Please note: The version of record of this article, first published in the *Journal of Applied Statistics*, is available online at Publisher's website: https://doi.org/10.1080/02664763.2024.2367148

A Bayesian High–dimensional Mediation Analysis for Multilevel Genome–wide Epigenetic Data

Xi Qiao[a], Duy Ngo[a], Bilinda Straight[b], Belinda L. Needham[c], Charles E. Hilton[d], and Amy Naugle[e]

[a]Statistics, Western Michigan University, Michigan, USA; [b] School of Environment, Geography, & Sustainability, Western Michigan University, Michigan, USA; [c]Epidemiology, University of Michigan, Michigan, USA; [d]Anthropology, University of North Carolina at Chapel Hill, North Carolina, USA; [e]Psychology, Western Michigan University, Michigan, USA.

ARTICLE HISTORY

Compiled March 29, 2024

ABSTRACT

Causal mediation analysis has increasingly become a popular practice in various clinical trials and epidemiological applications to evaluate whether an intermediate variable is on the pathway between the exposure of interest and a response. Previous mediation analyses in the literature mainly focused on settings with a single or low–dimensional mediators and single–level data. In this article, we propose a Bayesian causal mediation analysis method that can handle our multilevel intergenerational epigenetic mechanisms study (IEMS) with high–dimensional mediators. Specifically, we develop a Bayesian hierarchical model for data with such complexity, and then employ the Bayesian spike–and–slab priors on the exposure–mediator–outcome effect pathway to identify active mediators involved in mediation. We derive the natural indirect and direct effects based on our hierarchical model and provide statistical inference based on Markov chain Monte Carlo (MCMC) methods. Our simulation study demonstrates that our proposed Bayesian method outperforms other alternative methods in various scenarios. We further illustrate the utility of our method to IEMS to assess the causal mechanisms between maternal exposure to climate

CONTACT Duy Ngo. Email: duy.ngo@wmich.edu

extremes and offspring's growth outcomes through DNA methylation.

KEYWORDS

High-dimensional mediation analysis; Bayesian variable selection; Multilevel data modeling; Epigenetic study.

1. Introduction

In clinical trials and epidemiological studies, it is often of interest to evaluate an overall treatment or exposure effect on a response. Of even greater interest is to explain the underlying mechanism by which the effect of an exposure on the outcome is mediated through a casual intermediate variable or mediator. In practice, mediation analysis utilizes one or more measured mediators hypothesized to lie on the causal pathway between the exposure and the outcome. Typically, the mediation analysis involves the decomposition of the overall exposure effect into an indirect (mediation) effect, which is the effect of an exposure explained by a mediator, and a direct effect, which is the effect of an exposure unexplained by that mediator. The two commonly used approaches of mediation analysis include linear regression models within the framework of linear structural equation modeling (SEM) [5, 32, 33], and causal mediation analysis based on the counterfactual framework [1, 43]. Recent advances in causal inference have generalized and extended the mediation model intuitively developed in the SEM approach by precisely defining the indirect and direct effects using potential outcomes, giving the identification conditions of these effects, and lastly incorporating nonlinearities and interactions [19, 38, 40, 49, 53, 54].

To date most research in mediation analysis has been devoted to the case of a single mediator, with some attention given to the case of low dimensional mediators, meaning that there are typically only one or few mediators [2, 10, 20, 52]. However, high dimensional mediators often exist in substantive research. This article is motivated by an intergenerational epigenetic mechanisms study (IEMS) of the effects of maternal exposure to climate extremes such as drought on offspring DNA methylation (DNAm) in the Samburu people of northern Kenya [46], which is assessed by using the Infinium MethylationEPIC BeadChip array to measure methylation at about 850,000 cytosine–

phosphate-guanine (CpG) sites, resulting in high–dimensional data (see Figure 1). When the mediator space is high–dimensional, with larger numbers of mediators and correlation among them, estimating individual path coefficients in the standard way is not feasible because of difficulties modeling the appropriate relationship between variables in this setting [6]. Moreover, the standard estimation procedure becomes unstable when the number of mediators significantly exceeds the number of observations, known as a small–N–large–P problem. This is because the sample covariance matrix is singular, with at least $P - N$ of the smallest eigenvalues estimated to be zero, so its inverse will not exist, resulting in an over–inflated standard error[14, 23].

Some recent researchers have proposed methods to accommodate high-dimensional mediators. Zhang *et al.* [57] demonstrates the practical performance of the combination of the Sure Independent Screening approach (SIS)[11], Minimax Concave Penalty (MCP) techniques, and multiple testing procedure with controlled False Discovery Rate (FDR), to identify the subset of DNA methylation sites that mediate the association between smoking and reduced lung function. This method was implemented in R package *HIMA* and later was updated to *HIMA2* with the de-biased LASSO procedure to estimate the regression parameters [41]. To estimate and select many pathways effects simultaneously, Zhao and Luo [58] introduced a pathway LASSO method, a convex relaxation of the non-convex product function, for a sparse mediation model with structural equation modeling approach. Song *et al.* [45] developed a Bayesian inference method using Bayesian Sparse Linear Mixed Model (BSLMM), which imposes continuous shrinkage priors to identify the inactive and active mediators [59]. To account for the possible correlation among the mediators that mediate the association between exposure and outcome, Song *et al.* later [44] proposed two Bayesian hierarchical models, one with a Gaussian mixture prior for correlated mediator selection, and the other with a Potts mixture prior. However, the existing methods are not readily applicable with multilevel data.

What makes high–dimensional mediation analysis even more challenging is the complex structure of data. Multilevel data is often encountered in many disciplines such as medicine where patients are nested within hospital, education where students are nested within schools, or children within mother in our IEMS data (Figure 2). This

type of multilevel data violates the assumption of independence for traditional regression methods, so it will lead to biased estimates. Several mediation analysis methods [25–27, 56] for multilevel data are based on the frequentist approach and single or low dimensional mediators. Therefore, in this paper, we aim to develop a Bayesian causal mediation analysis method that can handle multilevel and high–dimensional mediators, for IEMS data.

Specifically, in the high-dimensional mediation setting, we propose a novel mediator identification procedure to detect active mediators that are involved in mediation by adopting Bayesian shrinkage priors to capture the sparsity of the exposure–mediator and mediator–outcome effect pathways. Our Bayesian approach serves as an extension of [45], and naturally adapts to hierarchically correlated effects from multilevel data through conditionally specified hierarchical priors, such as specifying the likelihood of the data given unknown random individual effects, determining the density of the population of random effects, and then providing priors (or hyperpriors) on the parameters of the population density. Note that the parameters obtained from random effects models do not necessarily have the same interpretation as under marginal or population–averaged models [15, 29, 37]. We focus on conditionally specified hierarchical and random effect models, and on MCMC estimation via conditional likelihood with random effects as part of the parameter set [9].

The paper is organized as follows. In Section 2, we introduce notation and briefly review the single mediator analysis. In Section 3, we present our proposed Bayesian causal mediation analysis for multilevel data with high-dimensional mediators. Next we evaluate and compare the performance of our method with univariate mediation analysis and other existing methods via numerical simulations in Section 4. We then apply our method to the IEMS dataset, and the discussion and conclusion are provided in Section 5. The proofs and algorithms are provided in the Supplemental materials.

