Published in final edited form as:

Biometrics. 2019 September ; 75(3): 745–756. doi:10.1111/biom.13053.

High dimensional mediation analysis with latent variables

Andriy Derkach[1], **Ruth M. Pfeiffer**[1], **Ting-Huei Chen**[2], **Joshua N. Sampson**[1]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, Maryland

[2]Department of Mathematics and Statistics, Laval University, Quebec City, Canada

Abstract

We propose a model for high dimensional mediation analysis that includes latent variables. We describe our model in the context of an epidemiologic study for incident breast cancer with one exposure and a large number of biomarkers (i.e., potential mediators). We assume that the exposure directly influences a group of latent, or unmeasured, factors which are associated with both the outcome and a subset of the biomarkers. The biomarkers associated with the latent factors linking the exposure to the outcome are considered “mediators.” We derive the likelihood for this model and develop an expectation-maximization algorithm to maximize an L1-penalized version of this likelihood to limit the number of factors and associated biomarkers. We show that the resulting estimates are consistent and that the estimates of the nonzero parameters have an asymptotically normal distribution. In simulations, procedures based on this new model can have significantly higher power for detecting the mediating biomarkers compared with the simpler approaches. We apply our method to a study that evaluates the relationship between body mass index, 481 metabolic measurements, and estrogen-receptor positive breast cancer.

Keywords

direct effect; factor analysis; mediation analysis; oracle property; penalized likelihood

1 | INTRODUCTION

In epidemiology, a mediation model aims to explain how an exposure (E) is associated with an outcome (Y). Traditionally, the model proposes that the exposure influences a single mediating variable (M), which, influences the outcome. More advanced models propose that the exposure influences a small set of mediating variables ($\boldsymbol{M} = (M_1,\ldots,M_p)$), which, in turn influence the outcome (VanderWeele and Vansteelandt, 2014; Assi et al., 2015; Steen et al., 2017). Here, we consider the scenario where the set of (putative) mediators is large and a study aims to identify the true set of mediators and to describe the underlying mediation model. Our motivation is an estrogen-receptor positive (ER+) breast cancer case-control

Correspondence Andriy Derkach, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD 20850. andriy.derkach@nih.gov.

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of the article.

study that measured body mass index (BMI) and 481 serum metabolites. The objective is to identify those metabolites that mediate the well-established relationship between high BMI and an increased risk of ER+ breast cancer (Moore et al., 2018).

In most discussions of mediation, the presumption is that the exposure directly influences a subset of conditionally independent mediators (Figure 1A), with more recent discussions (Daniel et al., 2015) allowing the mediators to be causally ordered (Figure 2A). Here, following our beliefs about the underlying biology, we presume that the exposure directly influences a group of conditionally independent latent, or unmeasured, factors ($\boldsymbol{F} = (F_1,...,F_q)$) which, in turn, influence both a subset of “mediating” biomarkers and the outcome (Figure 1B). In our motivating breast cancer study, for example, we might expect BMI to reduce the level of the sex hormone-binding globulin (SHBG) protein (Calle and Kaaks, 2004), which increases the availability of many of the measured hormones listed in Table 1 and the unmeasured, carcinogenic, hormones (i.e., estrogens) that cause breast cancer. In this example, the factor is a well-defined but unmeasured protein level. A more heuristic example might be evaluating the relationship between poverty, metabolites, and cancer, where poverty influences a number of distinct factors (e.g., consumption of specific foods, hours of sleep, proximity to sources of pollution, etc) that are each known to affect the levels of multiple metabolites and the risk of cancer.

Our goal is to formalize the latent variable model for mediation depicted in Figure 1B. We specify an L1-penalized version of the corresponding likelihood, propose an extension of the expectation-maximization (EM) algorithm used for sparse factor analysis (Hirose and Yamamoto, 2014; Srivastava et al., 2017) to obtain the maximum-likelihood estimates, and show that these estimates have the “oracle” property (Zou, 2006). We develop our estimation procedure for data from cohorts and retrospectively collected case-control studies. Furthermore, we show that accounting for latent variables, when the proposed model holds, can significantly increase a study’s power to detect “mediating” biomarkers (i.e., those biomarkers influenced by the mediating factors). Importantly, we note that our model is a simplification. The graph describing the true relationship between the exposure, biomarkers, and outcome is likely more complicated, where in addition to unmeasured confounders, the mediating factors can be causally ordered (Figure 2B), the graph can include bidirectional edges and cycles (Figure 2C), and the graph can include edges directly connecting the biomarkers (Figure 2D).

We note that the models in Figures 1B, 2B, and 2C are not distinguishable without imposing additional restrictions (Bai and Li, 2012) on the latent variables (e.g., conditional independence). Therefore, our independent factors are constructs of the model and we caution against the interpretation of the indirect effects through any specific factors. Nevertheless, with more modest assumptions, we can estimate and interpret the total indirect effect through all factors.

These methods extend procedures that handle latent factors affecting a small set of biomarkers (Muthén and Asparouhov, 2015; Albert et al., 2016) and add to the literature exploring high dimensional mediators. Zhang et al. (2016) model how epigenetic changes mediate the relationship between smoking and reduced lung function, assuming smoking directly affects methylation levels (i.e., Figure 1A). Huang and Pan (2016) test whether the expression levels of specific sets of genes mediate the relationship between miR-223 and glioblastoma by rotating the biomarkers and testing the resulting conditionally independent components. Chen et al. (2018) identify brain regions, from functional magnetic resonance imaging (fMRI) images, that link thermal response and self-reported pain by identifying orthogonal linear combinations of the biomarkers (i.e., fMRI voxels), known as "directions of mediation." We contrast these methods with our own approach further in Section 6.

The remainder of the paper is organized as follows. In Section 2, we describe the statistical model and the proposed EM algorithm. In Section 3, we state the theoretical properties of the resulting estimates. In Section 4, we describe two alternative approaches for identifying mediating biomarkers and estimating relevant parameters. In Section 5, we study the properties of the estimates obtained using the different approaches and apply our method to the motivating study of breast cancer. Section 6 concludes with a brief discussion.

2 | LATENT-VARIABLE MEDIATION ANALYSIS

2.1 | Overview

Our first goal is to propose a mediation model, where mediators are latent variables (Figure 1B). Our second goal is to provide a procedure to estimate the parameters in this model.

2.2 | Mediation model

We index subjects in the study by i, $i = 1,...,N$, and assume that the relationship between the exposure (E_i), factors ($\boldsymbol{F}_i$), biomarkers ($\boldsymbol{M}_i$), and outcome (Y_i) can be described by the directed acyclic graph in Figure 1B. Moreover, although not pictured, we allow for a set of baseline covariates ($\boldsymbol{X}_i$) that can influence E_i, $\boldsymbol{F}_i$, and Y_i. We then define $\boldsymbol{F}_i(e) = \left(F_{i1}(e), \ldots, F_{iq}(e)\right)'$, where $F_{ij}(e)$ is the value of the jth factor in subject i if E_i is set to e and $Y_i(e, \boldsymbol{F}_i(e'))$ is the value of Y_i if E_i is set to e and the vector of factors $\boldsymbol{F}_i$ is set to $\boldsymbol{F}_i(e')$. We further assume that sequential ignorability (Imai et al., 2010) holds, or more specifically, that

$$\{Y_i(e, \boldsymbol{f}), \boldsymbol{F}_i(e')\} \perp\!\!\!\perp E_i \mid \boldsymbol{X}_i = \boldsymbol{x} \tag{1}$$

$$Y_i(e, \boldsymbol{f}) \perp\!\!\!\perp \boldsymbol{F}_i(e') \mid E_i = e', \ \ \boldsymbol{X}_i = \boldsymbol{x}\,. \tag{2}$$

We note that, in contrast to standard models, our mediators are latent variables (Albert et al., 2016). These latent variables are unlikely to individually match up with the underlying, conditionally independent biologic variables (e.g., F_1 is the unmeasured level of SHBG in the motivating example and is conditionally independent of all other mediating factors). In reality, it is more likely that there is a set ($\boldsymbol{B} = (B_{i1},...,B_{iq})$) of interrelated biologic mediators (e.g., levels of SBHG, insulin, and cytokines) and the independent factors represent weighted combinations of these biological quantities (e.g., $F_j = \sum_{j'} w_{j'} B_{ij'}$). This truth suggests that we should focus and interpret the combined (i.e., through all factors) indirect effect defined in Equation (5), as opposed to factor or path specific indirect effects.

