Supplementary Material: The Blessings of Multiple Causes

A Connections to genome-wide association studies

Many methods from the research literature, especially around genome-wide association studies, can be reinterpreted as instances of the deconfounder algorithm. Each can be seen as positing a factor model of assigned causes (Section 4.1) and a conditional outcome model (Section 4.2).

The deconfounder justifies each of these methods as forms of multiple causal inference and, though predictive checks, points to how a researcher can usefully compare and assess them. Most of these methods were motivated by imagining true unobserved confounding structure. However, the theory around the deconfounder shows that a well-fitted factor model will capture confounders independent of a researcher imagining what they may be; see the question in Section 5.

Below we describe many methods from the GWAS literature and show how they can be viewed as deconfounder algorithms. The GWAS problem is described in Section 4.3.

Linear mixed models. The LMM is one the most popular classes of methods for analyzing GWAS (Yu et al., 2006; Kang et al., 2008; Yang et al., 2014; Lippert et al., 2011; Loh et al., 2015; Darnell et al., 2017). Seen through the lens of the deconfounder, an LMM posits a linear outcome model that depends on both the SNPs and a scalar latent factor Z_i.

In the LMM literature, Z_i is not explicitly drawn from a factor model; rather, $Z_{1:n}$ are from a multivariate Gaussian whose covariance matrix, called the "kinship matrix," is calculated from the observed SNPs $\boldsymbol{a}_{1:n}$. However, this is mathematically equivalent to posterior latent factors from a one-dimensional principal component analysis (PCA) model. Subject to its capturing the distribution of SNPs, the LMM is performing multiple causal inference with a deconfounder.

Principal component analysis. A related approach is to first perform (multi-dimensional) PCA on the SNP matrix and then to estimate an outcome model from the corresponding residuals (Price et al., 2006). This too is an instance of the deconfounder. As a factor model, PCA is described in Eq. 9. Fitting an outcome model to its residuals is equivalent to conditioning on the reconstructed assignments, Eq. 21.

Logistic factor analysis. Closely related to PCA is LFA (Song et al., 2015; Hao et al., 2015). LFA can be seen as the following factor model,

$$\begin{aligned} Z_i &\sim \mathcal{N}(0, I) \\ \pi_{ij} \mid Z_i &\sim \mathcal{N}(z_i^\top \theta_j, \sigma^2), \quad j = 1, \ldots, m, \\ A_{ij} \mid \pi_{ij} &\sim \text{Binomial}(2, \text{logit}^{-1}(\pi_{ij})), \quad j = 1, \ldots, m. \end{aligned}$$

If it captures the SNP matrix well, then Z_i can be viewed as a substitute confounder.

With LFA in hand, Song et al. (2015) use inverse regression to perform association tests. Their approach is equivalent to assuming an outcome model conditional on the reconstructed assignments $a(\hat{z}_i)$, again Eq. 21, and subsequently testing for non-zero coefficients.

In a variant of LFA, Tran and Blei (2017) use a neural-network based model of the unobserved confounder, connecting this model to a causal inference with a nonparametric structural equation model (Pearl, 2009). They take an explicitly causal view of the testing problem.

Mixed-membership models. Finally, many statistical geneticists use mixed-membership models (Airoldi et al., 2014) to capture the latent population structure of SNPs, and then condition on that structure in downstream analyses (Pritchard et al., 2000a,b; Falush et al., 2003, 2007). In genetics, a mixed-membership model is a factor model that captures latent ancestral populations. The latent variable Z_i is on the $K-1$ simplex; it represents how much individual i reflects each ancestral population. The observed SNP A_{ij} comes from a mixture of Binomials, where Z_i determines its mixture proportions.

Using these models, researchers use a linear outcome model conditional on z_i and devise tests for significant associations (Pritchard et al., 2000b; Song et al., 2015; Tran and Blei, 2017). The deconfounder justifies this practice from a causal perspective, and underlines the importance of finding a model of population structure that captures the per-individual distribution of SNPs.

B Can the causes be causally dependent among themselves?

When the causes are causally dependent, the deconfounder can still provide unbiased estimates of the potential outcomes. Its success relies on a valid substitute confounder.

