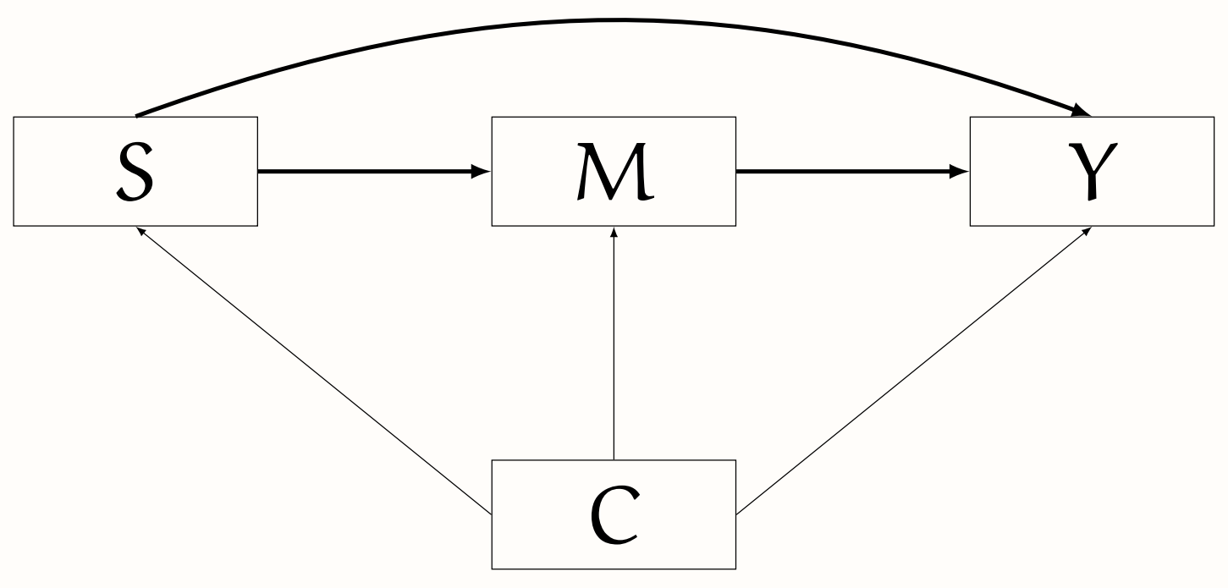
高維中介分析與其在表徵遺傳研究中的應用
On Aspects of High Dimensional Mediation Analysis with Applications to Epigenetic Studies

Students: Chin-Chun Yeh & Jing-Zhong Wang / Supervisors: Rajarshi Mukherjee, PhD & Wan-Chen Lee, ScD

Internship Department: Department of Biostatistics, Harvard T.H. Chan School of Public Health

Introduction

Mediation analysis plays an important role in biomedical and epidemiology research studies, especially to understand the mechanism that one variable is related to the others. What we do in mediation analysis is to estimate the effects of exposure variables on outcome variables, potentially mediated some intermediate variables, which are **mediators**. The mediator serves to clarify the nature of the relationship between the independent and dependent variables. The simplest case of mediation analysis with one mediator is the figure below and it has an exposure S, mediator M, confounder C, and outcome Y. The variable M mediates the effect of S on Y, in other words, S cause M then M cause Y.



In the case of high-dimensional mediation mechanism, we consider the following regression equations to assess the mediation effects:

$$E[M|S=s, C=c] = \gamma_0 S + \gamma_1 C + \epsilon_0 \tag{1}$$

$$E[Y|S=s, M=m, C=c] = \alpha_0 M + \alpha_1 S + \alpha_2 C + \epsilon_1 \tag{2}$$

Where the $S \in \mathbb{R}^{n\times p}$ as the exposures, $M \in \mathbb{R}^{n\times q}$ as the mediators, $Y \in \mathbb{R}^n$ as the outcomes. p is the number of exposures and q is the number of mediators. The sample size is recorded as n. α_1 is the parameter relating S and Y via the direct effect . Moreover, $\gamma_0 = (\gamma_{01}, \dots, \gamma_{0q})$ is the parameter vector relating the exposure to the mediator which also called **natural direct effect**, and $\alpha_0 = (\alpha_{01}, \dots, \alpha_{0q})$ is the parameter relating the mediator to the outcome. The **natural indirect effect** is denoted by the path $S \to M \to Y$, and in high dimensional case, natural individual indirect effect is denoted by $(\gamma_{01}\alpha_{01}, \dots, \gamma_{0q}\alpha_{0q})$; the natural global indirect effect is $\gamma_0^T\alpha_0$. Furthermore, $C \in \mathbb{R}^{n\times r}$ are the measured confounders, ϵ_0 and ϵ_1 are residuals.