We can now partition the total effect (TE) of changing the exposure from e to e' into the natural direct effect (NDE) and natural indirect effect (NIE): TE = NDE + NIE, where the indirect effect passes through those pathways captured by the latent factors:

$$\text{TE} = E[Y_i\{e', \boldsymbol{F}(e')\} - Y_i\{e, \boldsymbol{F}(e)\}], \tag{3}$$

$$\text{NDE} = E[Y_i\{e', \boldsymbol{F}(e)\} - Y_i\{e, \boldsymbol{F}(e)\}], \tag{4}$$

$$\text{NIE} = E[Y_i\{e', \boldsymbol{F}(e')\} - Y_i\{e', \boldsymbol{F}(e)\}]\,. \tag{5}$$

For binary Y_i, we focus on the mediation effects defined on the odds ratio scale (VanderWeele and Vansteelandt, 2014), where $\text{OR}_{\text{TE}} = \text{OR}_{\text{NDE}} \times \text{OR}_{\text{NIE}}$,

$$\text{OR}_{\text{TE}} = \frac{P[Y_i\{e', \boldsymbol{F}(e')\} = 1]}{1 - P[Y_i\{e', \boldsymbol{F}(e')\} = 1]} / \frac{P[Y_i\{e, \boldsymbol{F}(e)\} = 1]}{1 - P[Y_i\{e, \boldsymbol{F}(e)\} = 1]},$$

$$\text{OR}_{\text{NDE}} = \frac{P[Y_i\{e', \boldsymbol{F}(e)\} = 1]}{1 - P[Y_i\{e', \boldsymbol{F}(e)\} = 1]} / \frac{P[Y_i\{e, \boldsymbol{F}(e)\} = 1]}{1 - P[Y_i\{e, \boldsymbol{F}(e)\} = 1]},$$

and

$$\text{OR}_{\text{NIE}} = \frac{P[Y_i\{e', \boldsymbol{F}(e')\} = 1]}{1 - P[Y_i\{e', \boldsymbol{F}(e')\} = 1]} / \frac{P[Y_i\{e', \boldsymbol{F}(e)\} = 1]}{1 - P[Y_i\{e', \boldsymbol{F}(e)\} = 1]}\,.$$

We note that it has already been demonstrated (Albert et al., 2016) that these effects are not generally (e.g., nonparametrically) identifiable when the mediators are factors. However, these effects are identifiable under the parametric model represented in Figure 1B and discussed in the next section. Moreover, these effects are identifiable even without assumptions (e.g., independence) about the factors as discussed in Web Appendix A. Finally, path-specific or factor-specific indirect effects are not well-defined because the models of Figures 1b, 2a, 1b and 2c are not distinguishable.

2.3 | Parametric assumptions

Recall our notation. For subject i, $i = 1, \ldots, N$, let E_i be the exposure, Y_i be the outcome, and $\boldsymbol{M}_i = (M_{i1}, \ldots, M_{ip})'$ be a vector of biomarkers with $p >> N$, and $\boldsymbol{F}_i = (F_{i1}, \ldots, F_{iq})'$ be a vector of q latent mediators with $q << p$. We assume that the distribution of Y_i belongs to an exponential family,

$$f(Y_i; \zeta_i, \psi_Y) = \exp[\{Y_i\zeta_i - b(\zeta_i)\}/a(\psi_Y) + c(Y_i, \psi_Y)] \tag{6}$$

with

$$\zeta_i = \gamma_Y + \beta_{EY} E_i + \boldsymbol{\beta}'_{FY} \boldsymbol{F}_i\,. \tag{7}$$

We also assume that $\boldsymbol{F}_i$ and $\boldsymbol{M}_i$ are normally distributed:

$$\boldsymbol{F}_i = \boldsymbol{\beta}_{EF} E_i + \boldsymbol{e}_{f,i} \quad \text{with} \;\; \boldsymbol{e}_{f,i} \sim N\left(0, I_q\right), \tag{8}$$

where I_q is the q by q identity matrix and

$$\begin{aligned} &\boldsymbol{M}_i = \boldsymbol{\gamma}_M + \Lambda \boldsymbol{F}_i + \boldsymbol{e}_{m,i} \quad \text{with} \quad \boldsymbol{e}_{m,i} \sim N\left(0, \Psi^2\right), \\ &\Psi^2 = diag\left(\psi_1^2, \ldots, \psi_p^2\right). \end{aligned} \tag{9}$$

Under retrospective sampling (i.e., for case-control data), we rely on the additional assumption that $E \sim N\left(\gamma_E, \sigma_E^2\right)$. We note that $\boldsymbol{\beta}_{\mathrm{EF}} = \left(\beta_{\mathrm{EF},1}, \ldots, \beta_{\mathrm{EF},q}\right)'$ and $\boldsymbol{\beta}_{\mathrm{FY}} = \left(\beta_{FY,1}, \ldots, \beta_{FY,q}\right)'$ are vectors of length q, Λ is a $p \times q$ matrix and the (m, j)th element, denoted by λ_{mj}, represents the effect of the jth factor on the mth biomarker. Moreover, we define "mediating" biomarkers to be the set $\{m: \sum_j \left|\beta_{\mathrm{EF},j} \beta_{\mathrm{FY},j} \lambda_{mj}\right| \neq 0\}$ and therefore, when trying to identify the mediators, select the set $\left\{m: \sum_j \left|\hat{\beta}_{\mathrm{EF},j} \hat{\beta}_{\mathrm{FY},j} \hat{\lambda}_{mj}\right| \neq 0\right\}$.

The proposed setup accommodates outcomes from a variety of distributions, but for ease of exposition we focus on outcomes from either a binomial or a normal distribution. For a continuous Y we assume $\psi_Y = \sigma_Y^2$, $b(\zeta_i) = \zeta_i^2/2$ and $c(Y_i, \psi_Y) = -1/2\left\{Y_i/\sigma_Y^2 + \log\left(2\pi\sigma_Y^2\right)\right\}$ Equation (6) simplifies to

$$Y_i = \gamma_Y + \beta_{EY} E_i + \boldsymbol{\beta}'_{FY} \boldsymbol{F}_i + e_{y,i} \quad \text{with} \quad e_{y,i} \sim N\left(0, \sigma_y^2\right). \tag{10}$$

In this scenario, the pathway specific effects can be related to model parameters by

$$\mathrm{TE} = \left(\boldsymbol{\beta}'_{\mathrm{EF}} \boldsymbol{\beta}_{\mathrm{FY}} + \beta_{\mathrm{EY}}\right)(e' - e),$$

$$\mathrm{NDE} = \beta_{\mathrm{EY}}(e' - e),$$

$$NIE = \boldsymbol{\beta}'_{\mathrm{EF}} \boldsymbol{\beta}_{\mathrm{FY}}(e - e')$$

and we note that these effects are identifiable. Similar conclusions were drawn by Albert et al. (2016) in the case of a single latent mediator. For a binary Y we assume $\psi_Y = 1$ and $b(\zeta_i) = \log\{1 + \exp(\zeta_i)\}$ and rewrite Equation (6) in logistic form, $P(Y_i = 1; \zeta_i, \psi_Y) = \exp(\zeta_i)/\{1 + \exp(\zeta_i)\}$. When the outcome is rare, we can approximate the pathway-specific effects on the OR scale by $\mathrm{OR}_{\mathrm{NDE}} \approx \exp\{\beta_{\mathrm{EY}}(e' - e)\}$ and $\mathrm{OR}_{\mathrm{NIE}} \approx \exp\{\boldsymbol{\beta}'_{\mathrm{EF}} \boldsymbol{\beta}_{\mathrm{FY}}(e' - e)\}$, with modifications available to accommodate matched case-control studies (VanderWeele and Tchetgen Tchetgen, 2016) or interactions between the factors and exposure (VanderWeele and Vansteelandt, 2014).

In Web Appendix E, we consider extensions of Models (6–9) to accommodate additional covariates and extend the estimation procedure and theory presented to this setting. To estimate the vector of parameters $\boldsymbol{\theta} = \left(\gamma_Y, \boldsymbol{\beta}_{FY}, \beta_{EY}, \psi_Y, \boldsymbol{\beta}_{EF}, \boldsymbol{\gamma}_M, \Lambda, \Psi^2\right)'$ using the observed data, $(Y_i, \boldsymbol{M}_i, E_i)$ for $i = 1, \ldots, N$, we first assume q, the number of latent factors, is known. In Section 2.5, we discuss how to choose q in practice.

2.4 | Likelihood

Under prospective sampling, we derive the joint likelihood of $(Y_i, \boldsymbol{M}_i, \boldsymbol{F}_i)$ and $(Y_i, \boldsymbol{M}_i)$ while conditioning on E_i to avoid modeling the distribution of the exposure. Under retrospective sampling (i.e., case-control data), we derive the joint likelihood of $(Y_i, E_i, \boldsymbol{M}_i, \boldsymbol{F}_i)$ and $(Y_i, E_i, \boldsymbol{M}_i)$.

