde{t}_{ip} (\hat{y}_i - y_i) \\
&= \frac{w_Y}{n} \left(\tilde{\mathbf{T}}' (\hat{\mathbf{y}} - \mathbf{y}) \right)_p .
\end{aligned}$$

The claims on $\nabla_{\boldsymbol{\gamma}} f$ are therefore true in both cases. □

Appendix B Additional simulation results

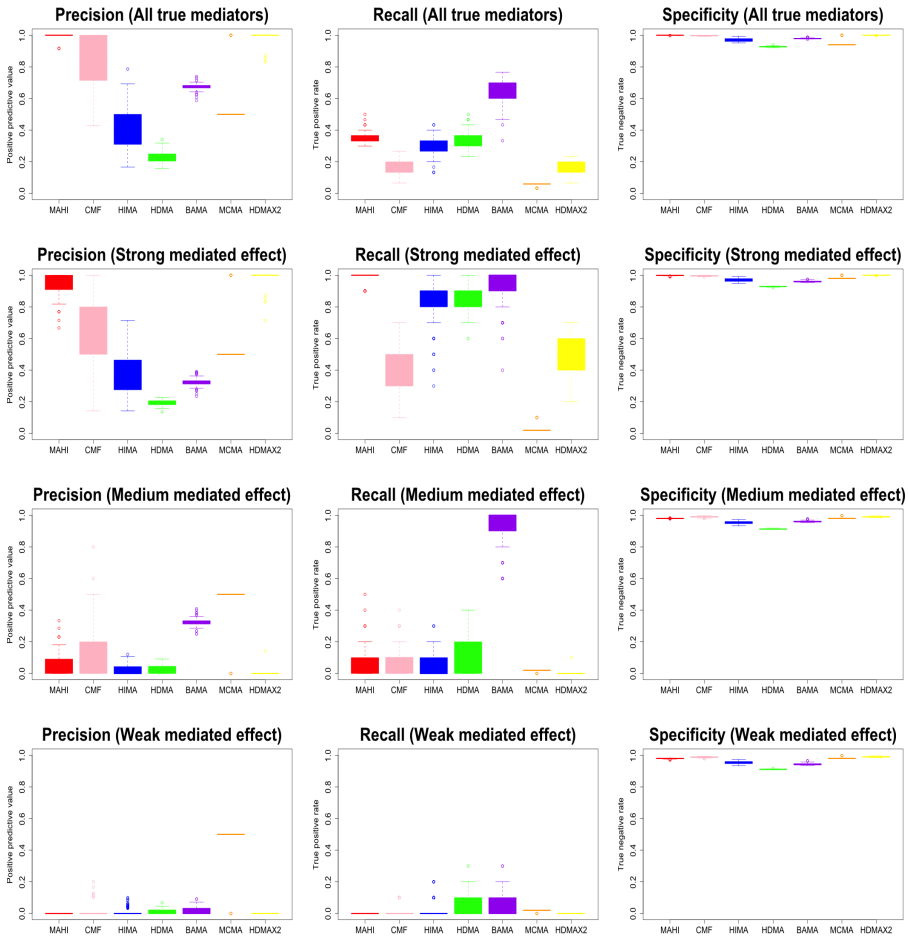

Fig. B1: Comparison of high-dimensional mediation analysis methods with regards to the ability to select the true mediators $M_1, \ldots, M_{30}$. The results are displayed in the form of boxplots showing the distribution over 100 replicates simulated with model (2) for continuous outcomes. Variables $M_1, \ldots, M_{10}$ are *strong* mediators, $M_{11}, \ldots, M_{20}$ *mild* mediators with *medium* indirect effects, and $M_{21}, \ldots, M_{30}$ *weak* mediators.