Methodology

In the previous study[1], PathwayLasso[2] and HIMA[3] are used to distinguish mediators. However, these methods are not designed for statistical inference. Debias Lasso and MIDA are then proposed to overcome this problem.

Debias Lasso[4]

$(X'X)^{-1}$ can be uninvertible when $q \gg n$. The core idea of Debias Lasso is to replace $D(X'X)^{-1}$ by Ω_I, which is defined in equation 4 below.

$$\begin{pmatrix}\hat{\beta}\\ \hat{\alpha}_1\end{pmatrix} = \begin{pmatrix}\hat{\sum}_{SS}^{-1}\hat{\sum}_{SM}\tilde{\alpha}_0\\ \tilde{\alpha}_1\end{pmatrix} + (I_2 \otimes \hat{\sum}_{SS}^{-1})\frac{1}{n}\hat{\Omega}_I X^T(Y - X\tilde{\alpha}),\ X = (M, S) \tag{3}$$

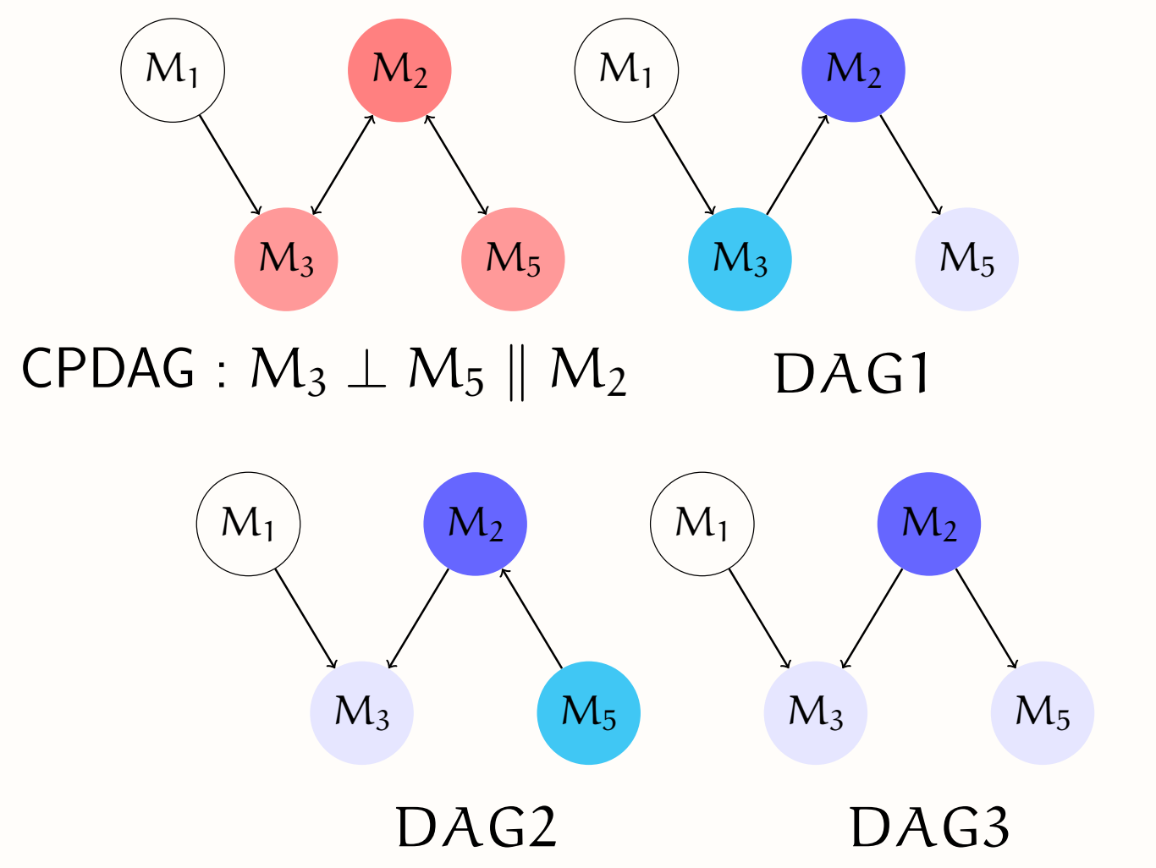
$$\hat{\Omega}_I = \operatorname{argmin}\|\Omega\|_1 \text{ subject to } \|\Omega\hat{\sum}_{XX} - \hat{D}\|_\infty \le \tau_n,\ \hat{D} = \begin{Bmatrix}\hat{\sum}_{SM} & 0\\ 0 & \hat{\sum}_{SS}\end{Bmatrix} \tag{4}$$