2.4.1 | Prospective likelihood—Here the full data likelihood for $(Y, \boldsymbol{M}, \boldsymbol{F})$ is

$$L_P^F(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(Y_i, \boldsymbol{M}_i, \boldsymbol{F}_i \mid E_i; \boldsymbol{\theta}), \tag{11}$$

where f is the product of the densities defined by Equations (6–9),

$$\begin{aligned} f(Y_i, \boldsymbol{M}_i, \boldsymbol{F}_i \mid E_i; \boldsymbol{\theta}) = {} & f_Y(Y_i \mid \boldsymbol{F}_i, E_i; \gamma_Y, \boldsymbol{\beta}_{\mathrm{FY}}, \beta_{\mathrm{EY}}, \psi_Y) \\ & \times f_M\left(\boldsymbol{M}_i \mid \boldsymbol{F}_i; \boldsymbol{\gamma}_M, \Lambda, \Psi^2\right) f_F(\boldsymbol{F}_i \mid E_i; \boldsymbol{\beta}_{\mathrm{EF}}) . \end{aligned} \tag{12}$$

We use f_M, f_Y, and f_F to denote implied distribution of $\boldsymbol{M}$, Y, and $\boldsymbol{F}$, respectively. However, the factors $\boldsymbol{F}_i$ are not observed. The likelihood for the observed data, $(Y_i, \boldsymbol{M}_i)$, is therefore

$$L_P^O(\boldsymbol{\theta}) = \prod_{i=1}^{N} \int_{\boldsymbol{F}} f(Y_i, \boldsymbol{M}_i, \boldsymbol{F}_i \mid E_i; \boldsymbol{\theta}) d\boldsymbol{F} . \tag{13}$$

Although $L_P^O(\boldsymbol{\theta})$ does not have a closed form in general, we show in the Web Appendix B that $L_P^O(\boldsymbol{\theta})$ is the product of normal distributions when Y_i is normally distributed.

2.4.2 | Retrospective likelihood—Under retrospective sampling, N_1 cases and N_0 controls are drawn from the population of cases and controls, respectively, and biomarkers and exposures are observed ($N_1 + N_0 = N$). The corresponding likelihood is

$$\begin{aligned} L_R^F(\boldsymbol{\theta}) = {} & \prod_{t\ \in\ \text{case}} f(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i Y_i = 1; \boldsymbol{\theta}) \\ & \times \prod_{t\ \in\ \text{control}} f(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i \mid Y_i = 0; \boldsymbol{\theta}) . \end{aligned} \tag{14}$$

Here, we assume $E_i \sim N\left(\gamma_E, \sigma_E^2\right)$ in the overall population. Although closed forms of the conditional distributions of (E_i, $\boldsymbol{M}_i$, $\boldsymbol{F}_i$) are generally not available, they can be approximated well when the outcome is rare in the general population, that is,

$$\begin{aligned} P(Y_i = 1 \mid E_i, \boldsymbol{F}_i; \gamma_Y, \beta_{EY}, \boldsymbol{\beta}_{\text{FY}}) \\ = \frac{\exp(\gamma_Y + \beta_{EY} E_i + \boldsymbol{\beta}_{\text{FY}}' \boldsymbol{F}_i)}{1 + \exp(\gamma_Y + \beta_{EY} E_i + \boldsymbol{\beta}_{\text{FY}}' \boldsymbol{F}_i)} \\ \approx \exp(\gamma_Y + \beta_{EY} E_i + \boldsymbol{\beta}_{\text{FY}}' \boldsymbol{F}_i) \end{aligned} \qquad (15)$$

Under the rare disease assumption, the distribution of (E_i, $\boldsymbol{M}_i$, $\boldsymbol{F}_i$) in controls is approximately equal to the distribution in the general population. Thus, under Models (6–9)

$$\begin{aligned} f(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i \mid Y_i = 0; \boldsymbol{\theta}) \approx f(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i; \boldsymbol{\theta}) \\ = \phi\left(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i; \boldsymbol{\mu}_0, \Sigma_{E,M,F}\right), \end{aligned} \qquad (16)$$

where $\phi\left(.; \boldsymbol{\mu}_0, \Sigma_{E,M,F}\right)$ is a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma_{E,M,F}$. The covariance matrix $\Sigma_{E,M,F}$ is defined in the Web Appendix B.2 and $\boldsymbol{\mu}_0 = (\mu_E, \boldsymbol{\mu}_M, \boldsymbol{\mu}_F)'$. Note that, $\Sigma_{E,M,F}$ is a function of the parameters used in Models (6–9). The distribution of (E_i, $\boldsymbol{M}_i$, $\boldsymbol{F}_i$) in cases under Model (15) is

$$f(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i \mid Y_i = 1; \boldsymbol{\theta}) \approx \phi\left(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i; \boldsymbol{\mu}_1, \Sigma_{E,M,F}\right) \qquad (17)$$

where

$$\begin{aligned} \mu_1 = \left(\mu_E^1, \boldsymbol{\mu}_M^1, \boldsymbol{\mu}_F^1\right)' = (\mu_E, \boldsymbol{\mu}_M, \boldsymbol{\mu}_F)' \\ + \Sigma_{E,M,F}(\beta_{EY}, \mathbf{0}', \boldsymbol{\beta}_{FY}')' \end{aligned}$$

Note that the distributions of (E_i, $\boldsymbol{M}_i$, $\boldsymbol{F}_i$) in cases and controls differ only in their means.

Based on the above approximations, the likelihood for the full data (E_i, $\boldsymbol{M}_i$, $\boldsymbol{F}_i$) is

$$\begin{aligned} L_R^F(\boldsymbol{\theta}) = \prod_{t \in \text{case}} \phi\left(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i; \boldsymbol{\mu}_1, \Sigma_{E,M,F}\right) \\ \times \prod_{t \in \text{controls}} \phi\left(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i; \boldsymbol{\mu}_0, \Sigma_{E,M,F}\right). \end{aligned} \qquad (18)$$

and the likelihood for the observed data is therefore easily shown to be

$$\begin{aligned} L_R^O(\boldsymbol{\theta}) = \prod_{t \in \text{case}} \phi\left(E_i, \boldsymbol{M}_i; \boldsymbol{\mu}_{1;M,E}, \Sigma_{M,E}\right) \\ \times \prod_{t \in \text{controls}} \phi\left(E_i, \boldsymbol{M}_i; \boldsymbol{\mu}_{0;M,E}, \Sigma_{M,E}\right), \end{aligned} \qquad (19)$$

where $\mu_{1;E,M} = \left(\mu_E^1, \boldsymbol{\mu}_M^1\right)'$, $\mu_{0;E,M} = (\mu_E, \boldsymbol{\mu}_M)'$, and $\Sigma_{E,M}$ is the appropriate submatrix of $\Sigma_{E,M,F}$. The effects of the exposure and the latent factors on the outcome are thus completely captured by the difference between the means of (E_i, $\boldsymbol{M}_i$) in cases and controls.

2.5 | Penalty to induce sparsity in the factors, *F*

To introduce sparseness in the factors $\boldsymbol{F}$ and the number of biomarkers associated with those factors, we maximize the penalized log-likelihood

$$\begin{aligned} \mathrm{PLL}(\boldsymbol{\theta}) = \log\left\{L^{O}(\boldsymbol{\theta})\right\} - \rho_{1N}\sum_{j=1}^{q} P\left(\beta_{FY,j}\right) \\ -\rho_{2N}\sum_{j=1}^{q} P\left(\beta_{EF,j}\right) - \rho_{3N}\sum_{m=1}^{p}\sum_{j=1}^{q} P\left(\lambda_{mj}\right), \end{aligned} \qquad (20)$$

where $L^{O}(\boldsymbol{\theta})$ is the likelihood defined in (13) or (19) and $P(\cdot)$ is the chosen penalty function. We use the adaptive lasso penalty, where $P(\varphi) = \frac{|\varphi|}{\left|\hat{\varphi}^{0}\right|}$ and $\hat{\varphi}^{0}$ is a square root consistent estimate of φ, but other options, such as SCAD or MC+ (Fan and Li, 2001; Zhang, 2010), are possible. In practice, we let $\hat{\beta}^{0}_{\mathrm{EF},j}$, $\hat{\beta}^{0}_{\mathrm{FY},j}$ and $\hat{\lambda}^{0}_{mj}$ be initial estimates from the observed likelihood $L^{O}(\boldsymbol{\theta})$. We allow the penalties (ρ_{1N}, ρ_{2N}, ρ_{3N}) to differ for each type of association. The penalized log-likelihood estimator $\hat{\boldsymbol{\theta}}_{P}$ is defined as

$$\hat{\boldsymbol{\theta}}_{P} \equiv \hat{\boldsymbol{\theta}}_{P}\left(\rho_{1N}, \rho_{2N}, \rho_{3N}\right) = \underset{\theta}{\operatorname{argmax}}\ \mathrm{PLL}(\boldsymbol{\theta})\,. \qquad (21)$$

Because $\hat{\boldsymbol{\theta}}_{P}$ cannot be expressed in a closed from, we develop an (EM) algorithm building upon the methods for sparse factor analysis (Hirose and Yamamoto, 2014; Srivastava et al., 2017). Although the EM algorithm is a significant contribution of this paper, we provide details in Web Appendix C to keep the main text focused.