2. A brief review of single mediator case

We first briefly review the standard causal mediation analysis with a single mediator (see [18] for a more detailed explanation). Let Y_i be an observed outcome for an

individual i, T_i denote a binary exposure or treatment, which equals one if individual i receives the treatment/exposed and is zero otherwise. Let M_i be an observed value of the mediator that may be on the pathway from the treatment to the outcome (as shown in Figure 3). Moreover, let $Y_i(t, m)$ denote the potential outcome under the treatment $T_i = t$ and the mediator $M_i = m$, and likewise, $M_i(t)$ is a potential mediator value with the observed treatment t. Under the formal causal mediation analysis [39, 43], the causal mediation effect or natural indirect effect for individual i given the treatment status $t \in \{0, 1\}$ is defined as $\text{NIE}_i(t) = Y_i(t, M_i(1)) - Y_i(t, M_i(0))$, which compares the potential outcome that would be observed when the individual i under treatment t and mediator is changed from $M(0)$ to $M(1)$. The unit–level direct effect of treatment T on outcome Y is $\text{NDE}_i(t) = Y_i(1, M_i(t)) - Y_i(0, M_i(t))$, which compares the potential outcome under treatment and control while the mediator M is at its natural level under treatment t. The total effect is the sum of the natural direct and indirect effect, i.e.

$$Y_i(1) - Y_i(0) = Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \text{NIE}_i(t) + \text{NDE}_i(1 - t).$$

From these unit–level quantities of interest, we can define the population average effect for each quantity such as $\text{ANIE}(t) = \text{E}\left[\text{NIE}_i(t)\right]$, and $\text{ANDE}(t) = \text{E}\left[\text{NDE}_i(t)\right]$. The goal of casual mediation analysis is, therefore, to decompose the total treatment effect into the direct and indirect effects, and these effects can be parameterized with two linear regressions separately,

$$M_i = \alpha_0 + \alpha_T T_i + \boldsymbol{\alpha_X}^T \boldsymbol{X_i} + \epsilon_{M_i}, \tag{1}$$

$$Y_i = \beta_0 + \beta_T T_i + \beta_M M_i + \boldsymbol{\beta_X}^T \boldsymbol{X_i} + \epsilon_{Y_i}, \tag{2}$$

where $\boldsymbol{X_i}$ is a $c \times 1$ vector of observed pre–treatment confounders, and ϵ_{M_i} and ϵ_{Y_i} are normally distributed, independent random noise variables. The effects α_T, β_M and β_T are illustrated in Figure 3. After fitting these two models, we can obtain $\hat{\alpha}_T \hat{\beta}_M$ as an estimate of the $\text{ANIE}(t)$, whereas the estimated coefficient $\hat{\beta}_T$ is an estimate of $\text{ANDE}(t)$. A good property of natural effects is that we can compute the ratio of

indirect effect and total effect to estimate the proportion of the effect that goes through the mediator. This is very useful in explaining how a specific intervention works.

In practice, we do not observe all the potential outcomes for each individual, so the individual level effects cannot be identified. Imai *et al.*[19] showed that the population average effects can be identified under the following sequential ignorability assumption,

$$\{Y_i(t^*, m), \;\; M_i(t)\} \perp T_i | \boldsymbol{X_i}, \tag{3}$$

$$Y_i(t^*, m) \perp M_i(t) | T_i = t, \boldsymbol{X_i}. \tag{4}$$

The assumption (3) states that given the observed pre–treatment confounders, there is no confounding between the outcome and exposure, and there is no unmeasured confounding between all mediators and the exposure. The second assumption (4) implies that there is no confounding between the outcome–mediator relationship after controlling for the exposure. For a high–dimensional mediator, researchers can fit two linear regressions (1) and (2) for each mediator. However, high–dimensional mediators should be fit in a single model rather than one at a time to improve power [52]. Therefore, we introduce our proposed Bayesian approach for the high–dimensional mediators in the following section.

3. Bayesian hierarchical model for multilevel data with high–dimensional mediators (BHMM)

Our focus is on the 1–1–1 mediation model [28], in which three variables, response Y, mediator M and exposure T, are measured at level–1 (see Figure 2). Suppose that we observe a data from N children (level–1) and for each child $i = 1, \ldots, n$ in mother $j = 1, \ldots, N$ (level–2), we observe the data $D_{ij} = (T_{ij}, \boldsymbol{M_{ij}}, Y_{ij}, \boldsymbol{X_{ij}})$, where T_{ij} represents the exposure indicator, $\boldsymbol{M_{ij}} = (M_{ij,1}, \ldots, M_{ij,p})^T$ is a $p \times 1$ vector of continuous DNAm mediators, Y_{ij} the continuous outcomes of interest and $\boldsymbol{X_{ij}}$ is a $c \times 1$ vector of measured pre–exposure confounders.

The level one model can be expressed as follows:

$$\boldsymbol{M_{ij}} = \boldsymbol{\alpha_{0_j}} + \boldsymbol{\alpha_T} T_{ij} + \boldsymbol{\alpha_X} \boldsymbol{X_{ij}} + \boldsymbol{\epsilon_{M_{ij}}}, \tag{5}$$

$$Y_{ij} = \beta_{0_j} + \beta_T T_{ij} + \boldsymbol{\beta_M}^T \boldsymbol{M_{ij}} + \boldsymbol{\beta_X}^T \boldsymbol{X_{ij}} + \epsilon_{Y_{ij}}, \tag{6}$$

where $\epsilon_{Y_{ij}} \sim \mathrm{N}(0, \sigma_Y^2)$ denotes a random error, and a random error vector $\boldsymbol{\epsilon_{M_{ij}}}$ is assumed to have a multivariate normal distribution (MVN), i.e. MVN$(\mathbf{0}, \boldsymbol{\Sigma_M})$. Here, $\boldsymbol{\Sigma_M}$ captures the correlation structure among the mediators. Furthermore, we assume that $\epsilon_{Y_{ij}}$ and $\boldsymbol{\epsilon_{M_{ij}}}$ are independent.

At level 2, the models for the coefficients in Equation 5 and (6) are as follows:

$$\boldsymbol{\alpha_{0_j}} = \boldsymbol{\alpha_0} + \boldsymbol{\nu_j}, \tag{7}$$

$$\beta_{0_j} = \beta_0 + \mu_j, \tag{8}$$

where $\mu_j \sim \mathrm{N}(0, \sigma_{\beta_0}^2)$ and $\boldsymbol{\nu_j} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma_{\alpha_0}})$ denote the error of intercepts. For simplicity and from our previous analysis of IEMS in [46, 48], we focus on a random intercept model, but one can easily incorporate other multilevel structures and include random slopes.

Assume that there are no interactions between exposure and mediators in Equation (5). Under the sequential ignorability conditions [52], we showed the detailed proof in the Supplemental materials that the conditional ANDE, conditional ANIE and

conditional ATE can be expressed as follows.

$$\begin{aligned}
\text{ANDE} &= \mathrm{E}\left[\mathrm{NDE}_{ij}(t)|\boldsymbol{\alpha_{0j}}, \beta_{0j}, \boldsymbol{X_{ij}}\right] \\
&= \mathrm{E}\left[Y_{ij}(1, \boldsymbol{M_{ij}}(t)) - Y_{ij}(0, \boldsymbol{M_{ij}}(t))|\boldsymbol{\alpha_{0j}}, \beta_{0j}, \boldsymbol{X_{ij}}\right] \\
&= \beta_T, \qquad (9) \\
\text{ANIE} &= \mathrm{E}\left[\mathrm{NIE}_{ij}(t)|\boldsymbol{\alpha_{0j}}, \beta_{0j}, \boldsymbol{X_{ij}}\right] \\
&= \mathrm{E}\left[Y_{ij}(t, \boldsymbol{M_{ij}}(1)) - Y_{ij}(t, \boldsymbol{M_{ij}}(0))|\boldsymbol{\alpha_{0j}}, \beta_{0j}, \boldsymbol{X_{ij}}\right] \\
&= \boldsymbol{\alpha_T}^T \boldsymbol{\beta_M}, \qquad (10) \\
\text{ATE} &= \mathrm{E}\left[\mathrm{TE}_{ij}(t)|\boldsymbol{\alpha_{0j}}, \beta_{0j}, \boldsymbol{X_{ij}}\right] \\
&= \mathrm{E}\left[\mathrm{NDE}_{ij}(t)|\boldsymbol{\alpha_{0j}}, \beta_{0j}, \boldsymbol{X_{ij}}\right] + \mathrm{E}\left[\mathrm{NIE}_{ij}(t)|\boldsymbol{\alpha_{0j}}, \beta_{0j}, \boldsymbol{X_{ij}}\right] \\
&= \beta_T + \boldsymbol{\alpha_T}^T \boldsymbol{\beta_M}. \qquad (11)
\end{aligned}$$

Equation 10 shows that the ANIE is the sum of the product of $\alpha_{T,k}$ and $\beta_{M,k}$ for $k = 1, \ldots, p$, and this product does not correspond to the NIE of a single kth mediator due to interrelated mediators. The kth mediator is an active mediator when both $\alpha_{T,k}$ and $\beta_{M,k}$ are non-zero, so there are three situations where the kth mediator will be identified as an inactive mediator: (1) $\alpha_{T,k}$ is non-zero but $\beta_{M,k}$ is zero; (2) $\alpha_{T,k}$ is zero but $\beta_{M,k}$ is non-zero; and (3) both $\alpha_{T,k}$ and $\beta_{M,k}$ are zero. Selecting the active mediators turns out to be a variable selection problem. Next, we introduce the Bayesian approach to identifying active mediators in the mediation models (5) and (6).