MIDA[5]

When estimating individual indirect effect, by only putting exposures, S, the mediator we are concerned, M_j, and the mediators directly affect the mediator,$Pa_{\zeta}(M_j)$ into the equation (2), we can get a more accurate estimate and also overcome the problem in high-dimensional data by reducing the number of variables.

$$\beta_{ij|C} := e^T_{1,|C|+1}((\sum\nolimits_0)_{(i,C)(i,C)})^{-1}(\sum\nolimits_0)_{(i,C)\ j},\ e_{1,|C|+1} = (1, 0^T_{|C|})^T \tag{5}$$

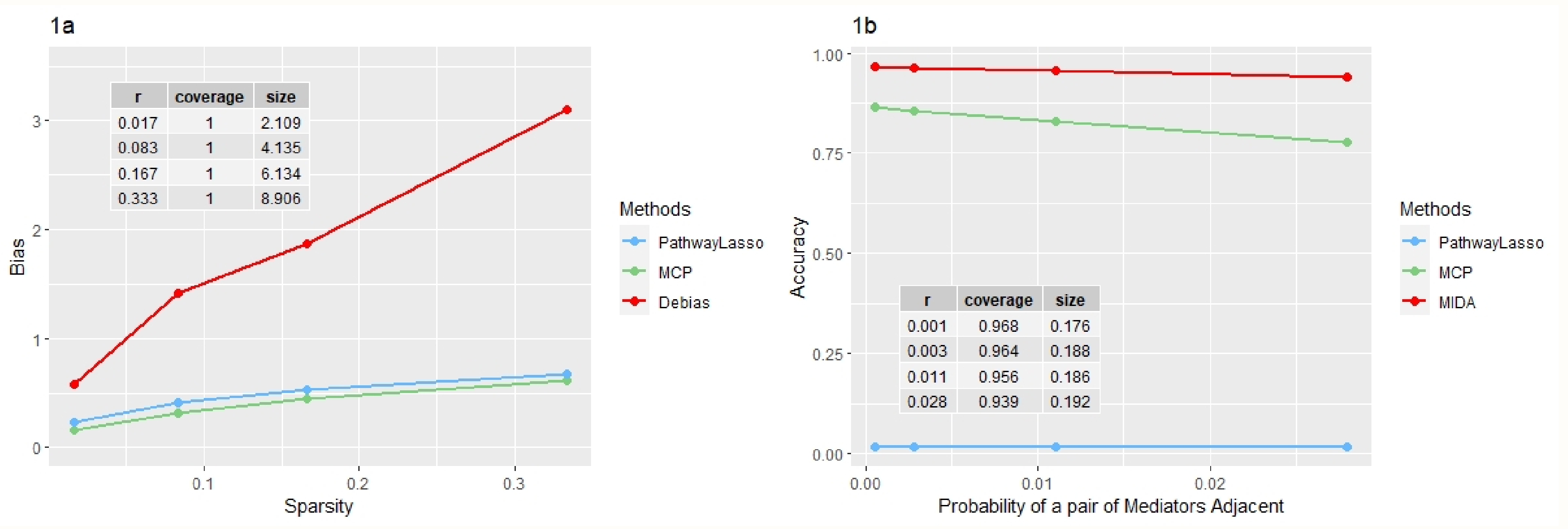
$$NIE_j = \beta_{SM_j}\beta_{M_jY|Pa_\zeta(M_j)\cup S} \tag{6}$$



The mediators which directly affect M_j called the *parent* of M_j. We use the concept of *DAG*, $\hat{\zeta}$, to get the parents of M_j. By estimating *CPDAG*, we can get a group of DAG, which have the same information of conditional independence.

