Joint Confounder Selection for Causal Mediation Analysis in High Dimensions

Chanmin Kim and Hyunwoo Kim

Department of Statistics
Sungkyunkwan University

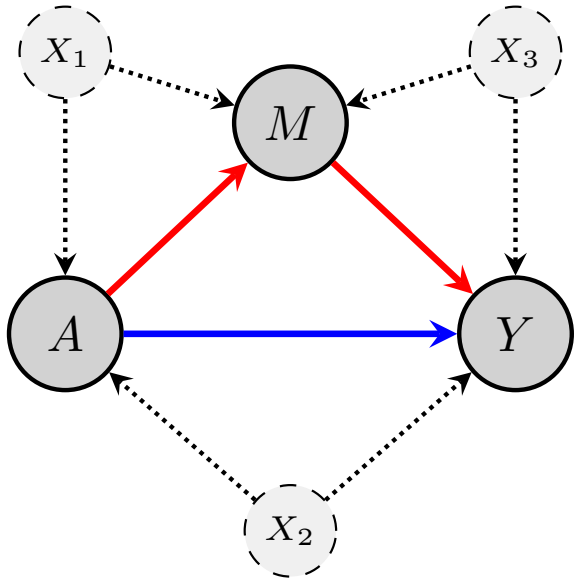
NeurIPS 2026

Mediation Needs Confounding Control on Three Pathways

Mediation analysis decomposes the effect of A on Y:

Direct effect: $A \to Y$ (not through M)

Indirect effect: $A \to M \to Y$



Identification requires adjusting for **three** types of confounders:

① A–M confounders (X_1)

② A–Y confounders (X_2)

③ M–Y confounders (X_3)

Gap

With hundreds or thousands of covariates, hand-picking confounders is infeasible. Data-driven selection exists for a single exposure–outcome pair, but *not for the three-pathway mediation setting.*

This Paper: Four Contributions

1. **A new criterion**: the *mediation disjunctive cause criterion*, designed for the three-pathway confounding structure of mediation analysis.

2. **Identification**: any covariate set satisfying the criterion is a *sufficient adjustment set* under sequential ignorability.

3. **A Bayesian nonparametric model**: Bayesian causal mediation forests with a joint sparsity-inducing prior linking the exposure, mediator, and outcome models (**BCMF-CS**).

4. **Consistency guarantees**: posterior consistency of confounder selection and of direct/indirect effect estimation, *including* $P \gg N$.

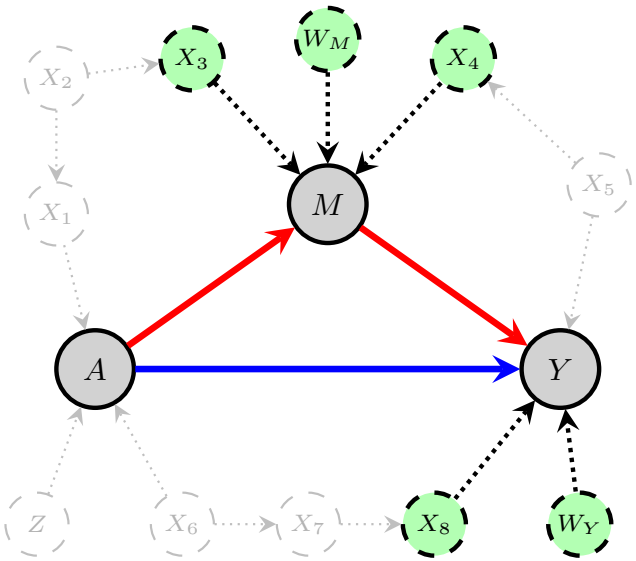
Mediation Disjunctive Cause Criterion

Definition (Mediation Disjunctive Cause Criterion)

A covariate set $\mathcal{C}$ satisfies the criterion if every open backdoor path π between any pair of $\{A, M, Y\}$ is intercepted by a direct cause in $\mathcal{C}$,

$$\mathcal{C} \cap \pi \cap \big(\mathrm{pa}(A) \cup \mathrm{pa}(M) \cup \mathrm{pa}(Y)\big) \neq \emptyset,$$

and $\mathcal{C}$ contains no pure instrument for A.