3.1. Prior Specification

In Bayesian framework, a spike–and–slab prior is used for variable selection or shrinkage estimation [22]. George and McCulloch [13] introduced a mixture of two normal distributions with a Bernoulli latent variable d. For example, the priors of the kth coefficient regression $\alpha_{T,k}$ in $\boldsymbol{\alpha_T}$ as follows

$$\alpha_{T,k}|d_{\alpha,k} \overset{i.i.d}{\sim} d_{\alpha,k}\mathrm{N}(0, \sigma^2_{\alpha_{T,1}}) + (1 - d_{\alpha,k})\mathrm{N}(0, \sigma^2_{\alpha_{T,0}}), \quad k = 1, \ldots, p, \qquad (12)$$

where $d_{\alpha,k} \sim \text{Bernoulli}(\theta_{\alpha,k})$, and the two normal distributions have the same zero mean but different variances, such as $\sigma^2_{\alpha_{T,1}}$ has a large value while $\sigma^2_{\alpha_{T,0}}$ is suitably small. Under this prior, if the kth mediator is active, $\alpha_{T,k}$ will be drawn from a zero–mean normal distribution with large variance, and the opposite occurs when $\alpha_{T,k}$ will be drawn from a point mass at 0, e.g. a zero–mean normal distribution with extremely small variance. The prior hierarchy for $\alpha_{T,k}$ is completed by choosing a prior for $\theta_{\alpha,k}$, and a common choice for this hyperparameter is beta distribution, i.e. $\theta_{\alpha,k} \sim \text{Beta}(a_{\alpha,k}, b_{\alpha,k})$. The values of $a_{\alpha,k}$ and $b_{\alpha,k}$ depend on our prior belief whether the corresponding mediator is active or not. For example, when $a_{\alpha,k} = b_{\alpha,k} = 1$, it leads to a non-informative prior, and thus we do not have favor for $\alpha_{T,k}$ being drawn from normal distribution with large variance or small variance and allow the data to determine. When $a_{\alpha,k}$ is large and $b_{\alpha,k}$ is small, we favor $\alpha_{T,k}$ to be kept in the model. We also assign the spike and slab prior to kth coefficient regression $\beta_{M,k}$ in $\boldsymbol{\beta_M}$ as below

$$\beta_{M,k}|d_{\beta,k} \overset{i.i.d}{\sim} d_{\beta,k}\text{N}(0, \sigma^2_{\beta_{M,1}}) + (1 - d_{\beta,k})\text{N}(0, \sigma^2_{\beta_{M,0}}), \quad k = 1, \ldots, p, \tag{13}$$

where $d_{\beta,k} \sim \text{Bernoulli}(\theta_{\beta,k})$ and $\theta_{\beta,k} \sim \text{Beta}(a_{\beta,k}, b_{\beta,k})$. Under this prior specification, $\beta_{M,k}$ exhibits properties similar to those of $\alpha_{T,k}$.

For hierarchical models, the use of a flat or excessively diffuse prior may lead to an improper posterior distribution [17]. To ensure propriety of the joint posterior distribution, we choose weakly informative prior distributions for the remaining parameters. For example, normal priors with zero mean and large variance for the regression coefficients and inverse–gamma priors with small parameter values for the variance components. For $\ell = 1, \cdots, c$, the prior distributions of $\{\boldsymbol{\alpha_{X,\ell}}, \boldsymbol{\alpha_0}\}$ are taken to be a multivariate normal distribution; normal distributions for $\{\beta_T, \beta_{X,\ell}, \beta_0\}$; inverse–gamma distributions for $\left\{\sigma^2_{\alpha_{T,1}}, \sigma^2_{\alpha_{T,0}}, \sigma^2_{\beta_{M,1}}, \sigma^2_{\beta_{M,0}}, \sigma^2_{\beta_0}, \sigma^2_Y\right\}$; and inverse–Wishart distributions for $\boldsymbol{\Sigma_{\alpha_0}}$ and $\boldsymbol{\Sigma_M}$. We will later discuss the specified prior distributions in our simulation study (Section 4).

3.2. Posterior Sampling Strategy

Based on the prior specifications in the previous sections, we employ the Gibbs sampling method based on fully conditional probability with conjugate distribution to obtain the posterior samples for our proposed BHMM. Latent variable d_β, d_α could be sampled from Bernoulli distribution with inclusion probability $P(d_{\alpha,k} = 1|\theta_{\alpha,k}, \alpha_{T,k}, \sigma^2_{\alpha_{T,0}}, \sigma^2_{\alpha_{T,1}})$ and $P(d_{\beta,k} = 1|\theta_{\beta,k}, \beta_{M,k}, \sigma^2_{\beta_{T,0}}, \sigma^2_{\beta_{T,1}})$ using conditional probability given the observed data and the other parameters [13]. We then adopt the posterior inclusion probability (PIP) which provides the probability of $\alpha_{T,k}$ and $\beta_{M,k}$ coming from the normal distribution with a larger variance (the slab). The PIP of $\alpha_{T,k}$ and $\beta_{M,k}$ can be estimated by averaging the inclusion probability [34].

Under the Bayesian variable selection framework, the kth mediator is active when both $\alpha_{T,k}$ and $\beta_{M,k}$ are from the normal distribution with larger variance, so both PIP of $\alpha_{T,k}$ and $\beta_{M,k}$ are at large value. Mediators with either or both PIP of $\alpha_{T,k}$ and $\beta_{M,k}$ are small will be identified as inactive mediator. In order to establish the appropriate thresholds of $\alpha_{T,k}$ and $\beta_{M,k}$, we first sort PIPs of $\alpha_{T,k}$ and $\beta_{M,k}$ in the descending order. The threshold values are then determined based on the PIP of $\alpha_{T,k}$ and $\beta_{M,k}$ that will ensure an overall false positive rate (FPR), the ratio between the number of false positive (FP) mediators and the true inactive mediators, i.e. True Negative (TN) + False Positive (FP) mediators (see Table 1 for mediator classification), is controlled at certain value, i.e., 0.05 [55]. FPR controls the number of the mis–classification of the true inactive mediators. If both $\alpha_{T,k}$ and $\beta_{M,k}$ are greater than the threshold, the kth mediator will be determined as active. Algorithm 1 presents the posterior sampling algorithm for parameters and latent variables for a total of R iterations (excluding B draws as burn–in). Detailed descriptions of the sampling distributions are provided in Supplemental materials (Section 2).

4. Simulation Study

We perform simulation studies to examine the performance of our proposed models. We generate data by mimicking the motivating IEMS data. In particular, the number of mother (level 2) is $N = 100$, and for each mother j for $j = 1, ...N$, there are

Algorithm 1 BHMM Posterior Sampling Algorithm

for each iteration from 1 to R **do**

Step 1: **for** kth mediator from 1 to p **do**

draw latent variable $d_{\alpha,k}$ from $f(d_{\alpha,k}|\theta_{\alpha,k}, \alpha_{T,k}, \sigma^2_{\alpha_{T,0}}, \sigma^2_{\alpha_{T,1}})$.

draw hyperparameter $\theta_{\alpha,k} \sim \text{Beta}(a_{\alpha,k} + d_{\alpha,k}, b_{\alpha,k} + 1 - d_{\alpha,k})$.

draw latent variable $d_{\beta,k}$ from $f(d_{\beta,k}|\theta_{\beta,k}, \beta_{M,k}, \sigma^2_{\beta_{T,0}}, \sigma^2_{\beta_{T,1}})$.

draw hyperparameter $\theta_{\beta,k} \sim \text{Beta}(a_{\beta,k} + d_{\beta,k}, b_{\beta,k} + 1 - d_{\beta,k})$.

end for

Step 2: Sample parameters in Equation 5 and 7.

draw $(\boldsymbol{\alpha_{0_j}}, \boldsymbol{\alpha_T}, \boldsymbol{\alpha_X}, \boldsymbol{\Sigma_M})$.

draw $(\boldsymbol{\alpha_0}, \boldsymbol{\Sigma_{\alpha_0}})$.

draw variance components $(\sigma^2_{\alpha_{T,0}}, \sigma^2_{\alpha_{T,1}})$ in spike–and–slab prior.