In practice, we specify the total number of factors, $q_{\max}$, to be 40 and choose values of ρ_{1N}, ρ_{2N}, and ρ_{3N} that minimize the extended Bayes information criterion (EBIC; Chen and Chen, 2008) defined as EBIC = $-2l\left\{\hat{\theta}_{P}\left(\rho_{1N}, \rho_{2N}, \rho_{3N}\right)\right\} + \log(N)df + 2\gamma\log(\tau)$, where $l(\hat{\boldsymbol{\theta}})$ is the observed log-likelihood (Equations (13) or (19)), df is the number of parameters with nonzero estimates, and $\boldsymbol{\tau}$ is the number of possible models with df nonzero parameters. We generally suggest that users start with a value of $q_{\max}$ that is likely to exceed the true number of factors influencing the biomarkers. In practice, we set $\gamma = 0.5$ in Equation (21). EBIC tends to outperform the more traditional selection criteria AIC and BIC in previous high dimensional settings (Chen and Chen, 2008; Srivastava et al., 2017) and in our simulations.

3 | THEORETICAL PROPERTIES OF THE ESTIMATES

We highlight key properties of the model, the EM algorithm, and the estimates, $\hat{\boldsymbol{\theta}}_{P}$, with proofs in the Supporting Information (see Web Appendix D). First, we show that the parameters $\boldsymbol{\theta}$ are identifiable under the following condition for factor analysis presented in Anderson and Rubin (1956),

Condition 1.

If any row of the loading matrix Λ *is deleted, there remain two disjoint submatrices of rank* q.

Proposition 1 (Identifiability).

If Condition 1 holds, then $\boldsymbol{\theta}$ *is identifiable, and* Λ, $\boldsymbol{\beta}_{\mathrm{FY}}$, $\boldsymbol{\beta}_{\mathrm{EF}}$ *are identifiable up to an orthogonal rotation*.

We note that the products of parameters $\Lambda\Lambda'$, $\|\boldsymbol{\beta}_{\mathrm{FY}}\|^2 = \boldsymbol{\beta}_{\mathrm{FY}}'\boldsymbol{\beta}_{\mathrm{FY}}$, $\|\boldsymbol{\beta}_{\mathrm{EF}}\|^2 = \boldsymbol{\beta}_{\mathrm{EF}}'\boldsymbol{\beta}_{\mathrm{EF}}$ and mixed products $\Lambda\boldsymbol{\beta}_{\mathrm{FY}}, \Lambda\boldsymbol{\beta}_{\mathrm{EF}}$ and $\beta_{\mathrm{EF}}'\boldsymbol{\beta}_{\mathrm{FY}}$ are uniquely identified. The identifiability of these terms is crucial when estimating direct and indirect effects (Section 2.2).

Our second property, building upon previous work of Hirose and Yamamoto (2014), Srivastava et al. (2017), states the properties the EM algorithm.

Proposition 2 (Convergence of the EM algorithm).

With each iteration of the proposed EM algorithm, the penalized log-likelihood (20) does not decrease,

$$\mathrm{PLL}\left(\hat{\theta}^{k}\right) \leq \mathrm{PLL}\left(\hat{\theta}^{k+1}\right), k \geq 1,$$

and the sequence of EM estimates $\hat{\boldsymbol{\theta}}^{k}$ *converges to a local maximum* $\hat{\boldsymbol{\theta}}_{P}^{*}$.

Our third property, building upon work of Zou (2006), is that the resulting estimates, $\hat{\boldsymbol{\theta}}_P = \left(\hat{\theta}_{P1}, \ldots, \hat{\theta}_{PW}\right)'$, where W is the total number of parameters, have the oracle property. Let $\boldsymbol{\theta}$ be the vector of all parameters (see Section 2.3), $A = \{j | \theta_j \neq 0\}$ index the set of parameters not equal to 0, $\hat{A}_N = \left\{j \mid \hat{\theta}_{Pj} \neq 0\right\}$ index the set of parameters with nonzero estimates; based on our dataset of N subjects and let $\hat{\boldsymbol{\theta}}_P^S = \left\{\hat{\theta}_{Pj}: j \in A\right\}$ be the vector of estimates for the nonzero parameters $\boldsymbol{\theta}^S = \left\{\theta_j: j \in A\right\}$.

Proposition 3 (Oracle Property).

Suppose that $\rho_{kN}/\sqrt{N} \rightarrow 0$ *and* $\rho_{kN} \rightarrow \infty$ *for* $k \in \{1, 2, 3\}$. *Then we obtain*

1. consistency of the selection of nonzero effects:

$$\lim_{N \to \infty} P\left(\hat{A}_N = A\right) = 1, \tag{22}$$

2. asymptotic normality for the nonzero effects:

$$\lim_{N \to \infty} \sqrt{N}\left(\hat{\boldsymbol{\theta}}_P^S - \boldsymbol{\theta}^S\right) \rightarrow_d \quad N\left(0, \mathcal{I}_{\hat{\theta}^S}^{-1}\right), \tag{23}$$

where $\mathbb{I}_{\theta^s}$ is Fisher's information matrix for the true model (i.e., excluding zero coefficients).

4 | ALTERNATIVE METHODS TO IDENTIFY SUBSETS OF MEDIATORS

Here, we propose two alternative, two-step approaches to identify the subset of "mediating" biomarkers. Recall for latent variable mediation analysis (LVMA), we identify the mediating biomarkers to be the set $\left\{m: \sum_j \left|\hat{\beta}_{EF,j}\hat{\beta}_{FY,j}\hat{\lambda}_{mj}\right| \neq 0\right\}$.

4.1 | Individual marker mediation analysis (IMA)

We first consider testing each biomarker individually using Sobel's test (Sobel, 1982). For a continuous outcome, we can test biomarker m by fitting two linear regression models

$$Y_i = \gamma_Y^* + \beta_{\mathrm{MY},m}^* M_{m,i} + \beta_{\mathrm{EY},m}^* E_i + \epsilon_Y^*$$

$$M_{m,i} = \gamma_M^* + \beta_{\mathrm{EM},m}^* E_i + \epsilon_{M,m}^*$$

and then calculating the P value, p_m, for the test statistic $Z_m = \hat{\beta}_{\mathrm{MY},m}^* \hat{\beta}_{\mathrm{EM},m}^* / \hat{\sigma}_{\beta\beta}$, where $\hat{\sigma}_{\beta\beta}^2$ is the estimated variance of the product $\hat{\beta}_{\mathrm{MY},m}^* \hat{\beta}_{\mathrm{BM},m}^*$ and is calculated using the Taylor-Series expansion (Sobel, 1982). Logistic regression would be used when Y is binary. We can then identify the set of "mediating" biomarkers as those with a P value below a specified threshold.

4.2 | Two-step mediation analysis (TMA)

We next consider a two-step approach where, in the first step, we identify the latent factors underlying the biomarkers and, in the second step, we test whether each of those latent factors are mediators. Specifically, in the original-version (TMAO), we first perform sparse factor analysis on $\boldsymbol{M}$, ignoring E and Y:

$$\underset{\Lambda, \Psi^2}{\arg\max}\left[\sum_{i=1}^{N} \log\left\{\phi\left(\boldsymbol{M}_i; \Lambda, \Psi^2\right)\right\} - \rho \sum_{m=1}^{p} \sum_{j=1}^{q} \frac{\left|\lambda_{mj}\right|}{\left|\hat{\lambda}_{mj}^0\right|}\right], \tag{24}$$

where $\phi\left(\cdot\,; \Lambda, \Psi^2\right)$ is a multivariate normal distribution with mean $\mathbf{0}$ and variance $\Lambda\Lambda' + \Psi^2$ and obtain our estimated factors $\left(\hat{F}_{i1}, \ldots, \hat{F}_{iq}\right) = \hat{\Lambda}'\left(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}^2\right)^{-1}\boldsymbol{M}_i$. Again, we choose the penalty, ρ, which minimizes the EBIC. In the second step, we use Sobel's test (Sobel, 1982), to individually test each of the estimated factors. The "mediating" biomarkers are those with a nonzero loading on the selected factors.