- ✓ **One direct cause per path** suffices (green): no need to observe the whole chain.
- ✓ Pure instruments (Z) are excluded.

Theorem (Sufficiency)

Any covariate set satisfying the criterion is a **sufficient adjustment set** for the natural direct and indirect effects.

BCMF-CS: Three Forests, One Shared Sparsity Prior

Bayesian **C**ausal **M**ediation **F**orest with **C**onfounder **S**election: BART priors on each function

$$\Pr(A_i = 1) = \Phi\{\mu(\boldsymbol{X}_i)\}, \qquad \text{(Exposure)}$$

$$M_i = \mu_M(\boldsymbol{X}_i, \hat{\pi}_i) + \tau_M(\boldsymbol{X}_i)\, A_i + \epsilon_i, \qquad \text{(Mediator)}$$

$$Y_i = \mu_Y(\boldsymbol{X}_i, \hat{\pi}_i, \hat{m}_{0i}, \hat{m}_{1i}) + \tau_Y(\boldsymbol{X}_i)\, A_i + \tau'_Y(\boldsymbol{X}_i)\, M_i + \eta_i. \qquad \text{(Outcome)}$$

Key idea: one common vector of splitting probabilities

$$\boldsymbol{s} \sim \text{Dir}\left(\tfrac{\alpha}{P+3}, \ldots, \tfrac{\alpha}{P+3}\right) \quad \text{shared by } \mu,\ \mu_M,\ \mu_Y.$$

Effect of the shared prior

Covariates that matter across the exposure, mediator, and outcome models accumulate evidence in $\boldsymbol{s}$; noise variables are shrunk away, and *direct causes* are favored over their correlated ancestors.

Direct effect: $\hat{\delta}(\boldsymbol{x}) = \hat{\tau}_Y(\boldsymbol{x})$; indirect effect: $\hat{\zeta}(\boldsymbol{x}) = \hat{\tau}'_Y(\boldsymbol{x})\, \hat{\tau}_M(\boldsymbol{x})$.

Theoretical Guarantees

Target set: $\mathcal{C}^\star = \mathrm{pa}(M) \cup \mathrm{pa}(Y)$, the direct causes of the mediator and the outcome (assumed observed); $\mathcal{S}$: the selected set.

1. **Selection consistency**: the posterior concentrates on the correct set,

$$P(\mathcal{S} = \mathcal{C}^\star \mid \mathcal{D}) \xrightarrow{P} 1 \quad \text{as } N \to \infty.$$

2. **High dimensions**: the same holds when P_N grows with N, even $P_N \gg N$, provided $\log P_N = o(N)$.

3. **Effect consistency**: for the direct and indirect effects $\Psi = (\delta, \zeta)$,

$$P\big(\|\Psi - \Psi_0\|_2 > \xi \;\big|\; \mathcal{D}\big) \xrightarrow{P} 0 \quad \text{for any } \xi > 0.$$

Simulations

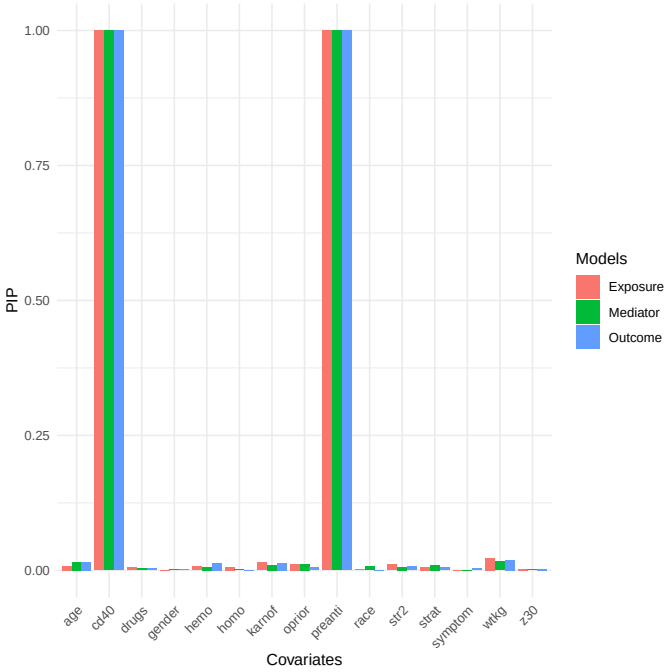
$N = 250$, $P = 100$ to 2000, 200 replicates. Baselines: **BCMF** (no selection), **DCMF** (separate selection in each model), **LSEM** (linear SEM).