Step 3: Sample parameters in Equation 6 and 8.

draw $(\beta_{0_j}, \beta_T, \boldsymbol{\beta_M}, \boldsymbol{\beta_X}, \sigma^2_Y)$.

draw $(\beta_0, \sigma^2_{\beta_0})$.

draw variance components $(\sigma^2_{\beta_{M,0}}, \sigma^2_{\beta_{M,1}})$ in spike–and–slab prior.

end for

two children (level 1) $n = 2$ with one non-exposed subject $T_{1j} = 0$ and one exposed subject $T_{2j} = 1$. We consider two covariates $\boldsymbol{X_{ij}} = (X_{ij,1}, X_{ij,2})^T$, where $X_{ij,1}, X_{ij,2}$ are independently generated from $N(0, 1)$.

Given the exposure and covariates, the p mediators $\boldsymbol{M_{ij}} = (M_{ij,1}, \ldots, M_{ij,p})^T$ are generated from Equation 5. We examine mediators under different correlations, which is achieved by simulating $\boldsymbol{\epsilon_{M_{ij}}}$ from multivariate normal distribution with mean of $\mathbf{0}$ and variance-covariance structure $\boldsymbol{\Sigma_M}$ as follows

$$\boldsymbol{\Sigma_M} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \cdots & \rho & 1 \end{bmatrix},$$

where σ^2 denotes the variance of mediators, and ρ represents the correlation between different mediators. We first examine the setting with $\rho = 0$, where the mediators are generated independently, then further increase ρ to 0.05, 0.1 and 0.2. Among p mediators, we assume 10% of them are true active mediators that mediate the association between exposure T and outcome Y, i. e., both $\alpha_{T,k}$ and $\beta_{M,k}$ are non-zero. The rest of inactive mediators are generated as below: 10% mediators with non-zero

$\alpha_{T,k}$ and zero $\beta_{M,k}$; 10% mediators with zero $\alpha_{T,k}$ and non-zero $\beta_{M,k}$; 70% mediators with zero for both $\alpha_{T,k}$ and $\beta_{M,k}$.

For each scenario of correlation $\rho \in \{0, 0.05, 0.1, 0.2\}$, two sets of effect size are considered for comparison, i.e., effect size A: $\alpha_{T,k} = \beta_{M,k} = 1$, and effect size B: $\alpha_{T,k} = \beta_{M,k} = 0.3$. Here the setting of effect size B are based on previous estimated coefficient in our previous paper [46, 47]. Moreover, we set $\sigma^2 = 0.01$, and assume that the random intercept $\boldsymbol{\alpha_{0j}}$ are generated from multivariate normal distribution with mean $\boldsymbol{\alpha_0} = \mathbf{1}$, variance-covariance matrix of $\boldsymbol{\Sigma_{\alpha_0}} = 0.01\boldsymbol{I_p}$, where $\boldsymbol{I_p}$ denotes the p by p identity matrix. Finally, we generate outcome from Equation 6, where error terms $\epsilon_{Y_{ij}} \sim \mathrm{N}(0, 0.01)$, and random intercept $\beta_{0j} \sim \mathrm{N}(\beta_0, 0.01)$, where $\beta_0 = 1$. We set other regression coefficients $\boldsymbol{\alpha_X}$ as a p by c matrix of 1 in Equation 5, β_X as a vector of 1, and β_T as 1 in Equation 6.

For random error terms in the Equation 5 and 7, we assume $\boldsymbol{\Sigma_M}$, $\boldsymbol{\Sigma_{\alpha_0}} \sim$ Inverse-Wishart$(p, \boldsymbol{I_p})$, and $\sigma^2_{\beta_0}$, $\sigma^2_Y \sim$ inverse–gamma$(10^{-4}, 10^{-4})$. For the remaining regression coefficients, we assign the non–informative multivariate normal priors MVN$(\mathbf{0}, 10^4\boldsymbol{I_p})$ for $\boldsymbol{\alpha_0}$, $\boldsymbol{\alpha_{X,\ell}}$, and a normal prior N$(0, 10^4)$ for β_0, β_T, $\boldsymbol{\beta_{X,\ell}}$, for $\ell = 1, \ldots, c$. For the specification of hyperparameters, they are chosen to reflect vague prior knowledge about the parameters. Particularly, we set Beta$(1, 1)$ prior for $\theta_{\alpha,k}$, $\theta_{\beta,k}$ associated with the spike-and-slab priors, an inverse–gamma$(2, 1)$ for $\sigma^2_{\alpha_{T,1}}$, $\sigma^2_{\beta_{M,1}}$, and an inverse–gamma$(2, 10^{-4})$ for $\sigma^2_{\alpha_{T,0}}$, $\sigma^2_{\beta_{M,0}}$. To assess the impact of these hyperparameter choices on the BHMM's performance, we further conducted a sensitivity analysis. As shown in Section 3 of Supplemental materials, the analysis results are not sensitive to different hyperparameter values in term of power. Our MCMC implementation as described in Algorithm 1 is run for a total of $R = 120,000$ iterations with the first $B = 40,000$ samples discarded as burn–in.

For comparison, we consider the following relevant alternative approaches: (1) the Horseshoe prior [8], (2) the Bayesian Sparse Linear Mixed Model (BSLMM); and (3) the conventional univariate mediation analysis (UMA). We employ the horseshoe prior in our BHMM, instead of the spike–and–slab prior, for handling sparsity. The horseshoe prior assumes a global parameter that shrinks all the parameters towards zero and a half–Cauchy prior on the local shrinkage parameter that allows some parameters to

escape the global shrinkage. The full prior specification for $\beta_{M,k}$ is

$$\beta_{M,k}|\lambda_{M,k}, \tau_\beta \sim \mathrm{N}(0, \lambda^2_{M,k}\tau^2_\beta), \tag{14}$$

$$\lambda_{M,k}, \tau_\beta \sim \text{Half–Cauchy}(0, 1), \tag{15}$$

where the scale parameter of 1 for the half–Cauchy distribution as the default choice given in Carvalho et al. [8]. We use similar prior specification for $\alpha_{T,k}$. BSLMM is a Bayesian shrinkage approach for high dimension mediator by using continuous shrinkage priors to the indirect effects [45], but it does not account for correlation among observations within a cluster. The UMA approach is a frequentist approach by fitting two linear regressions (defined in Equations 5 and 6) for each mediator individually. We use the R package *bama* to perform BSLMM model, and the *mediation* package to run UMA approach (using quasi-Bayesian Monte Carlo method with 5000 sample draws for variance estimation [45, 50]).

To evaluate the performance of these methods, we compute the power of BHMM, Horseshoe, UMA, and BSLMM via the true positive rate (TPR), which is the ratio between the number of TP mediators and the true active mediators, i.e. TP + FN mediators (Table 1), at a controlled FPR of 0.05. To determine the threshold for UMA, we rank the p–value for average indirect effect of each mediator in the ascending order, and then select the p–value threshold that will control the overall FPR at 0.05. Mediator with p–value less than the threshold will be determined as active. With low–dimensional mediators $p = 100$ and high–dimensional mediators $p = \{200, 500, 1000\}$, the power are computed over 100 simulated data sets at FPR of 0.05.

Figure 4 shows that our BHMM model using spike–and–slab prior has comparable performance to the horseshoe prior for the number of mediators $p = 100, 200$, and has greater power than the horseshoe for a larger number of mediators $p = 500, 1000$. It is noticed that the powers of four methods decrease as p increases. Moreover, the BSLMM and UMA are inferior to our proposed BHMM model and horseshoe prior across all scenarios, especially for higher dimension of mediators. For example, under the settings $\rho = 0, p = 1000$ and small effect size B, the powers of BHMM and horseshoe are 0.62 and 0.6, while the BSLMM and UMA yields lower power of 0.32 and 0.40, respectively.

Comparing the cases with the same effect size but different correlation ρ, our proposed approach is observed to have competitive performance with the horseshoe prior, and our BHMM outperforms the BSLMM and UMA. It is noticed that BSLMM yields smaller power than UMA as the effect size increases. For a small effect size B, the higher correlation ρ may lead to relatively worse performance in BSLMM and UMA because these methods ignore the correlation among observations within a cluster, resulting in underestimated standard errors and thus inflation of the Type I error rate. With the largest $\rho = 0.2$, we observe a tendency for the UMA to increase the number of false negative because UMA tests $\beta_{M,k}$ at very small significance levels to reduce the large number of false positives, resulting in dramatically reduced power and an increased number of false negatives. When the FPR is controlled at 0.05, all mediators will be detected as inactive mediators.