Note there are cases where a valid substitute confounder cannot exist. For example, consider a cause A_1 that causally affects A_2 according to $A_1 \sim \mathcal{N}(0,1), A_2 = A_1 + \epsilon, \epsilon \sim \mathcal{N}(0,1)$. In this case, a substitute confounder Z must satisfy $Z \overset{a.s.}{=} A_1$ or $Z \overset{a.s.}{=} A_2$, because it needs to render the two causes conditionally independent. But such a Z does not satisfy overlap.

On the other hand, causal dependence among the causes does not necessarily imply the nonexistence of a valid substitute confounder. Consider a different mechanism for the causal relationship between A_1 and A_2,

$$\begin{aligned} A_1 &\sim \mathcal{N}(0,1), \\ A_2 &= |A_1| + \epsilon, \quad \epsilon \sim \mathcal{N}(0,1). \end{aligned}$$

Here $Z \overset{a.s}{=} |A_1|$ is a valid substitute confounder; it satisfies overlap and renders A_1 conditionally independent of A_2.

Empirically, it is hard to detect the nonexistence of a valid substitute confounder without knowing the functional form of how the causes are structurally dependent. Insisting on using the deconfounder in this case results in limited overlap and high variance causal estimates downstream.

To illustrate this phenomenon, we repeat the experiments in Section 6.1 with the same confounder a_{age} but three causes: $a_{\text{mar}}, a_{\text{exp}}$ and an additional cause $a_{\text{mar+}}$. We assume $a_{\text{mar+}}$ causally depend on a_{mar}, where

$$a_{\text{mar+}} = a_{\text{mar}} + \varepsilon_{i,\text{mar+}}, \qquad \varepsilon_{i,\text{mar+}} \sim \mathcal{N}(0, 0.1^2). \tag{43}$$

	Check	Bias$^2 \times 10^{-2}$	Variance $\times 10^{-2}$	MSE $\times 10^{-2}$
No control	–	41.89	0.01	41.90
Control for age (oracle)	–	22.57	0.01	22.57
Control for 1-dim z_{line}	✓	29.98	16.97	46.96
Control for 1-dim $a(z_{\text{line}})$	✓	28.01	18.49	46.50
Control for 1-dim z_{quad}	✓	**25.10**	16.70	**41.80**
Control for 1-dim $a(z_{\text{quad}})$	✓	27.46	15.77	43.23

Table 5: Total bias and variance of the estimated causal coefficients β_{exp} and β_{mar} when there is a third cause dependent on a_{mar}. The nonlinear factor model outperforms linear factor model. The deconfounder estimate has much higher variance than usual (e.g., Table 4) when two of the causes are dependent.

It implies that theoretically there exists no substitute confounders that can both satisfy overlap and render the causes conditionally independent.

We simulate the outcome from

$$y_i = \beta_{\text{mar}} a_{\text{mar},i} + \beta_{\text{exp}} a_{\text{exp},i} + \beta_{\text{age}} a_{\text{age},i} + \beta_{\text{mar+}} a_{\text{mar+},i} + \varepsilon_i, \tag{44}$$

where $\varepsilon_i \sim \mathcal{N}(0,1)$. We generate the true causal coefficients from

$$\beta_{\text{mar}} \sim \mathcal{N}(0,1) \quad \beta_{\text{exp}} \sim \mathcal{N}(0,1) \quad \beta_{\text{age}} \sim \mathcal{N}(0,1) \quad \beta_{\text{mar+}} \sim \mathcal{N}(0,1). \tag{45}$$

Nevertheless, we apply the deconfounder to this data. We model the three causes with one-dimensional linear and quadratic factor model; both pass the predictive check, with a predictive score of 0.28 and 0.20. Table 5 shows the bias and variance of the deconfounder estimate of β_{mar} and β_{exp}. With causally dependent causes (Table 5), the deconfounder estimates have much larger variance than usual (Table 4); it signals that the substitute confounder we constructed is close to breaking overlap. That said, the deconfounder is still able to correct for a substantial portion of confounding bias.

Finally, we recommend applying the deconfounder to non-causally dependent causes. A valid substitute confounder is guaranteed to exist in this case; it will both satisfy overlap and render the causes conditionally independent of each other.

C Causal identification with a quadratic factor model and a linear outcome model

We establish causal identification when the true causal model is composed of a quadratic factor model and a linear outcome model.