In a modified version (TMAR), we perform sparse factor analysis (24) on $\boldsymbol{M}_R$, where the jth row of $\boldsymbol{M}_R$ are residuals after regressing the jth row of M on E. Specifically, we substitute $\boldsymbol{M}_R$ for $\boldsymbol{M}$ in Equation (24) to obtain $\widehat{\Lambda}_R$ and $\widehat{\Psi}_R$, and then let $\left(\widehat{F}_{i1}, \ldots, \widehat{F}_{iq}\right)' = \widehat{\Lambda}'_R\left(\widehat{\Lambda}_R\widehat{\Lambda}'_R + \widehat{\Psi}^2_R\right)^{-1}\boldsymbol{M}_i$. Under prospective sampling, $\boldsymbol{M}$, Y, and E need to be centered by their corresponding sample means because intercepts of are not identifiable directly from $\boldsymbol{M}$ or $\boldsymbol{M}_R$ by these two approaches. Under retrospective sampling, $\boldsymbol{M}$ and the exposure E are centered by their means in the controls. TMAR is conceptually similar to the mediation approach proposed by Huang and Pan (2016) and to surrogate variable analysis (Leek and Storey, 2007); that it is $\boldsymbol{M}_R$, not $\boldsymbol{M}$, that is accurately described by factor analysis.

4.2.1 | Remarks—

TMAO and TMAR are computationally faster than the full joint model LVMA and, for TMAO, it is straight-forward to calculate P values for the tested factors. However, by ignoring E and Y, less information is available for identifying factors and, ultimately, detecting mediators. Second, TMAO incorrectly assumes independence of factors. TMAR, or allowing for correlated factors (Hirose and Yamamoto, 2014), attempts to handle this issue, but these methods perform poorly in small samples. Third, in case-control studies, the assumption that the biomarkers are normally distributed is violated and Section 2.4.2 shows that much of the information is contained in the difference between the group means. Fourth, by not explicitly modeling the latent variables, the terms, β_{EY} and $\beta'_{\mathrm{EF}}\beta_{\mathrm{FY}}$ used to estimate the direct and indirect effects are biased because the imputed mediators contain measurement error (Carroll et al., 2006; le Cessie et al., 2012; Valeri et al., 2014).

5 | SIMULATIONS

5.1 | Data generation

We compared the properties of LVMA, IMA, and TMA (O and R) in simulated data with $N = 300$ or $N = 500$ observations. We assumed that there were fifteen factors ($q = 15$) each affecting twenty unique biomarkers, $\boldsymbol{M}$. The first factor was a mediator (i.e., $\beta_{\mathrm{EF},1} \neq 0$ and $\beta_{\mathrm{FY},1} \neq 0$), the next four factors were associated only with E (i.e., $\beta_{\mathrm{EF},j} \neq 0$ and $\beta_{\mathrm{FY},j} = 0$ for $j = 2,...,5$) and the last ten factors were associated with neither E nor Y (i.e., $\beta_{\mathrm{EF},j} = 0$ and $\beta_{\mathrm{FY},j} = 0$ for $j = 6,...,15$). We then added an additional ninety independent normally distributed biomarkers so that $p = 390$. Data were simulated based on the model in Section 2.3, with binary outcomes assuming a logistic link and continuous outcomes assuming normality. For case-control studies, we sampled an equal number of cases and controls (i.e., $N_1 = N_0 = 150$ or $N_1 = N_0 = 250$) from a larger population prospectively simulated. We varied the effect of the exposure on the mediating factor ($\beta_{\mathrm{EF},1} \in \{0.4, 0.5\}$), the effect of exposure-related factors on their constituent biomarkers ($\lambda_1 \in \{0.25, 0.3\}$), and the effect of exposure-unrelated factors on their constituent biomarkers ($\lambda_0 \in \{0.4, 0.5\}$). Here, $\lambda_{mj} = \lambda_1$ if $j \leq 5$ (when $\lambda_{mj} \neq 0$) and $\lambda_{mj} = \lambda_0$ if $6 \leq j \leq 15$ (when $\lambda_{mj} \neq 0$). Other parameters/distributions were set as follows: $\beta_{\mathrm{EY}} = \beta_{\mathrm{FY},1} = 0.3$, $\beta_{\mathrm{EF},2} = \cdots = \beta_{\mathrm{EF},5} = 0.7$, $\gamma_{M,1} = \cdots = \gamma_{M,p} = 0.5$, $\psi_1^2 = \cdots = \psi_p^2 = 1$, $\gamma_Y = 4.6$ (i.e., $P(Y = 1 | E = 0, F_1 = 0) = 0.01$ for binary outcomes), and $E \sim N(0.5, 1)$. In Web Appendix F, we describe the simulation setup and corresponding results for scenarios when there are no latent variables and the exposure directly affects individual biomarkers.

For each parameter and sample size combination, we simulated 1000 datasets. Then we applied each of the four methods and calculated the average number of true positives (TPs) selected, the average number of false positives (FPs) selected, and the average estimate of the conditional exposure effect, β_{EY}. Here, the biomarker m is a “true positive” (TP) if it is identified as a mediator and $\sum_j \left|\beta_{\mathrm{EF},\,j}\beta_{\mathrm{FY},\,j}\lambda_{mj}\right| \neq 0$. It is a “false positive” (FP) if it is identified as a mediator but $\sum_j \left|\beta_{\mathrm{EF},\,j}\beta_{\mathrm{FY},\,j}\lambda_{mj}\right| = 0$. For valid comparison between LVMA and the other methods, we selected P value thresholds for these methods so that the FP remained constant. Note, the requirement for a biomarker to be “selected” as a mediator is defined in Section 2. For IMA, TMAO, and TMAR, we estimated the conditional exposure effect as the coefficient for the exposure in a model for the outcome that also included all selected biomarkers or factors.

5.2 | Results

We summarize our results in Figures 3. First, for most settings, LMVA tended to have, the largest TP rate (panels (A) of Figures 3-5). Recall, we chose significance thresholds for the other three methods so that the average FP rate (see panels (B) of Figures 3-5) was similar across methods. TMAR and TMOR tended to perform similarly. The one exception is that TMAR performed poorly when we reduced the effect of exposure-related factors to λ_1 = 0.25 (see Figure 5). IMA consistently performed poorly (e.g., on average it had four-five times lower TP rate). The latter reinforces the need to use some form of latent variable analysis when the exposure does not directly affect biomarkers individually.

The pronounced when the effect of factors on associated biomarkers is small (Figure 5) and when the outcome was a binary variable (Figures 4 and 5). When comparing TMAO to TMAR, neither clearly performed better. Their relative performance strongly depends on sample size, with TMAR performing better as the sample sizes increased and the effect of the exposure could be better estimated. This advantage was further magnified when the effects of the factor on the biomarkers increased.

As expected, TMAR, TMAO, and IMA produced biased estimates of the conditional effect of exposure; compare the results in panel (C) of Figures 3-5 to the true direct effect, β_{EF} = 0.3. Although LVMA had smaller bias, the estimates of the conditional effect of exposure using LVMA still exceeded 0.35 in most scenarios. When we increased the sample size or the effect size so that TP = 20 (see Web Figures 5, 8, and 21), the bias in LVMA, but not for the other methods, disappeared.

LVMA was robust to the number of specified factors (see Web Figures 11 and 12). Therefore, in practice, we suggest specifying a relatively large number of factors. Furthermore, we found that using EBIC was slightly preferable to using AIC or BIC, but the two two-step methods, TMAO and TMAR, were far more sensitive to the choice of selection criteria (see Web Figures 5-10). LVMA was also robust to violations of the assumption that the error terms for the biomarkers were normally distributed (see Web Figures 13-18). Finally, when we decreased exposure effects on nonmediating factors to $\beta_{\mathrm{EF,2}} = \cdots = \beta_{\mathrm{EF,5}}$ = 0.4, the TP of LVMA, TMAO, and TMAR become similar. Web Figures 19-21 shows that, if we also decrease the effect of the mediating factors on the biomarkers to λ_1 = 0.25, LVMA regains its advantages. Note, in this latter setting the exposure becomes important for identifying factors.

5.3 | Data example

Our motivating study aims to identify metabolites that mediate the known relationship between high BMI and the increased risk of ER+ breast cancer. This study, nested inside the prostate, lung, colorectal, and ovarian cancer screening study (PLCO), includes 410 (ER+) breast cancers and 410 controls matched on study age (±2 years), date of blood collection (±3 months), and hormone therapy use at baseline. The study collected serum samples at the first follow-up visit, 1-year after baseline, and using these specimen, measured 481 known serum metabolites (<kDA). Metabolite levels were log transformed. Details on the study are in Moore et al. (2018).