The MIDA algorithm can be breakdown as follow: 1. for $j \in (1, \dots, q)$,obtain the vector of residuals $r_j = (r_j^{(1)}, \dots, r_j^{(n)})$ from the regression of M_j on S. 2.Apply PC algorithm on data $r_1, \dots, r_j$ to obtain an estimate $\hat{\zeta}'$ of the CPDAG. 3.For each $j \in (1, \dots, q)$, obtain a set of possible causal effects $\hat{\Theta}_{M_jY}(\hat{\zeta}') := \{\beta_{M_iY|Pa_\zeta(M_j)\cup S}, \zeta \in MEC(\hat{\zeta}')\}$, $NIE_j = \hat{\beta}_{SM} \times aver(\hat{\Theta}_{M_jY}(\hat{\zeta}'))$. We can get the statistical inference of the quantity,NIE_j.

Simulation



r	coverage	size
0.017	1	2.109
0.083	1	4.135
0.167	1	6.134
0.333	1	8.906

r	coverage	size
0.001	0.968	0.176
0.003	0.964	0.188
0.011	0.956	0.186
0.028	0.939	0.192

Figure 1: **(a): The bias and 95% confident interval coverage probability and length of Debias Lasso given the ratio of true mediators (b): The accuracy of distinguishing the true mediators and 95% confident interval coverage probability and length of MIDA given the probability of a pair of mediators adjacent to each other**

In situation 1 we manipulate the ratio of true mediators to see the performance of Debias Lasso. In situation 2, we study the performance of MIDA while the complexity among the mediators is getting larger. Debias Lasso can always get a confident interval including the true value, and the performance is good when the ratio of the true mediator is small enough. MIDA can actually distinguish the true mediator very well even when the expected number of mediators pairs is large.

Application

In the application, we study how toxicant classes impact the gestational age at delivery, and how diverse toxicological mechanisms mediate the effect. There are 38 toxicants including 4 classes, phthalates class, phenols, and parabens class, polycyclic aromatic hydrocarbons class, and trace metal class. The 61 biomarkers can be divided into 7 groups, including the cyclooxygenase pathway, cytochrome p450 pathway, lipoxygenase pathway, the parent compound, oxidative stress, protein damage, and inflammatory. The outcome is the gestational age at the final visit. There are also 6 covariates including race, maternal age, BMI, specific gravity, private health insurance, education.