We first write down the causal model:

$$Z = \epsilon_Z, \tag{46}$$
$$A_1 = \alpha_{10} + \alpha_{11} Z + \alpha_{12} Z^2 + \epsilon_{A1}, \tag{47}$$
$$A_2 = \alpha_{20} + \alpha_{21} Z + \alpha_{22} Z^2 + \epsilon_{A2}, \tag{48}$$
$$Y = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \gamma Z + \epsilon_Y, \tag{49}$$

where all the errors $\epsilon_Z, \epsilon_{A1}, \epsilon_{A2}, \epsilon_Y$ are independent zero-mean Gaussian with a fixed but unknown variance.

We note that all variables Z, A_1, A_2, Y are scalars in this example; only A_1, A_2, Y are observable; Z is unobserved.

To prove identification, we show that the causal parameters β_1 and β_2 are both functions of the moment generating function of (A_1, A_2, Y).

we first rewrite Y:

$$\begin{aligned} Y &= (\beta_0 + \beta_1\alpha_{10} + \beta_2\alpha_{20}) + (\beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma) \cdot Z + (\beta_1\alpha_{12} + \beta_2\alpha_{22}) \cdot Z^2 + \beta_1\epsilon_{A1} + \beta_2\epsilon_{A2} + \epsilon_Y, \\ &= (\beta_1\alpha_{12} + \beta_2\alpha_{22}) \cdot \left(Z + \frac{\beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma}{2 \cdot (\beta_1\alpha_{12} + \beta_2\alpha_{22})} \right)^2 + \beta_1\epsilon_{A1} + \beta_2\epsilon_{A2} + \epsilon_Y \\ &+ \left(\beta_0 + \beta_1\alpha_{10} + \beta_2\alpha_{20} - \left(\frac{\beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma}{2 \cdot (\beta_1\alpha_{12} + \beta_2\alpha_{22})} \right)^2 \right) \end{aligned}$$

In other words, the observed random variable Y is a sum of a constant, a non-central χ^2 random variable and a zero mean Gaussian random variable $\beta_1\epsilon_{A1} + \beta_2\epsilon_{A2} + \epsilon_Y$.

For notation simplicity, we denote the constants with separate symbols:

$$B_0 \triangleq \beta_0 + \beta_1\alpha_{10} + \beta_2\alpha_{20}, \tag{50}$$
$$B_1 \triangleq \beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma, \tag{51}$$
$$B_2 \triangleq \beta_1\alpha_{12} + \beta_2\alpha_{22}. \tag{52}$$

Therefore, we have

$$Y = B_0 + B_1 \cdot Z + B_2 \cdot Z^2 + \epsilon_Y, \tag{53}$$

where $(\frac{Z}{\sigma_Z} + \frac{B_1}{2B_2\sigma_Z})^2$ is a non-central χ^2 random variable with the non-centrality parameter $\lambda = \left(\frac{B_1}{2B_2\sigma_Z}\right)^2$ and degree of freedom $k = 1$. (σ_Z^2 is the variance of Z.)

We leverage this property to identify the distribution ϵ_Y. Notice the moment generating function of A_1, A_2, Y is

$$M_{A_1,A_2,Y}(t_1, t_2, t_3) \tag{54}$$
$$= \mathbb{E}[\exp(t_1A_1 + t_2A_2 + t_3Y)] \tag{55}$$

$$= \exp(B_0 t_3 + \alpha_{10} t_1 + \alpha_{20} t_2) \tag{56}$$

$$\cdot \mathbb{E}\left[\exp\left((\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3) \cdot Z + (\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \cdot Z^2\right)\right] \tag{57}$$

$$\cdot \mathbb{E}\left[t_1 \epsilon_{A1} + t_2 \epsilon_{A2} + t_3(\beta_1 \epsilon_{A1} + \beta_2 \epsilon_{A2} + \epsilon_Y)\right] \tag{58}$$

$$= \exp\left(B_0 t_3 + \alpha_{10} t_1 + \alpha_{20} t_2 - (\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \cdot \left(\frac{\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3)}{2(\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3)}\right)^2\right) \tag{59}$$

$$\cdot \mathbb{E}\left[\exp\left((\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3)\sigma_Z^2 \cdot \left(\frac{\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3}{2(\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3)\sigma_Z} + \frac{Z}{\sigma_Z}\right)^2\right)\right] \tag{60}$$

$$\cdot \mathbb{E}\left[t_1 \epsilon_{A1} + t_2 \epsilon_{A2} + t_3(\beta_1 \epsilon_{A1} + \beta_2 \epsilon_{A2} + \epsilon_Y)\right] \tag{61}$$