We modeled the data using LVMA with $q_{\max} = 40$ factors adjusting for the three matching variables (see Web Appendix E). The model identified only a single factor associated with both BMI ($\hat{\beta}_{\mathrm{EF},\,1} = 0.035$) and risk of breast cancer ($\hat{\beta}_{\mathrm{FY},\,1} = 0.14$). This factor had 111 nonzero loadings but only 16 of these loadings had an absolute value larger than 0.4, with the majority of remaining metabolites having loadings below 0.01. In Table 1, we list these 16 metabolites. In Web Figure 22, we display loadings for all metabolites from LVMA and standard factor analysis. Of interest, many of these metabolites are products of estrogen metabolism, suggesting estrogen metabolism does partially explain why increased BMI is associated with increased risk of ER+ breast cancer. However, most of the effect of BMI was not mediated by this factor. When estimating the TE, NDE, and NIE on the OR scale, we find $\mathrm{OR}_{TE} = \exp(0.039)$, $\mathrm{OR}_{\mathrm{NDE}} = \exp(0.034)$, and $\mathrm{OR}_{\mathrm{NIE}} = \exp(0.005)$ suggesting that the estrogen pathways explains only a small fraction of the TE of BMI.

We also applied TMAO, TMAR, and IMA to the data, adjusting for the matching variables. TMAO and TMAR did not detect statistically significant factors mediating the relationship between BMI and the risk of ER+ breast cancers (Web Figures 23 and 24). Similarly, IMA did not identify statistically significant metabolites mediating the relationship (Web Figure 25). Further details are in Web Appendix G.

6 | DISCUSSION

We proposed a latent variable model for high dimensional mediation analysis (LVMA). Our theoretical results show that the model parameters are identifiable, and LVMA estimates those parameters that have the so-called oracle properties of consistency and efficiency. Our simulation results further show that using LVMA, when appropriate, can significantly increase the number of discovered mediators. LVMA, by considering all variables simultaneously, efficiently estimates all parameters in the model. Specifically, using LVMA, we better estimate the mediating factors by using additional information about the exposure and outcome, as opposed to only using the information about the biomarkers. However, under model misspecification, such as when the exposure directly affects individual biomarkers, the assumption of latent variables can reduce the power to detect such biomarkers.

We highlight a couple of features of our method. First, we extend current literature on mediation analysis with a single latent mediator (Muthén and Asparouhov, 2015; Albert et al., 2016) to handle multiple latent mediators. Second, although some recent studies have started exploring penalized structural equation modeling (Jacobucci et al., 2016), these methods were not designed to handle the $p > > n$ setting. Third, we extend our mediation model and, more generally, sparse factor analysis to accommodate case-control sampling.

None of the previously published methods (Boca et al., 2013; Huang and Pan, 2016; Zhang et al., 2016; Zhao and Luo, 2016; Chen et al., 2018 explicitly assumed the latent structure illustrated in Figure 1, but two methods did so implicitly. Huang and Pan (2016) take an approach similar to TMAR but do not impose any sparsity on the factors. Their objective is also fundamentally different as they are testing whether the set of all biomarkers are mediators as opposed to trying to identify the subset that is mediators. Chen et al. (2018) aims to identify linear combinations of biomarkers that are associated with both the exposure and outcome. However, their approach does not allow for factors that are only associated with the exposure or the outcome. Therefore, the biomarkers associated with only one of those variables get mistakenly included in the "direction of mediation."

Several problems remain to be addressed in future work. Our latent model in (6–9) does not detect the existence of biomarkers that directly mediate the effect of E on Y. As evidenced by a large number of nonzero loadings in our application, the shrinkage of loadings for unrelated biomarkers may not be satisfactorily using EBIC. Some of our assumptions could also be violated in real-world examples. The factors need not be independent conditional on the exposure; the biomarkers need not be normally distributed; for retrospective sampling, the exposure needs to be normally distributed. We note the latter may be accommodated by the semiparametric approach based on Qin (1998) but that the discussion is beyond the scope of this paper. Despite these limitations, we believe the newly proposed LVMA offers a novel important tool for detecting biological mediators.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This study utilized the computational resources of the NIH HPC Biowulf cluster.

REFERENCES

Albert JM, Geng C. and Nelson S. (2016) Causal mediation analysis with a latent mediator. Biometrical Journal, 58, 535–548. [PubMed: 26363769]

Anderson TW and Rubin H. (1956) Statistical Inference in Factor Analysis. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry, 111–150, University of California Press, Berkeley, CA. https://projecteuclid.org/euclid.bsmsp/1200511860

Assi N, Fages A, Vineis P, Chadeau-Hyam M, Stepien M. and Duarte-Salles Tet al. (2015) A statistical framework to model the meeting-in-the-middle principle using metabolomic data: Application to hepatocellular carcinoma in the epic study. Mutagenesis, 30, 743–753. [PubMed: 26130468]

Bai J. and Li K. (2012) Statistical analysis of factor models of high dimension. The Annals of Statistics, 40, 436–465.

Boca SM, Sinha R, Cross AJ, Moore SC and Sampson JN (2013) Testing multiple biological mediators simultaneously. Bioinformatics, 30, 214–220. [PubMed: 24202540]

Calle EE and Kaaks R. (2004) Overweight, obesity and cancer: Epidemiological evidence and proposed mechanisms. Nature Reviews Cancer, 4, 579. [PubMed: 15286738]

Carroll RJ, Ruppert D, Stefanski LA and Crainiceanu CM (2006) Measurement error in nonlinear models: A modern perspective. New York: CRC Press.

Chen J. and Chen Z. (2008) Extended bayesian information criteria for model selection with large model spaces. Biometrika, 95, 759–771.

Chen OY, Crainiceanu C, Ogburn EL, Caffo BS, Wager TD and Lindquist MA (2018) High-dimensional multivariate mediation with application to neuroimaging data. Biostatistics, 19, 121–136. [PubMed: 28637279]

Daniel R, De Stavola B, Cousens S. and Vansteelandt S. (2015) Causal mediation analysis with multiple mediators. Biometrics, 71, 1–14. [PubMed: 25351114]

Fan J. and Li R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96, 1348–1360.

Hirose K. and Yamamoto M. (2014) Estimation of an oblique structure via penalized likelihood factor analysis. Computational Statistics & Data Analysis, 79, 120–132.

Huang Y-T and Pan W-C (2016) Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. Biometrics, 72, 402–413. [PubMed: 26414245]

Imai K, Keele L. and Tingley D. (2010) A general approach to causal mediation analysis. Psychological Methods, 15, 309–334. [PubMed: 20954780]

Jacobucci R, Grimm KJ and McArdle JJ (2016) Regularized structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal, 23, 555–566. [PubMed: 27398019]

le Cessie S, Debeij J, Rosendaal FR, Cannegieter SC and Vandenbrouckea JP (2012) Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. Epidemiology, 23, 551–560. [PubMed: 22526092]

Leek JT and Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genetics, 3, e161.

Moore SC, Playdon MC, Sampson JN, Hoover RN, Trabert B. and Matthews CEet al. (2018) A metabolomics analysis of body mass index and postmenopausal breast cancer risk. Journal of the National Cancer Institute, 110, 588–597. [PubMed: 29325144]

Muthén B. and Asparouhov T. (2015) Causal effects in mediation modeling: An introduction with applications to latent variables. Structural Equation Modeling: A Multidisciplinary Journal, 22, 12–23.

Qin J. (1998) Inferences for case-control and semiparametric twosample density ratio models. Biometrika, 85, 619–630.

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

Srivastava S, Engelhardt BE and Dunson DB (2017) Expandable factor analysis. Biometrika, 104, 649–663. [PubMed: 29430037]

Steen J, Loeys T, Moerkerke B. and Vansteelandt S. (2017) Flexible mediation analysis with multiple mediators. American Journal of Epidemiology, 186, 184–193. [PubMed: 28472328]

Valeri L, Lin X. and VanderWeele TJ (2014) Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. Statistics in Medicine, 33, 4875–4890. [PubMed: 25220625]

VanderWeele TJ and Tchetgen Tchetgen EJ (2016) Mediation analysis with matched case-control study designs. American Journal of Epidemiology, 183, 869–870. [PubMed: 27076669]

VanderWeele T. and Vansteelandt S. (2014) Mediation analysis with multiple mediators. Epidemiologic Methods, 2, 95–115. [PubMed: 25580377]

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

Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B. and Yoon Get al. (2016) Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics, 32, 3150–3154. [PubMed: 27357171]

Zhao Y, and Luo X. (2016). Pathway lasso: Estimate and select sparse mediation pathways with high dimensional mediators. arXiv preprint arXiv:1603.07749.

Zou H. (2006) The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101, 1418–1429.