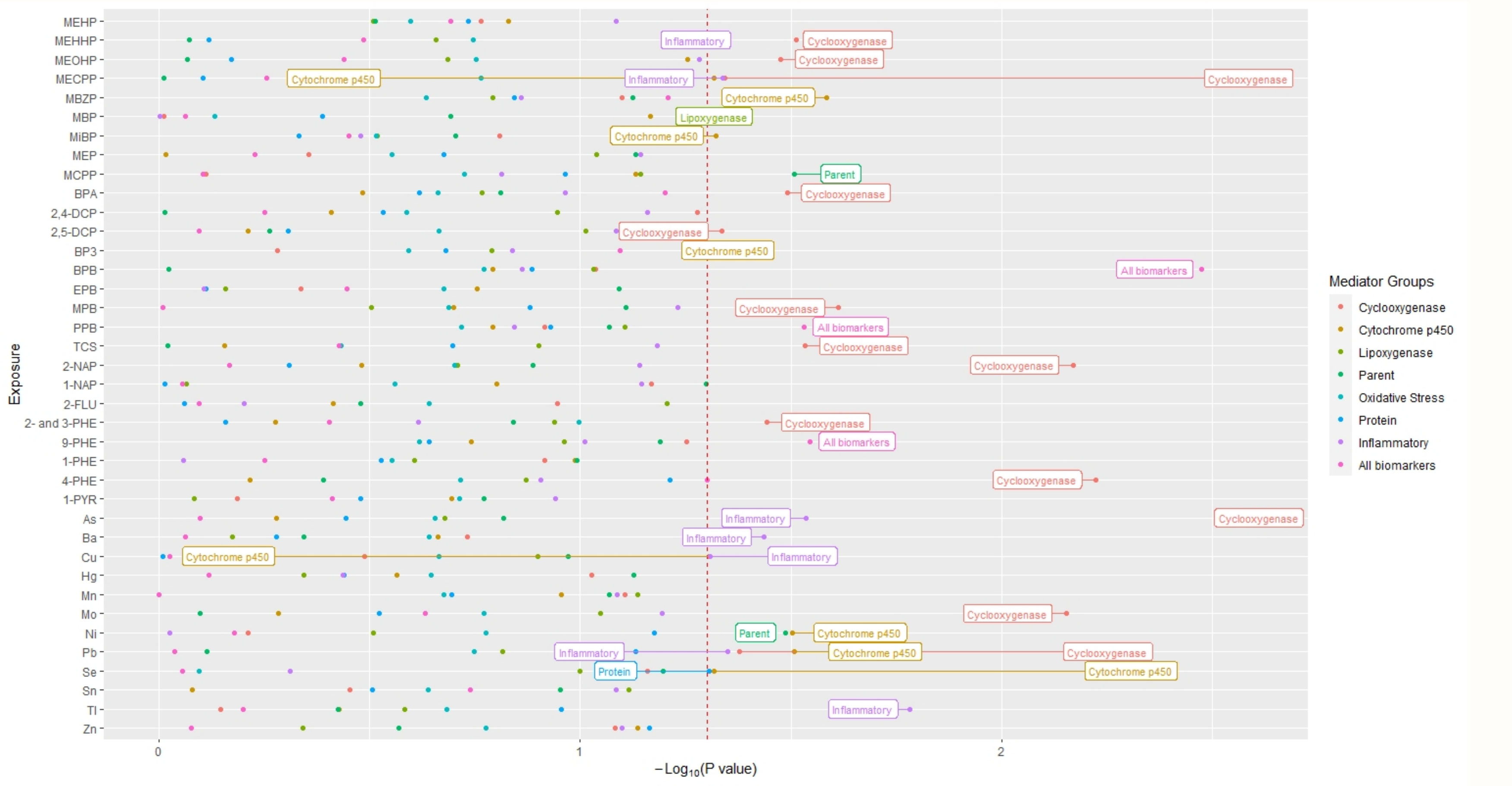


Figure 2: **Estimated $-\log_{10}(P\ value)$ of global indirect effect for gestational age at delivery of all the endogenous groups**