$$= \exp\left(B_0 t_3 + \alpha_{10} t_1 + \alpha_{20} t_2 - (\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \cdot \left(\frac{\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3}{2(\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3)}\right)^2\right) \tag{62}$$

$$\cdot \frac{\exp(\frac{\lambda t}{1-2t})}{(1-2t)^{1/2}} \tag{63}$$

$$\cdot \exp(\frac{1}{2}(t_1 + t_3\beta_1)^2 \sigma_{A_1}^2) \exp(\frac{1}{2}(t_2 + t_3\beta_2)^2 \sigma_{A_2}^2) \exp(\frac{1}{2} t_3 \sigma_Y^2) \tag{64}$$

$$= \exp\left(B_0 t_3 + \alpha_{10} t_1 + \alpha_{20} t_2 - (\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \cdot \left(\frac{\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3}{2(\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3)}\right)^2\right) \tag{65}$$

$$\cdot \frac{\exp(\frac{\lambda t}{1-2t})}{(1-2t)^{1/2}} \tag{66}$$

$$\cdot \exp(\frac{1}{2}(t_1 \sigma_{A_1}^2 + \beta_1^2 \sigma_{A_1}^2 t_3^2 + 2\beta_1 \sigma_{A_1}^2 t_1 t_3 + t_2 \sigma_{A_2}^2 + \beta_2^2 \sigma_{A_2}^2 t_3^2 + 2\beta_2 \sigma_{A_2}^2 t_2 t_3 + \sigma_Y^2 t_3), \tag{67}$$

where $t = (\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3)\sigma_Z^2$ and $\lambda = \left(\frac{\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3}{2(\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3)\sigma_Z}\right)^2$.

Notice that the ratio of the coefficients in front of t_3^2 and $t_1 t_3$ is β_1. Hence we can identify β_1 from the moment generating function of the unobserved random variables A_1, A_2, Y. The reason is the incongruence between exponential functions, polynomial functions, and square root functions, i.e. exponential functions can not be written as polynomials and others. The other components of the moment generating functions Eqs. 65 and 66 do not contain the terms t_3^2 and $t_1 t_3$.

The high-level intuition behind the above calculation is the incongruence between the nonlinear (quadratic) factor model and the linear outcome model. More specifically, the variance due to ϵ_Y in the linear outcome model cannot be attributed wrongfully to the causes and the confounder $\beta_1 A_1 + \beta_2 A_2 + \gamma Z$; the former is Gaussian while the latter is non-Gaussian except when $\alpha_{12} = \alpha_{22} = 0$. (This incongruence does not hold for the linear factor model and the linear outcome model.)

For the same reason, we can identify the other causal parameter β_2.

This result can be extended to other nonlinear factor models and linear outcome models.

D Detailed Results of the GWAS Study

In this section, we present tables of results from the GWAS study in Section 6.2.

Tables 6 to 10 contain the result under the high SNR setting.

	Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
No control	—	49.66	39.39
Control for confounders*	—	40.27	31.09
(G)LMM	—	46.22	37.81
PPCA	0.13	46.05	36.01
PF	0.15	44.58	36.30
LFA	0.14	43.02	36.65
GMM	0.01	47.33	40.24
DEF	0.18	**41.05**	**33.88**

Table 6: GWAS high-SNR simulation I: Balding-Nichols Model. ("Control for all confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms (G)LMM; DEF performs the best among the five factor models. Predictive checking offers a good indication of when the deconfounder fails.

Tables 11 to 15 contain the result under the low SNR setting.

E Detailed Results of the Movie Study

In this section, we present tables of results from the movies study in Section 6.3.

F Proof of Lemma 1

Proof sketch. First assume the Kallenberg construction in Eq. 37. This form shows that the assigned causes $(A_{i1},\dots,A_{im})$ are captured by functions of Z_i and randomization variables U_{ij}. This fact, in turn, implies that the randomness in $(A_{i1},\dots,A_{im})|Z_i$ comes from the randomization variables which are (by definition) independent of $Y_i(\boldsymbol{a})$. Therefore $(A_{i1},\dots,A_{im})$ is conditionally independent of Y_i given Z_i, i.e., unconfoundedness holds. Now assume that unconfoundedness holds. We prove that this assumption implies a Kallenberg construction by building on the randomization variable construction of conditional distributions (Kallenberg, 1997). □

Proof. For notation simplicity, we suppress the i subscript in this proof.