	Base (P=100)		Instruments		P=2000 $\gg N$	
Indirect effect	Abs. bias	Coverage	Abs. bias	Coverage	Abs. bias	Coverage
BCMF-CS	**0.25**	**0.90**	**0.49**	**0.59**	**0.43**	**0.62**
BCMF	0.45	0.69	0.84	0.17	1.11	0.06
DCMF	0.34	0.76	0.81	0.22	1.00	0.02
LSEM	1.08	0.39	1.05	0.39	1.06	0.34

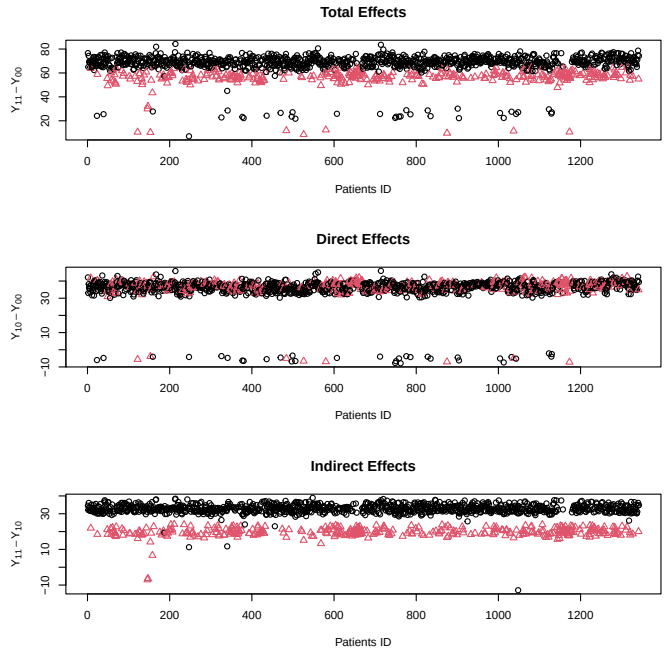
- Best interval coverage in every scenario and the lowest error overall.
- The advantage is largest where selection is hardest: instruments and $P \gg N$.
- With correlated covariates, it recovers *direct causes* rather than their ancestors (PIP ≥ 0.999 vs. < 0.13).

Application: ACTG175 HIV Trial

$N = 1{,}342$ patients (complete cases). A: zidovudine only vs. other regimens; M: CD4 count at 20 weeks; Y: CD4 count at 96 weeks; 15 baseline covariates.



Posterior inclusion probabilities



Individual total / direct / indirect effects

- Only **two** of 15 covariates selected: baseline CD4 (`cd40`) and prior antiretroviral exposure (`preanti`).
- Indirect effect differs by race; direct effect does not.

Takeaways

A complete pipeline for mediation analysis in high dimensions

1. **What to adjust for**: the mediation disjunctive cause criterion covers all three confounding pathways.
2. **How to find it**: BCMF-CS selects confounders jointly across the exposure, mediator, and outcome models through one shared sparse prior.
3. **Why it works**: posterior consistency of selection and of direct/indirect effects, even when $P \gg N$.

Thank you!

Please visit our poster.

chanmin.kim@skku.edu