In Supplemental documents (Section 3 and 4), the convergence of the MCMC chain is assessed by examining the trace and density plots of individual parameters, and the MCMC chain has converged. For computational costs, our proposed method is competitive with other Bayesian methods, and our method's computation is affordable for high–dimensional setting $p = 1000$ (3.936 seconds). Overall, our simulation results show that the BHMM achieves highest power (greater than 0.6), and our proposed BHMM yields greater power than other alternative methods when there is correlation between mediators.

5. BHMM Application to IEMS

We apply our proposed method to the same IEMS dataset as that in our previous papers [46–48], which provide more detailed information on data collection and data description. Our interest is to investigate whether the DNA methylation (DNAm) would mediate the association between maternal exposure to extreme drought and child body weight, as well as the association between maternal exposure to hotter subregion and child tibial length. In the analysis of [46, 48], they first carried out a screening step to obtain candidate mediators by running mediator–exposure regression for each mediator and selecting mediators whose coefficients of exposure has p–values

less than 0.05 after adjustment for multiple testing. The authors then performed UMA on these candidate mediators to identify which CpG mediates the relationship between the maternal exposure to climate extremes and child growth.

For the relationship between maternal exposure to drought and child body weight, Straight *et al.*[48] found that among 16 candidate CpGs, cg03771070 at gene *AKAP7* is identified as an active mediator, which is involved in insulin secretion and cardiac function, among other functions [3, 7]. In our analysis, our outcome is child body weight z score, the exposure variable is drought indicator (1=drought exposed, 0= unexposed), and we consider the following confounding covariates as in [48]: age, sex, two cell-type proportions (Epithelial (Epi) and Fibroblast (Fib)), and three stress variables of forced work (husbands or male kin forcing women to work too hard during pregnancy), denied food (denying them food during pregnancy), and mother's lifetime maternal trauma. The estimated unadjusted Intraclass Correlation Coefficient (ICC) suggests that mother accounts for 19% of the variance of child body weight [30]. We then apply the proposed BHMM to analyze the data with a natural choice of 0.5 as PIP threshold [4, 36], and we also detect cg03771070 as an active mediator. The estimated ANIE is -0.29 with the 95% highest posterior density (HPD) interval $(-0.48, -0.11)$, and estimated ANDE is -0.22 with 95% HPD interval $(-0.64, 0.19)$, resulting in 57% of the total effect effect of maternal exposure to drought on the child body weight that can be explained by DNAm. Figure 5 is an illustration of mediation between maternal exposure to drought and child body weight through cg03771070. Figure S2 (as depicted in Section 6 of the Supplemental materials) shows the trace and density plots of the posterior draws for the natural indirect effect via cg03771070, and it indicates that the MCMC chain has converged.

For the relationship between maternal exposure to hotter subregion and offspring tibial length, the outcome is child tibial length z score, an exposure variable is subregion (1=hotter subregion, 0=cooler subregion), and using the same confounders as in the drought model. We found 33% variance of child tibial length outcome were explained by mother. Among 639 candidate mediators, we found six CpGs as active mediators: cg10928038, cg19699973, cg22882310, cg23990814 are hypermethylated between hotter versus cooler subregion; and another two cg08290892, cg13735602 are

hypomethylated. Note that cg23990814 is also detected as significant mediator in [47]. The estimated ANIE is -0.32 with 95% HPD $(-0.54, -0.12)$, while the estimated ANDE is 0.80 with 95% HPD $(0.35, 1.24)$ for ANDE. The estimated indirect effect and direct effect have the opposite sign, indicating the inconsistent mediation where the total effect of maternal exposure to hotter subregion on child tibial length is suppressed by the DNAm patterns [31]. Figure 6 is an illustration of mediation between maternal exposure to hotter subregion and child tibial length through six active CpGs. Table 2 presents estimated ANIE of each mediator with 95% HPD and nearest gene of the identified CpGs and the corresponding biological interpretation of each active CpG mediator. Figure S3 (as depicted in Section 6 of the Supplemental materials) shows the convergence of the MCMC chains for the natural indirect effect mediated by active CpG sites. In addition, we compute the mean absolute percent error (MAPE) to measure the model assessment performance of BHMM and UMA. The results are summarized in Section 5 of the Supplemental materials. In our previous work using UMA [47], we identified eight significant mediators impacting the relationship between subregion and child tibial length Z score. The MAPE values range from 7.69 to 49.55 for the mediator models, and vary from 125.13 to 182.41 for the outcome models. With the BHMM, we detected six significant mediators, and the MAPE is 35.16 for the mediator model and 152.31 for the outcome model.

6. Discussion

In this paper, we develop a Bayesian causal mediation analysis for multilevel data with high–dimensional mediators. We demonstrate through simulation study that the proposed BHMM outperforms other alternative methods in various scenarios, especially for correlated mediators. We also applied our methods to IEMS dataset, we found that cg03771070 at gene *AKAP7* mediates the association between in utero exposure to severe drought and offspring body weight, and six CpG sites mediate the association between hotter versus cooler climate and child tibial growth.

The advantages of our proposed BHMM are: (1) *Model simplicity and efficiency.* Multilevel data is easily modeled under the Bayesian hierarchical framework, which si-

multaneously incorporates both individual-level and group-level models. It is more efficient in inference for parameters by compromising between complete pooling all groups and no–pooling. Moreover, since hierarchical modeling combines information from multilevel variations, it is feasible to use all the data to perform inference for groups with small sample size; (2) *Estimation simplicity.* Utilization of MCMC makes it possible to obtain the estimation of parameters in complex multilevel models. Through Gibbs sampling method with conjugate prior distributions, one can derive posterior distribution of parameters based on fully conditional distribution given data and other parameters. (3) *Shrinkage efficiency.* When the number of mediators is greater than sample size, and under the assumption that only a small proportion of the mediators are active, the Bayesian spike–and–slab prior approach can efficiently capture the sparsity and shrink the inactive ones towards zero, thus it gains more power than UMA in the sense of detecting active mediators. (4) *Joint indirect effect.* Our proposed method simultaneously analyzes the multiple mediators, allowing one to account for the correlation among the mediators, and makes it possible to examine the joint direct effect of exposure explained by selected active mediators without making any path-specific or ordering assumptions on mediators. Compared to the frequentist approach, which is based on the asymptotic properties of the data, the main advantage of our hierarchical Bayesian framework is that it allows inferences in each cluster to be driven by all the data rather than only the data in that particular cluster. Therefore, each cluster helps to increase the precision of the estimates of the other clusters and of the overall population. Moreover, our Bayesian approach can integrate information and prior knowledge available at different scales and provide a flexible Bayesian model to explicitly quantify the modeling uncertainty of the outcome, which accounts for smaller sample sizes and complex structures such as multilevel data [35].

Although our method can jointly analyze high dimensional mediators in the multilevel data setting, indirect effect is assessed at one level rather than multiple levels. One may hope to extend our method to more complex multilevel setting so that the mediation effect could decomposed into upper/group level mediation effect, and lower/individual level mediation effect. With more random effect introduced to the model, the number of parameters needed to be estimated will significantly increase,

so will the computation cost. Future development of new algorithms/models are necessary to effectively characterize the hierarchy and sparsity of the mediation effect, efficiently identify the active mediators and estimate the regression coefficient when sample size is much lower than the number of parameters.

The proposed high-dimensional multilevel mediation analysis relies on the counterfactual framework of mediation, and estimates natural direct and indirect effect under the sequential ignorability assumptions. Our future work will develop sensitivity analyses to quantify the degree to which violation of the assumption would change the results. One possible approach is to perform sensitivity analysis of the correlation between the residual of mediator model and outcome model. Under the sequential ignorability, the correlation is expected to be zero, thus the magnitude of this correlation denotes the departure from ignorability assumption [19].

7. Software

The simulation studies and data analysis were carried out using R version 4.0.2. The R code seamlessly integrated by C++ using R package *Rcpp* is available on *Github* (`https://github.com/XiQiao2023/BHMM`).

Acknowledgements

The authors thank the other investigators Charles Owuor Olungah, Claudia Lalancette and The Epigenomics Core at The University of Michigan for their contribution in the data collection and processing. We are also grateful for all the students and the staff for their lab and filed work, and the Samburu participants in the IEMS study.

Ethics Statement

All data collection and analysis methods conformed to the principles stated in the Declaration of Helsinki and were approved by Western Michigan University Human Subjects Institutional Review Board [Protocol #17-05-09] and Kenya's National Commission for Science, Technology & Innovation. All recruitment and informed consent

materials were translated and back translated by a multilingual team that included Samburu community partners. The study was explained in the Samburu vernacular at community meetings and to parents and child participants at each data collection visit, with consent (and assent for child participants) obtained at each visit.