We assume $\mathcal{Z}$ is a measurable space and $\mathcal{A}_j, j = 1,\dots,m$ are Borel spaces.

	Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
No control	—	68.78	38.16
Control for confounders*	—	60.29	32.76
(G)LMM	—	65.25	35.41
PPCA	0.15	65.98	36.11
PF	0.17	64.25	34.79
LFA	0.17	64.00	37.08
GMM	0.02	67.23	35.40
DEF	0.20	**63.73**	**33.71**

Table 7: GWAS high-SNR simulation II: 1000 Genomes Project (TGP). ("Control for all confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms (G)LMM; DEF performs the best among the five factor models. Predictive checking offers a good indication of when the deconfounder fails.

We first prove the necessity. Assume that $A_j = f_j(Z, U_j), j = 1, \ldots, m$, where $f_j, j = 1, \ldots, m$ are measurable and

$$(U_1, \ldots, U_m) \perp\!\!\!\perp (Z, Y(a_1, \ldots, a_m)) \tag{68}$$

for all $(a_1, \ldots, a_m)$. By Proposition 5.18 in Kallenberg (1997), Eq. 68 implies

$$(U_1, \ldots, U_m) \perp\!\!\!\perp_Z Y(a_1, \ldots, a_m),$$

and so

$$(Z, U_1, \ldots, U_m) \perp\!\!\!\perp_Z Y(a_1, \ldots, a_m)$$

by Corollary 5.7 in Kallenberg (1997). It implies

$$(A_1, \ldots, A_m) \perp\!\!\!\perp_Z Y(a_1, \ldots, a_m)$$

for all $(a_1, \ldots, a_m) \in \mathscr{A}_1 \otimes \cdots \otimes \mathscr{A}_m$. The last step is because A_j's are measurable functions of $(Z, U_1, \ldots, U_m)$.

Now we prove the sufficiency. Assume that $Y(a_1, \ldots, a_m) \perp\!\!\!\perp_Z (A_1, \ldots, A_m)$. Marginalizing out all but one A_j gives

$$Y(a_1, \ldots, a_m) \perp\!\!\!\perp_Z A_j, j = 1, \ldots, m.$$

By Theorem 5.10 in Kallenberg (1997), there exists a measurable function $f_j : \mathscr{Z} \times [0,1] \to \mathscr{A}_j$ and a Uniform[0,1] random variable $\tilde{U}_j$ satisfying $\tilde{U}_j \perp\!\!\!\perp (Z, Y(a_1, \ldots, a_m))$ such that the random variable $\tilde{A}_j = f_j(Z, \tilde{U}_j)$ satisfies

$$\tilde{A}_j \stackrel{d}{=} A_j \text{ and } (\tilde{A}_j, Z) \stackrel{d}{=} (A_j, Z).$$

Moreover, we have

$$\tilde{A}_j \perp\!\!\!\perp_Z Y(a_1, \ldots, a_m)$$

with the same argument as the above necessity part.

	Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
No control	—	77.35	45.93
Control for confounders*	—	67.53	39.43
(G)LMM	—	74.38	42.79
PPCA	0.14	74.45	43.27
PF	0.14	71.40	42.75
LFA	0.13	72.11	42.34
GMM	0.03	76.27	46.88
DEF	0.16	**69.86**	**41.61**

Table 8: GWAS high-SNR simulation III: Human Genome Diversity Project (HGDP). ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms (G)LMM; DEF performs the best among the five factor models. Predictive checking offers a good indication of when the deconfounder fails.

Hence, by Proposition 5.6 in Kallenberg (1997),

$$P(\tilde{A}_j \in \cdot \mid Z, Y(a_1,\ldots,a_m)) = P(\tilde{A}_j \in \cdot \mid Z) = P(A_j \in \cdot \mid Z) = P(A_j \in \cdot \mid Z, Y(a_1,\ldots,a_m)),$$

and so

$$(\tilde{A}_j, Z, Y(a_1,\ldots,a_m)) \stackrel{d}{=} (A_j, Z, Y(a_1,\ldots,a_m)).$$

By Theorem 5.10 in Kallenberg (1997), we may choose some random variable U_j such that

$$U_j \stackrel{d}{=} \tilde{U}_j \text{ and } (\tilde{A}_j, Z, Y(a_1,\ldots,a_m), U_j) \stackrel{d}{=} (A_j, Z, Y(a_1,\ldots,a_m), \tilde{U}_j).$$