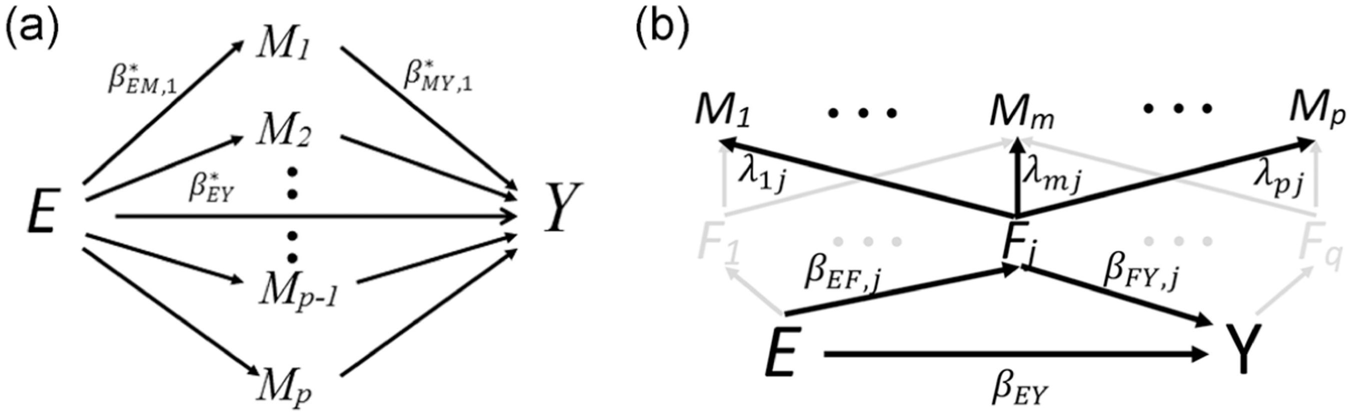


FIGURE 1.

Causal graphs of the mediation models. A, Traditional mediation analysis where the exposure influences the biomarkers. B, Latent-variable mediation analysis where the exposure influences a set of q latent, or unmeasured, factors and those factors influence both the biomarkers and the outcome. To simplify the figure, we highlight the arrows and notation for only a single factor. In the sparse scenario, we expect most effects to be 0 (i.e., λ, β_{EY}, and β_{FY}, are usually 0). We define the $p \times q$ matrix, Λ, so that the m, jth entry is λ_{mj}

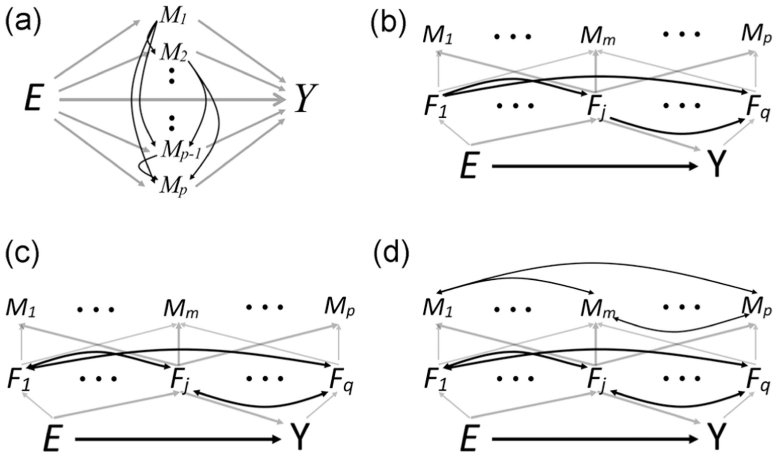
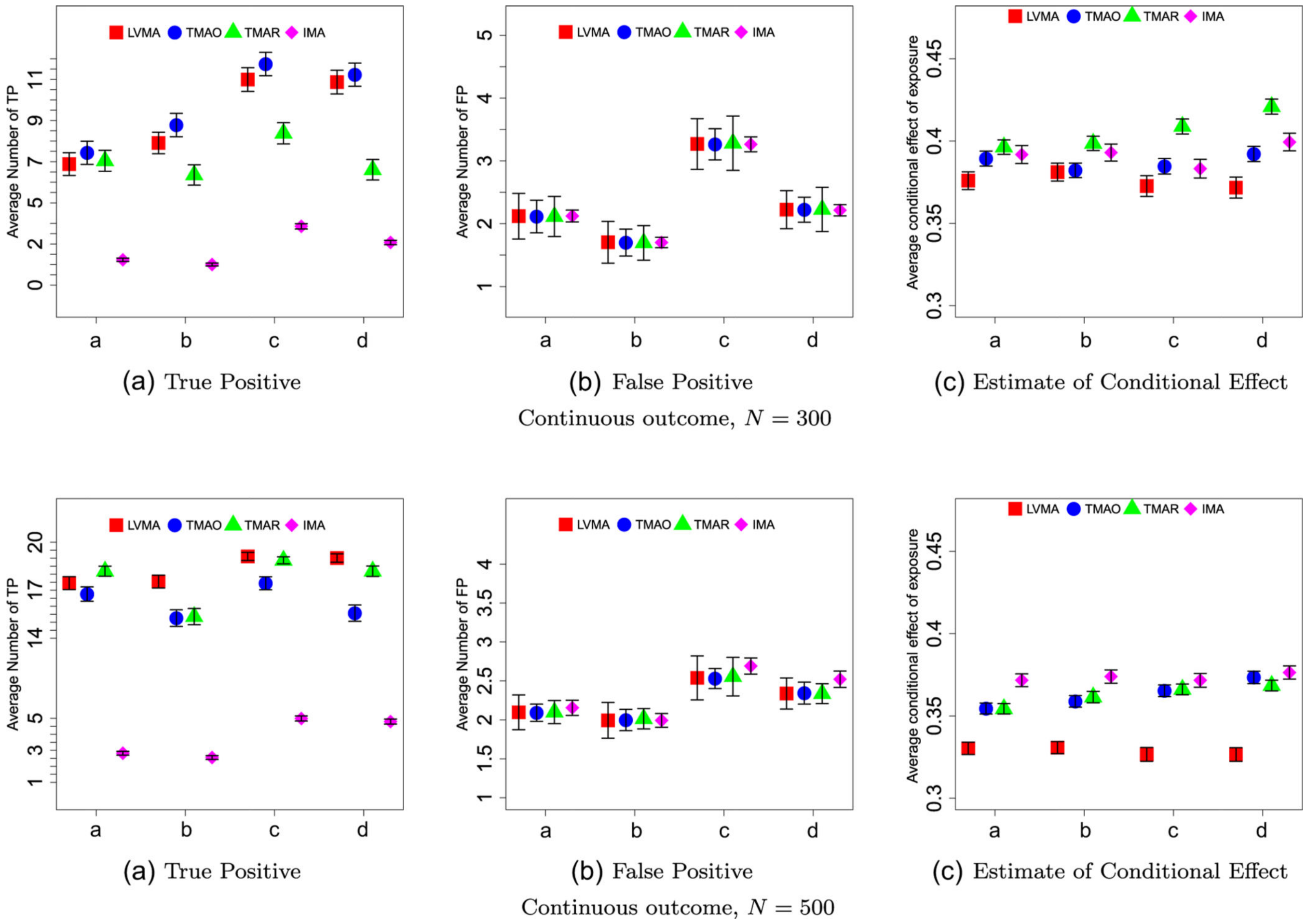


FIGURE 2.

Alternative causal graphs of the mediation models. A, Causally ordered mediation analysis where the exposure influences the biomarkers. B, Latent-variable mediation analysis where the exposure influences a set of q causally ordered latent factors. C, Latent-variable mediation analysis where the exposure influences a set of q bidirectionally connected latent factors. D, Latent-variable mediation analysis where the exposure influences a set of q bidirectionally connected latent factors and those factors influence bidirectionally connected biomarkers