References

[1] J.M. Albert, *Mediation analysis via potential outcomes models*, Statistics in medicine 27 (2008), pp. 1282–1304.

[2] J.M. Albert and S. Nelson, *Generalized causal mediation analysis*, Biometrics 67 (2011), pp. 1028–1038.

[3] A. Asthana, C. Gaughan, B. Dong, S.R. Weiss, and R.H. Silverman, *Specificity and mechanism of coronavirus, rotavirus, and mammalian two-histidine phosphoesterases that antagonize antiviral innate immunity*, mBio 12 (2021), pp. e01781–21.

[4] M.M. Barbieri and J.O. Berger, *Optimal predictive model selection*, The Annals of Statistics 32 (2004), pp. 870 – 897.

[5] R. Baron and D. Kenny, *The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations*, Journal of Personality and Social Psychology 51 (1986), pp. 1173–1182.

[6] M.G. Blum, L. Valeri, O. François, S. Cadiou, V. Siroux, J. Lepeule, and R. Slama, *Challenges raised by mediation analysis in a high-dimension setting*, Environmental Health Perspectives 128 (2020), p. 055001.

[7] G.K. Carnegie, C.K. Means, and J.D. Scott, *A-kinase anchoring proteins: From protein complexes to physiology and disease*, IUBMB Life 61 (2009), pp. 394–406.

[8] C.M. Carvalho, N.G. Polson, and J.G. Scott, *The horseshoe estimator for sparse signals*, Biometrika 97 (2010), pp. 465–480.

[9] S. Chib and B.P. Carlin, *On mcmc sampling in hierarchical longitudinal models*, Statistics and Computing 9 (1999), pp. 17–26.

[10] R.M. Daniel, B.L. De Stavola, S. Cousens, and S. Vansteelandt, *Causal mediation analysis with multiple mediators*, Biometrics 71 (2015), pp. 1–14.

[11] J. Fan and J. Lv, *Sure independence screening for ultra-high dimensional feature space*, J Roy Stat Soc B 70 (2007).

[12] M.T. Flores-Dorantes, Y.E. Díaz-López, and R. Gutiérrez-Aguilar, *Environment and gene*

association with obesity and their impact on neurodegenerative and neurodevelopmental diseases, Frontiers in neuroscience 14 (2020), pp. 863–863.

[13] E.I. George and R.E. McCulloch, *Variable selection via gibbs sampling*, Journal of the American Statistical Association 88 (1993), pp. 881–889.

[14] T. Hastie, R. Tibshirani, J.H. Friedman, and J.H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer, 2009.

[15] P.J. Heagerty and S.L. Zeger, *Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors)*, Statistical Science 15 (2000), pp. 1–26.

[16] N. Heard-Costa, M. Zillikens, K. Monda, A. Johansson, T. Harris, M. Fu, T. Haritunians, M. Feitosa, T. Aspelund, G. Eiriksdottir, M. Garcia, L. Launer, A. Smith, B. Mitchell, P. McArdle, A. Shuldiner, S. Bielinski, E. Boerwinkle, F. Brancati, E. Demerath, J. Pankow, A. Arnold, Y. Chen, N. Glazer, B. McKnight, B. Psaty, J. Rotter, N. Amin, H. Campbell, U. Gyllensten, C. Pattaro, P. Pramstaller, I. Rudan, M. Struchalin, V. Vitart, X. Gao, A. Kraja, M. Province, Q. Zhang, L. Atwood, J. Dupuis, J. Hirschhorn, C. Jaquish, C. O'Donnell, R. Vasan, C. White, Y. Aulchenko, K. Estrada Gil, B. Hofman, F. Rivadeneira, A. Uitterlinden, J. Witteman, B. Oostra, R. Kaplan, V. Gudnason, J. O'Connell, I. Borecki, C. Duijn, L. Cupples, C. Fox, and K. North, *Nrxn3 is a novel locus for waist circumference: A genome-wide association study from the charge consortium*, PLoS genetics 5 (2009), pp. e1000539–e1000539.

[17] J.P. Hobert and G. Casella, *The effect of improper priors on gibbs sampling in hierarchical linear mixed models*, Journal of the American Statistical Association 91 (1996), pp. 1461–1473.

[18] K. Imai, L. Keele, D. Tingley, and T. Yamamoto, *Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies*, American Political Science Review 105 (2011), pp. 765–789.

[19] K. Imai, L. Keele, and T. Yamamoto, *Identification, inference and sensitivity analysis for causal mediation effects*, Statistical science 25 (2010), pp. 51–71.

[20] K. Imai and T. Yamamoto, *Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments*, Political Analysis 21 (2013), pp. 141–171.

[21] I. Imoto, I. Sonoda, Y. Yuki, and J. Inazawa, *Identification and characterization of human pknox2, a novel homeobox-containing gene*, Biochemical and biophysical research communications 287 (2001), pp. 270–276.

[22] H. Ishwaran and J.S. Rao, *Spike and slab variable selection: frequentist and bayesian strategies*, The Annals of Statistics 33 (2005), pp. 730–773.

[23] A. Javanmard and A. Montanari, *Confidence intervals and hypothesis testing for high-dimensional regression*, J. Mach. Learn. Res. 15 (2014), p. 2869–2909.

[24] S.T. Johnson, Y. Chu, J. Liu, and D.R. Corey, *Impact of scaffolding protein tnrc6 paralogs on gene expression and splicing*, RNA (Cambridge) 27 (2021), pp. 1004–1016.

[25] D. Kenny, D. Kash, and N. Bolger, *Data analysis in social psychology*, The handbook of social psychology 1 (1998), pp. 233–265.

[26] D. Kenny, J. Korchmaros, and N. Bolger, *Lower level mediation in multilevel models*, Psychological methods 8 (2003), pp. 115–28.

[27] J.L. Krull and D.P. Mackinnon, *Multilevel mediation modeling in group-based intervention studies*, Evaluation review 23 (1999), pp. 418–444.

[28] J.L. Krull and D.P. MacKinnon, *Multilevel modeling of individual and group level mediated effects*, Multivariate behavioral research 36 (2001), pp. 249–277.

[29] Y. Lee and J.A. Nelder, *Conditional and marginal models: another view*, Statistical Science 19 (2004), pp. 219 – 238.

[30] D. Liljequist, B. Elfving, and K. Skavberg Roaldsen, *Intraclass correlation – a discussion and demonstration of basic features*, PLOS ONE 14 (2019), pp. 1–35.

[31] D.P. MacKinnon, J.L. Krull, and C.M. Lockwood, *Equivalence of the mediation, confounding and suppression effect*, Prevention science 1 (2000), pp. 173–181.

[32] D.P. MacKinnon and J.H. Dwyer, *Estimating mediated effects in prevention studies*, Evaluation review 17 (1993), pp. 144–158.

[33] D.P. MacKinnon, C.M. Lockwood, J.M. Hoffman, S.G. West, and V. Sheets, *A comparison of methods to test mediation and other intervening variable effects.*, Psychological methods 7 (2002), p. 83.

[34] G. Malsiner-Walli and H. Wagner, *Comparing spike and slab priors for bayesian variable selection*, Austrian Journal of Statistics 40 (2018), pp. 241–264.

[35] M. Miočević, D.P. MacKinnon, and R. Levy, *Power in bayesian mediation analysis for small sample research*, Structural equation modeling: a multidisciplinary journal 24 (2017), pp. 666–683.

[36] J.S. Morris, P.J. Brown, R.C. Herrick, K.A. Baggerly, and K.R. Coombes, *Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models*, Biometrics 64 (2008), pp. 479–489.

[37] J.M. Neuhaus, J.D. Kalbfleisch, and W.W. Hauck, *A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data*, International Statistical Review/Revue Internationale de Statistique (1991), pp. 25–35.

[38] J. Pearl, *Interpretable conditions for identifying direct and indirect effects*, Tech. Rep., California Univ Los Angeles Dept of Computer Science, 2012.

[39] J. Pearl, *Direct and indirect effects*, CoRR abs/1301.2300 (2013).

[40] J. Pearl, *Direct and indirect effects*, in *Probabilistic and Causal Inference: The Works of Judea Pearl*, 2022, pp. 373–392.