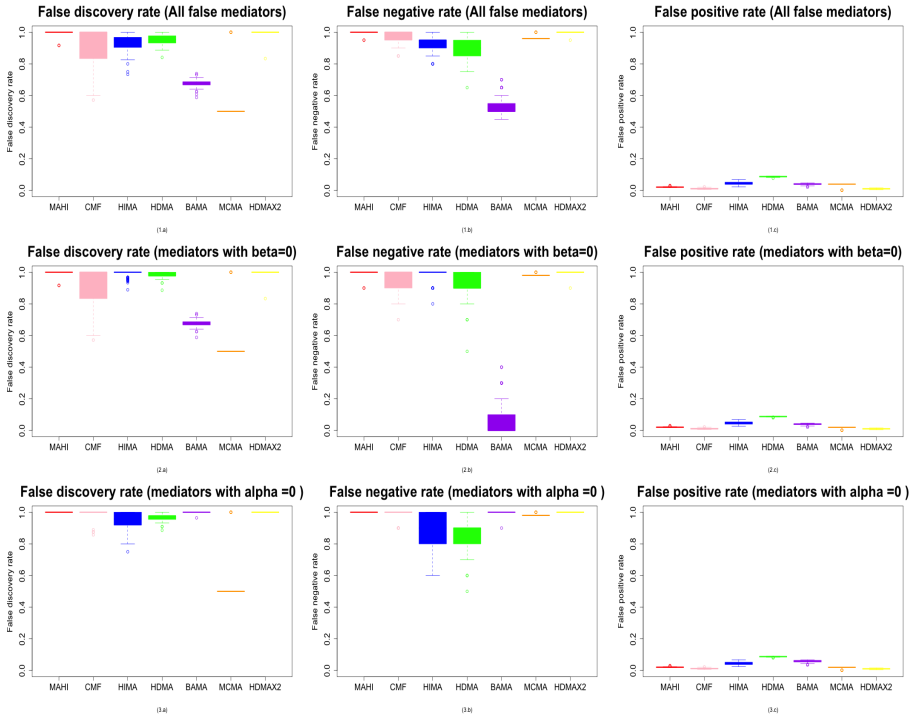

Fig. B2: Comparison of high-dimensional mediation analysis methods with regards to the selection of *false* mediators (variables $M_{31}, \ldots, M_{50}$). The results are displayed in the form of boxplots showing the distribution over 100 replicates simulated with model (2) for continuous outcomes.