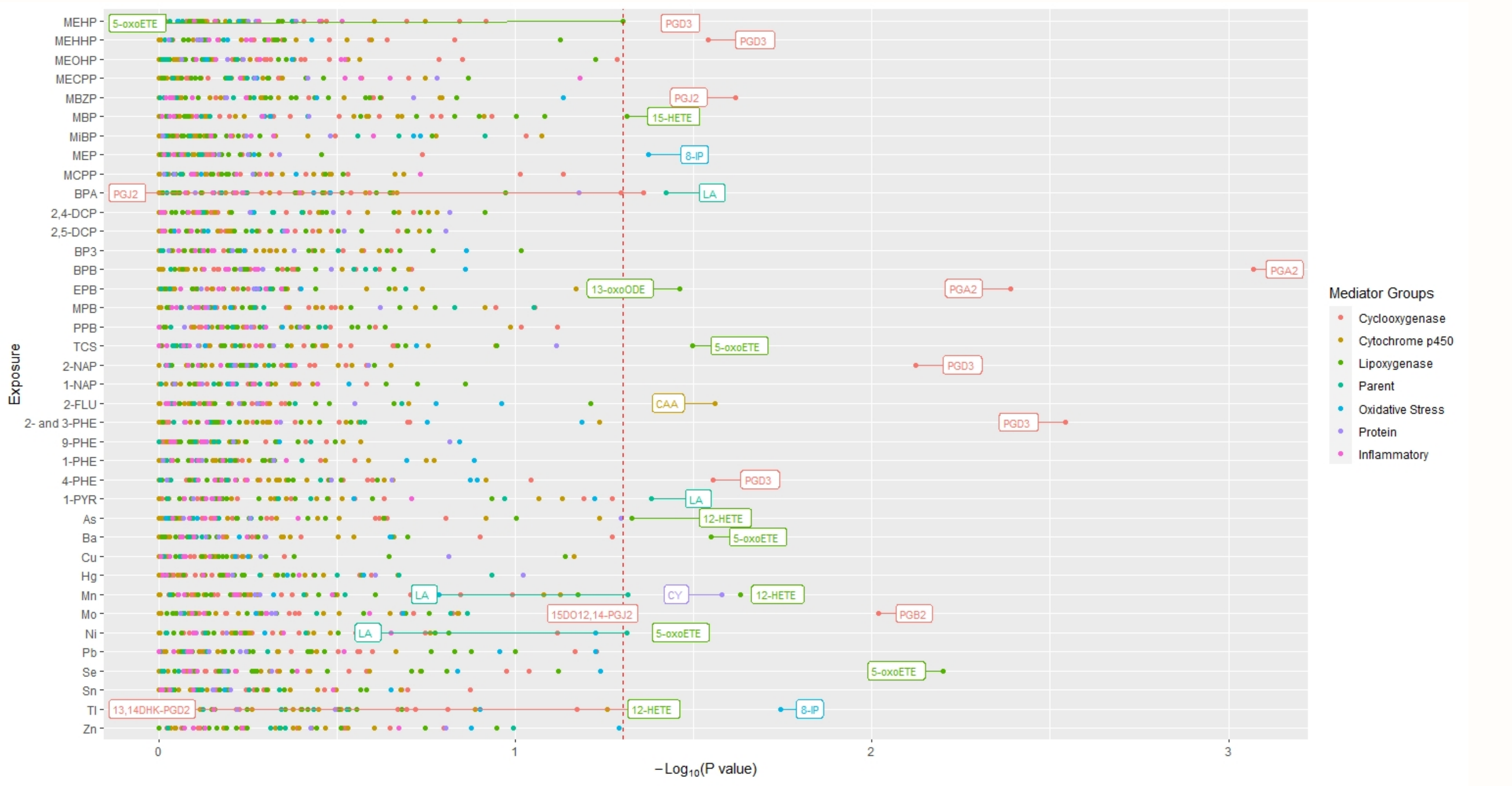


Figure 3: **Estimated $-\log_{10}(P\ value)$ of individual indirect effect for gestational age at delivery of all the biomarkers**

The global indirect effects are mostly driven by cyclooxygenase, cytochrome p450, and inflammatory pathway. The Cyclooxygenase group is the only group that can both drive some of the global effect and some of the individual indirect effect. Most of the mediator groups can only drive one of them. Protein and oxidative stress groups can hardly cause any effect. We infer that most of the endogenous group can only mediate the impact of toxicant by either a whole group or single biomarkers, except the cyclooxygenase group.

Summary and Conclusions

Debiased Lasso and HIMA make the statistic inference in high-dimensional mediation analysis possible, the methods have wide applicability in practice. Several other issues may complicate the estimate of high dimensional mediation effects. e.g. multiple high dimensional exposures, general outcomes, interact effects, different data types variables, or non-linearly structure of the estimation. The methods among these topics remain to be developed to loosen the constrain and assumptions of the data.

References

[1] Aung, M. T. et al., Application of an analytical framework for multivariate mediation analysis of environmental data, *Nature Communications (2020)*

[2] Zhao, Y. & Luo, X., Pathway Lasso: Estimate and Select Sparse Mediation Pathways with High Dimensional Mediators, *(2016)*

[3] Zhang, H. et al., Estimating and testing high-dimensional mediation effects in epigenetic studies, *Bioinformatics 32, 3150–3154 (2016)*

[4] Zhou. R. R. et al., Estimation and inference for the indirect effect in high-dimensional linear mediation models, *Biometrika (2020)*

[5] Chakrabortty, A., Nandy, P. & Li, H., Inference for Individual Mediation Effects and Interventional Effects in Sparse High-Dimensional Causal Graphical Models, *(2021)*

[6] Our presentation slide on Google Drive: `https://reurl.cc/1o0Mm9`