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

[42] J. Rahajeng, S.S.P. Giridharan, B. Cai, N. Naslavsky, and S. Caplan, *Important relationships between rab and mical proteins in endocytic trafficking*, World journal of biological chemistry 1 (2010), pp. 254–264.

[43] J.M. Robins and S. Greenland, *Identifiability and exchangeability for direct and indirect effects*, Epidemiology (1992), pp. 143–155.

[44] Y. Song, X. Zhou, J. Kang, M.T. Aung, M. Zhang, W. Zhao, B.L. Needham, S.L.R. Kardia, Y. Liu, J.D. Meeker, J.A. Smith, and B. Mukherjee, *Bayesian hierarchical models for high-dimensional mediation analysis with coordinated selection of correlated mediators*, Statistics in Medicine 40 (2021), pp. 6038–6056.

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

[46] B. Straight, C.E. Hilton, A. Naugle, C.O. Olungah, D. Ngo, X. Qiao, and B.L. Needham, *Drought, psychosocial stress, and ecogeographical patterning: Tibial growth and body shape in samburu (kenyan) pastoralist children*, American Journal of Biological Anthropology 178 (2022), pp. 574–592.

[47] B. Straight, C.E. Hilton, A. Naugle, C.O. Olungah, D. Ngo, X. Qiao, and B.L. Needham, *Dna methylation as a mediator of the association between maternal exposure to regional climate variation and child growth and adiposity*, 2023. Unpublished manuscript.

[48] B. Straight, X. Qiao, D. Ngo, C.E. Hilton, C.O. Olungah, A. Naugle, C. Lalancette, and B.L. Needham, *Epigenetic mechanisms underlying the association between maternal climate stress and child growth: characterizing severe drought and its impact on a kenyan*

community engaging in a climate change-sensitive livelihood, Epigenetics (2022), pp. 1–13.

[49] T.R. Ten Have and M.M. Joffe, *A review of causal estimation of effects in mediation analyses*, Statistical Methods in Medical Research 21 (2012), pp. 77–107.

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

[51] P. Uysal-Onganer and R.M. Kypta, *Wnt11 in 2011 - the regulation and function of a non-canonical wnt*, Acta Physiologica 204 (2012), pp. 52–64.

[52] T. VanderWeele and S. Vansteelandt, *Mediation analysis with multiple mediators*, Epidemiologic methods 2 (2014), pp. 95–115.

[53] T.J. VanderWeele, *Invited commentary: structural equation models and epidemiologic analysis*, American journal of epidemiology 176 (2012), pp. 608–612.

[54] T.J. VanderWeele and S. Vansteelandt, *Conceptual issues concerning mediation, interventions and composition*, Statistics and its Interface 2 (2009), pp. 457–468.

[55] X. Xu and M. Ghosh, *Bayesian variable selection and estimation for group lasso*, Bayesian Analysis 10 (2015).

[56] Y. Yuan and D.P. Mackinnon, *Bayesian mediation analysis*, Psychological Methods 14 (2009), pp. 301–322.

[57] H. Zhang, Y. Zheng, Z. Zhang, T. Gao, B. Joyce, G. Yoon, W. Zhang, J. Schwartz, A. Just, E. Colicino, P. Vokonas, L. Zhao, J. Lv, A. Baccarelli, L. Hou, and L. Liu, *Estimating and testing high-dimensional mediation effects in epigenetic studies*, Bioinformatics 32 (2016), pp. 3150–3154.

[58] Y. Zhao and X. Luo, *Pathway lasso: Estimate and select sparse mediation pathways with high dimensional mediators*, arXiv: Machine Learning (2016).

[59] X. Zhou, P. Carbonetto, and M. Stephens, *Polygenic modeling with bayesian sparse linear mixed models*, PLoS genetics 9 (2013), pp. e1003264–e1003264.

Table 1. Classification of mediators identification.

	True Active Mediators	True Inactive Mediators
Identified Active Mediators	True Positive (TP)[a]	False Positive (FP)[b]
Identified Inactive Mediators	False Negative (FN)[c]	True Negative (TN)[d]

[a]The TP mediators are the ones truly mediating the association between exposure and outcome and are successfully identified as active by our model. [b]The FP mediators are defined as the ones with true mediation effect of zero and are incorrectly identified as active. [c]The FN mediators are the ones truly mediating the association between exposure and outcome but are incorrectly identified as inactive mediators. [d]The TN mediators are the ones with true mediation effect of zero and are successfully identified as inactive mediators.

Table 2. Active mediators between maternal exposure to hotter/cooler subregion and child tibial length.

IlmnID	Estimate	95% HPD	Nearest Gene	Biological Function
cg08290892	-0.07	(-0.15, -0.01)	WNT11	WNT11 (Wingless-Type MMTV Integration Site Family, Member 11) is a highly conserved gene and member of the secreted signaling protein encoding WNT family. It is thought to play a role in the development of the skeleton, as well as the kidney, heart, and lung [51].
cg10928038	-0.04	(-0.13, -0.01)	PKNOX2	PKNOX2 (Knotted 1 Homeobox 2 Protein) is a highly conserved gene and a member of the three-amino-acid loop extension (TALE). It is thought to serve as a nuclear transcription factor (regulating other genes) [21].
cg13735602	-0.12	(-0.23, -0.02)	Not Available	Not Available
cg19699973	0.05	(0.01, 0.11)	TNRC6C	TNRC6C (Trinucleotide Repeat-Containing Gene 6C Protein) is a scaffolding protein involved in miRNA-mediated gene silencing. It is thought to play a substantial role in gene expression but more research is needed [24].
cg22882310	-0.05	(-0.12, -0.01)	NRXN3	NRXN3 (Neurexin 3) plays a role in nervous system function and has been linked to body mass index, waist circumference, and obesity in genome-wide association studies [12, 16].
cg23990814[a]	-0.11	(-0.21,-0.02)	MICALL2	MICAL-like protein 2 is involved in cellular processes, including actin cytoskeleton organization [42].

[a]cg23990814 is also identified as significant mediator between subregion of residence and tibial length in [47]

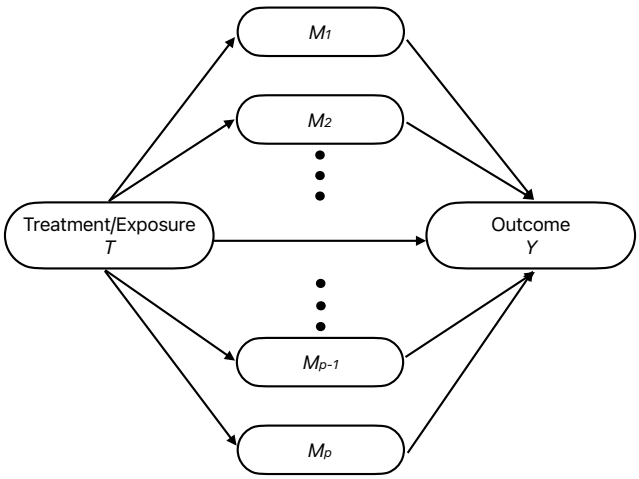

Figure 1. An illustration of high–dimensional mediation analysis.

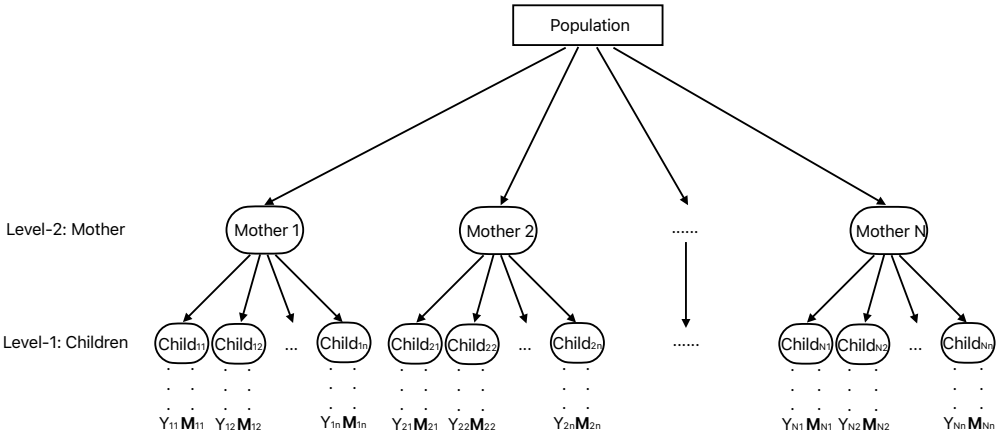

Figure 2. An illustration of multilevel data.