(a) True Positive (b) False Positive (c) Estimate of Conditional Effect

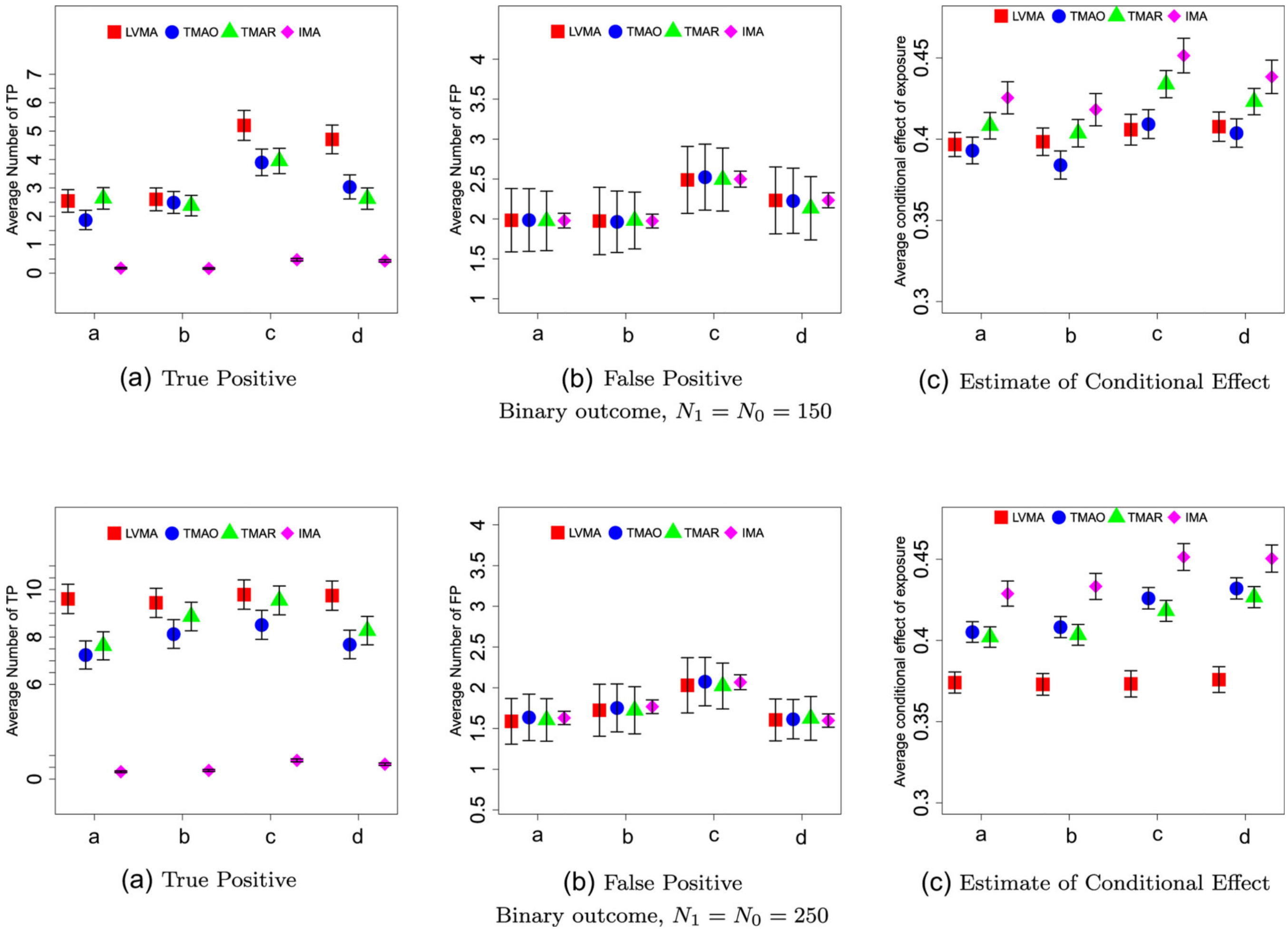
Continuous outcome, $N = 300$

(a) True Positive (b) False Positive (c) Estimate of Conditional Effect

Continuous outcome, $N = 500$

FIGURE 3.

Continuous outcome and large factor effects ($\lambda_1 = 0.3$). The panels, labeled A-C, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red = LVMA; blue = TMAO; green = TMAR; purple = IMA) and for four scenarios (a: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; b: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; c: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; d: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with N = 300 and N = 500 subjects, respectively. The whiskers show two standard errors around the average estimates.



(a) True Positive

(b) False Positive

(c) Estimate of Conditional Effect

Binary outcome, $N_1 = N_0 = 150$

(a) True Positive

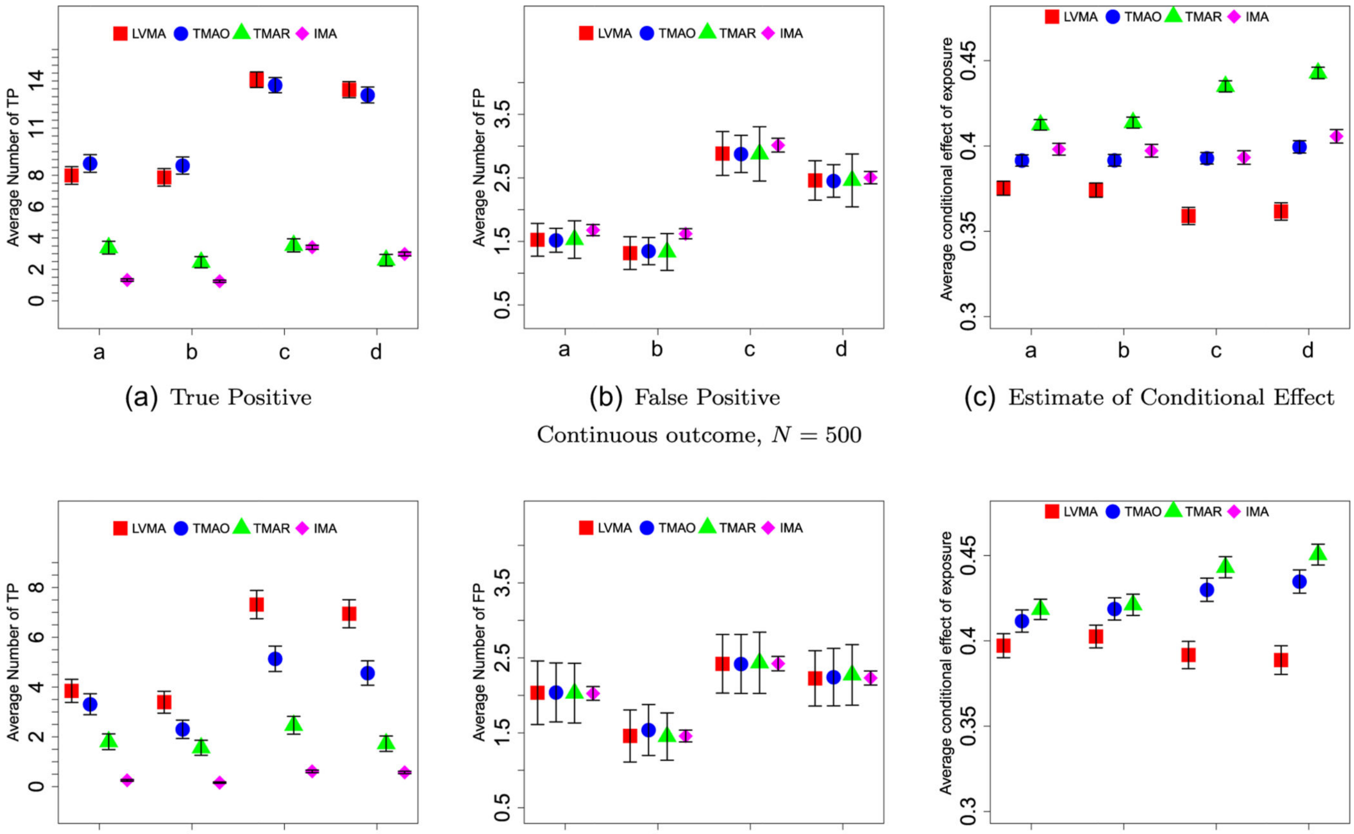
(b) False Positive

(c) Estimate of Conditional Effect

Binary outcome, $N_1 = N_0 = 250$

FIGURE 4.

Binary outcome and large factor effects ($\lambda_1 = 0.3$). The panels, labeled A-C, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red = LVMA; blue = TMAO; green = TMAR; purple = IMA) and for four scenarios (a: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; b: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; c: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; d: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with N = 300 and N = 500 subjects, respectively. The whiskers show two standard errors around the average estimates.



(a) True Positive (b) False Positive (c) Estimate of Conditional Effect

Continuous outcome, $N = 500$

(a) True Positive (b) False Positive (c) Estimate of Conditional Effect

Binary outcome, $N_1 = N_0 = 250$

FIGURE 5.

Small factor effects ($\lambda_1 = 0.25$). The panels, labeled A-C, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red = LVMA; blue = TMAO; green = TMAR; purple = IMA) and for four scenarios (a: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; b: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; c: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; d: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with continuous and binary outcomes, respectively (N = 500). The whiskers show two standard errors around the average estimates.

TABLE 1

Metabolites linking BMI and breast cancer

Metabolite	**Loading (λ_{mj})**
5α-Pregnan-3β,20-α-diol monosulfate (2)	0.46
5α-Pregnan-3β,20-α-diol disulfate	0.49
Etiocholanolone glucuronide	0.49
16α-Hydroxydehydroepiandrosterone 3-sulfate	0.51
Epiandrosterone sulfate	0.60
Androsterone sulfate	0.61
4-Androsten-3β,17-β-diol monosulfate (2)	0.61
Pregnen-diol disulfate	0.63
4-Androsten-3β,17-α-diol monosulfate (3)	0.64
5α-Androstan-3β,17-β-diol disulfate	0.65
21-Hydroxypregnenolone disulfate	0.65
Pregnen steroid monosulfate	0.65
4-Androsten-3β,17-β-diol monosulfate (1)	0.67
4-Androsten-3β,17-β-diol disulfate (2)	0.70
4-Androsten-3β,17-β-diol disulfate (1)	0.70
Dehydroisoandrosterone sulfate (DHEA-S)	0.75

This list includes those metabolites that were strongly affected ($\gamma > 0.4$) by the factor mediating increased BMI and ER+ breast cancer in PLCO.