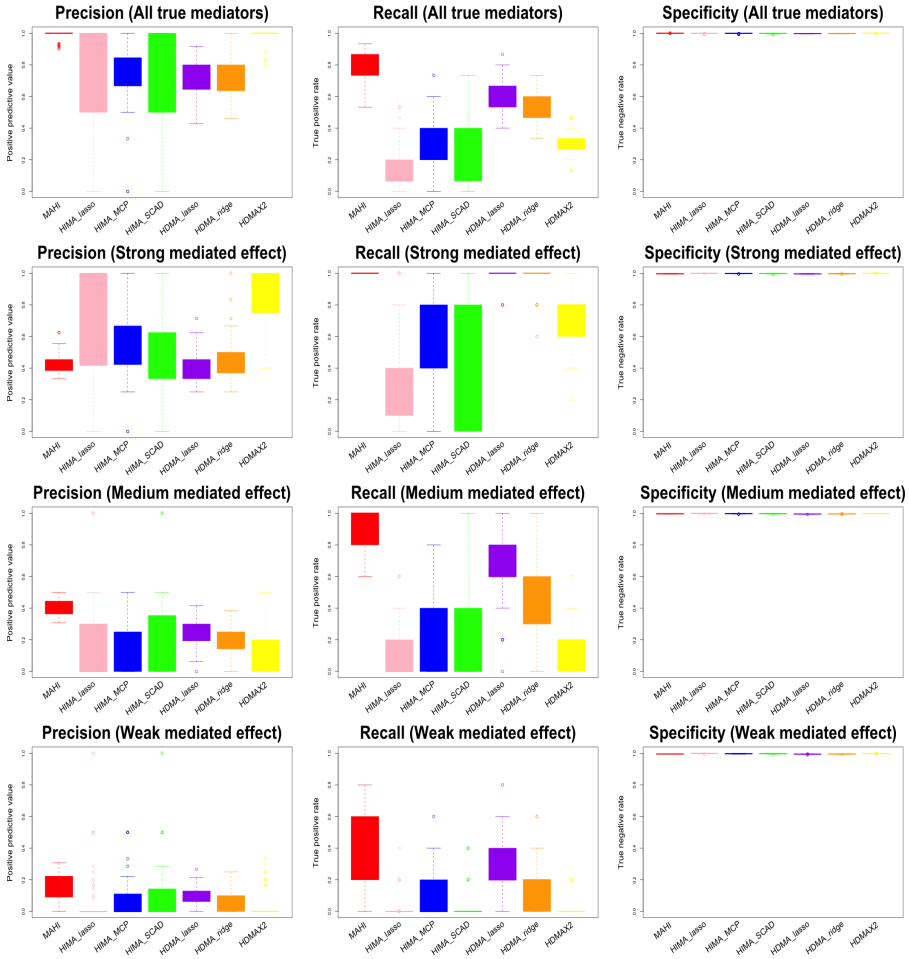

Fig. B3: Comparison of high-dimensional mediation analysis methods with regards to the ability to select true mediators (variables $M_1, \ldots, M_{15}$). The results are displayed in the form of boxplots showing the distribution over 100 replicates with binary outcomes simulated with model (3).

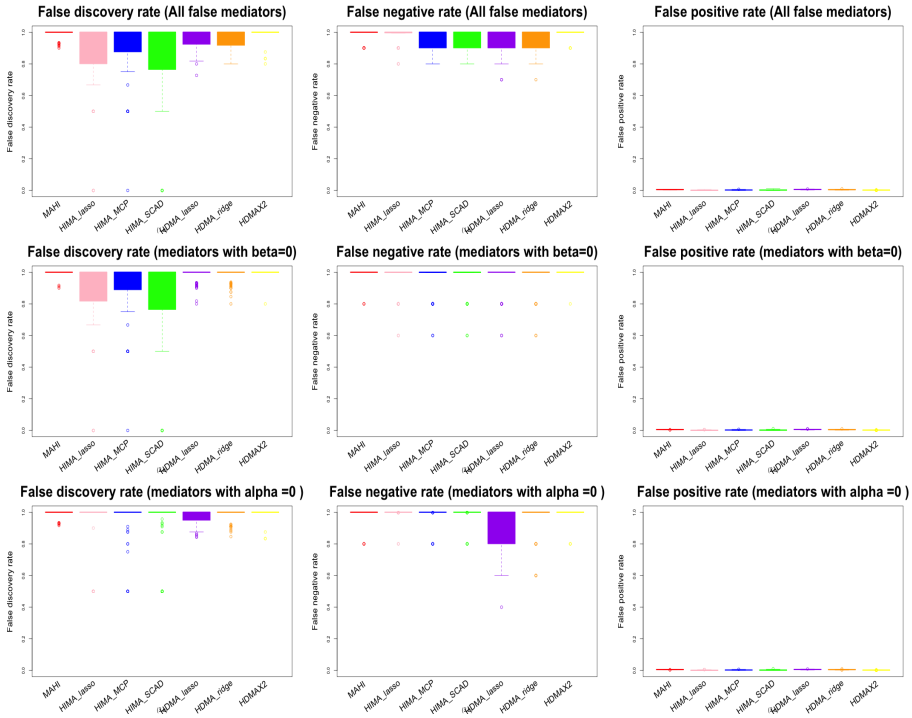

Fig. B4: Comparison of high-dimensional mediation analysis methods with regards to the selection of **false mediators** $M_{16}, \ldots, M_{25}$. The results are displayed in the form of boxplots showing the distribution over 100 replicates with binary outcomes simulated with model (3).