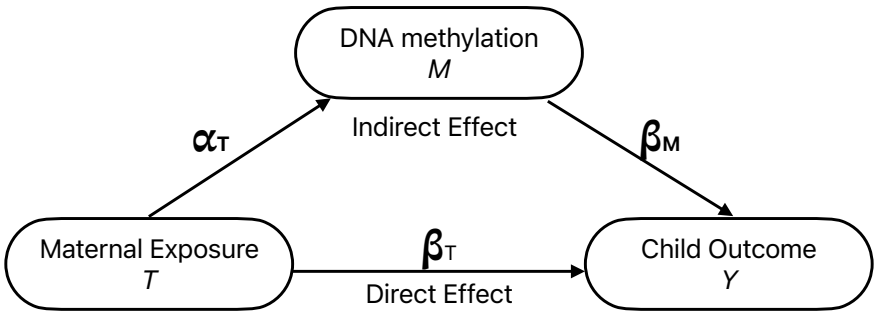

Figure 3. An illustration of single mediation model.

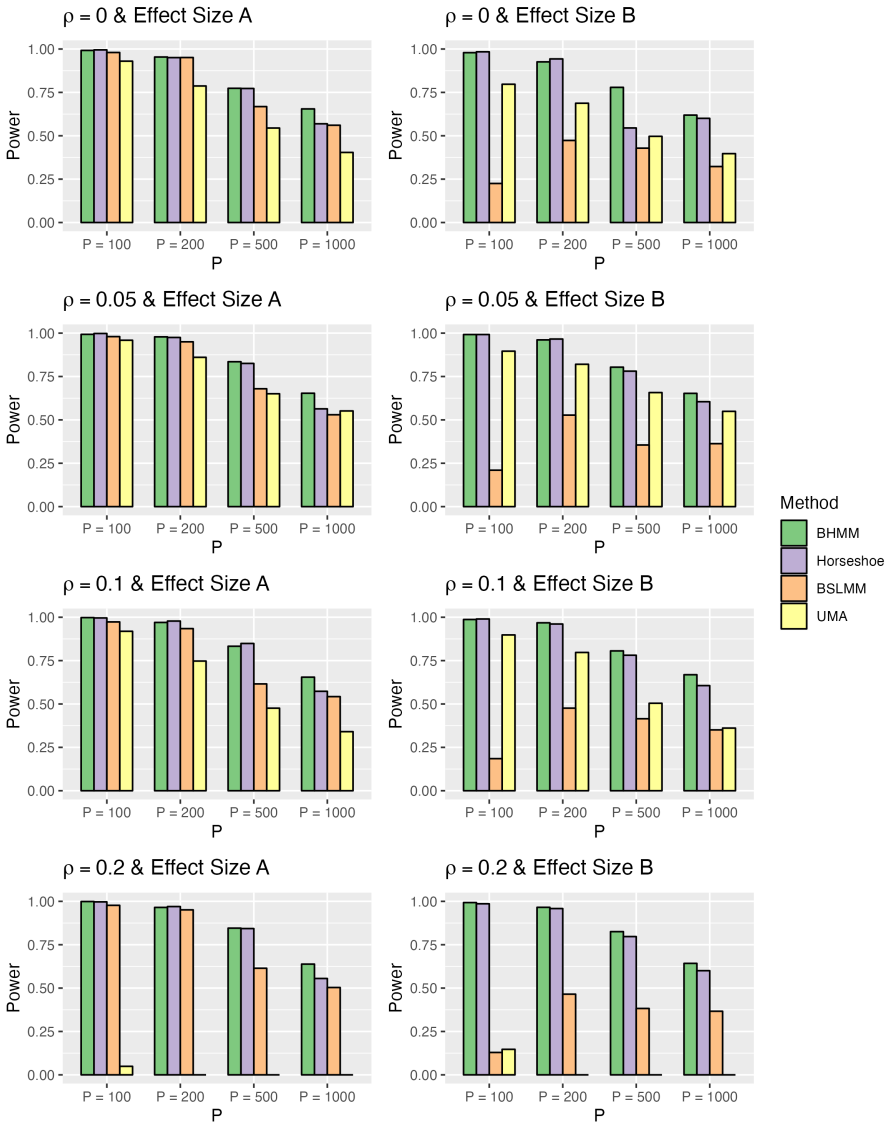

Figure 4. Power comparison of four methods BHMM, Horseshoe, BSLMM, UMA. The power are summarized over 100 simulated data sets at false positive rate (FPR) of 0.05.

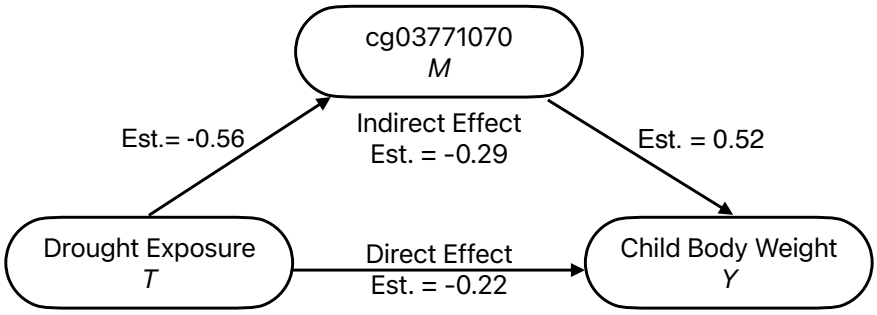

Figure 5. An illustration of estimated indirect effect and direct effect of maternal exposure to drought on child body weight through cg03771070.

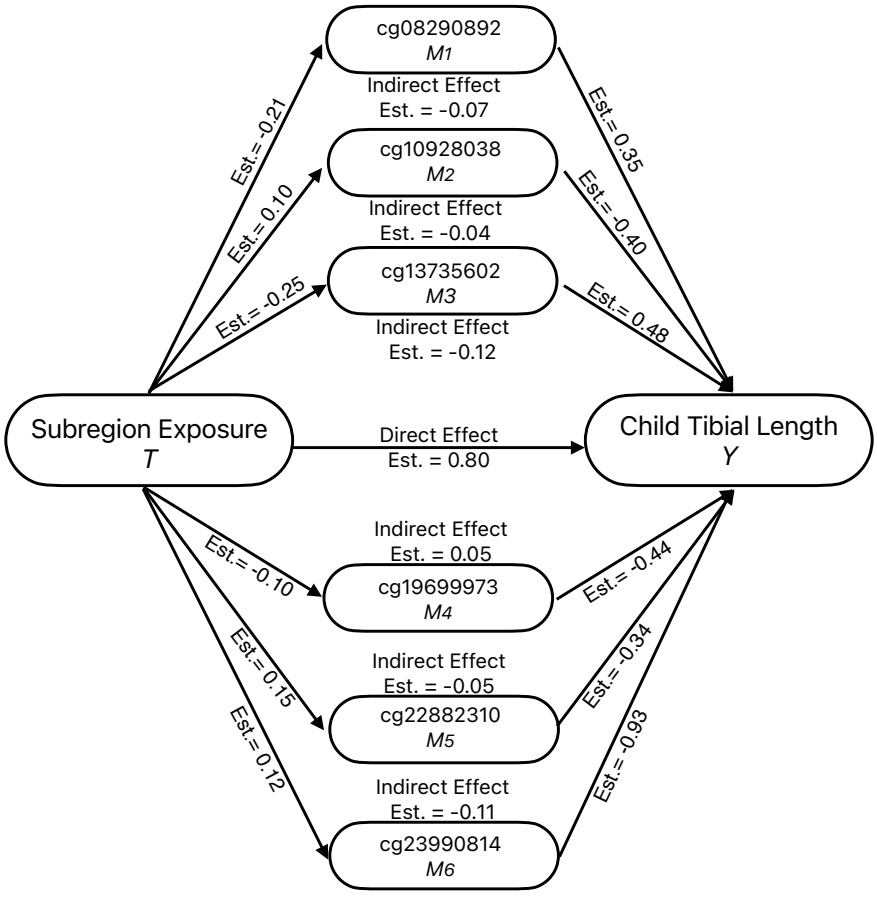

Figure 6. An illustration of estimated indirect effect and direct effect of maternal exposure to hotter/cooler subregion on child tibial length through six identified mediators.