In particular, we have

$$U_j \perp\!\!\!\perp (Z, Y(a_1,\ldots,a_m))$$

and

$$(A_j, f_j(Z, U_j)) \stackrel{d}{=} (\tilde{A}_j, f_j(Z, \tilde{U}_j).$$

Since

$$\tilde{A}_j = f_j(Z, \tilde{U}_j)$$

and the diagonal in S^2 is measurable, we have

$$A_j \stackrel{a.s.}{=} f_j(Z, U_j).$$

We then show $(U_1,\ldots,U_m) \perp\!\!\!\perp (Z, Y(a_1,\ldots,a_m))$. By Theorem 5.10 in Kallenberg (1997), there exists a measurable function $g_1 : \mathscr{Y} \times \mathscr{Z} \times [0,1] \to [0,1]$ and a Uniform[0,1] random variable $\hat{U}_1$ satisfying $\hat{U}_1 \perp\!\!\!\perp (Y(a_1,\ldots,a_m), Z)$ and

$$(Y(a_1,\ldots,a_m), Z, U_1) \stackrel{d}{=} (Y(a_1,\ldots,a_m), Z, g_1(Y(a_1,\ldots,a_m), Z, \hat{U}_1)).$$

Moreover, by

$$U_1 \perp\!\!\!\perp_Z Y(a_1,\ldots,a_m),$$

we have

$$g_1(Y(a_1,\ldots,a_m),Z,\hat{U}_1) \perp\!\!\!\perp_Z Y(a_1,\ldots,a_m)$$

there exists some measurable function $g_1' : \mathcal{Z} \times [0,1] \to [0,1]$ such that

$$g_1(Y(a_1,\ldots,a_m),Z,\hat{U}_1) = g_1'(Z,\hat{U}_1)$$

and

$$\hat{U}_1 \perp\!\!\!\perp (Z, Y(a_1,\ldots,a_m)).$$

In other words, we have

$$(Y(a_1,\ldots,a_m),Z,U_1) \stackrel{d}{=} (Y(a_1,\ldots,a_m),Z,g_1'(Z,\hat{U}_1)).$$

Repeating these steps, we again have from Theorem 5.10 in Kallenberg (1997) that there exists a measurable function $g_2 : \mathcal{Y} \times \mathcal{Z} \times [0,1]^2 \to [0,1]$ and a Uniform[0,1] random variable $\hat{U}_2$ satisfying

$$\begin{aligned}&(Y(a_1,\ldots,a_m),Z,U_1,U_2)\\ \stackrel{d}{=}\ &(Y(a_1,\ldots,a_m),Z,g_1'(Z,\hat{U}_1),g_2(Y(a_1,\ldots,a_m),Z,\hat{U}_1,\hat{U}_2))\end{aligned}$$

and

$$\hat{U}_2 \perp\!\!\!\perp (Z, Y(a_1,\ldots,a_m), \hat{U}_1).$$

Again by

$$U_1 \perp\!\!\!\perp_Z Y(a_1,\ldots,a_m),$$

we have a measurable function $g_2' : \mathcal{Z} \times [0,1]^2 \to [0,1]$ that satisfies

$$\begin{aligned}&(Y(a_1,\ldots,a_m),Z,U_1,U_2)\\ \stackrel{d}{=}\ &(Y(a_1,\ldots,a_m),Z,g_1'(Z,\hat{U}_1),g_2'(Z,\hat{U}_1,\hat{U}_2)).\end{aligned}$$

Repeating these steps m times, we have

$$\begin{aligned}&(Y(a_1,\ldots,a_m),Z,U_1,U_2,\ldots,U_m)\\ \stackrel{d}{=}\ &(Y(a_1,\ldots,a_m),Z,g_1'(Z,\hat{U}_1),g_2'(Z,\hat{U}_1,\hat{U}_2),\ldots,g_m'(Z,\hat{U}_1,\hat{U}_2,\ldots,\hat{U}_m))\end{aligned}$$

with

$$\hat{U}_j \perp\!\!\!\perp (Z, Y(a_1,\ldots,a_m), \hat{U}_1,\ldots,\hat{U}_{j-1}), j = 1,\ldots,m.$$

We notice that the right side of the equation have conditional independence property

$$(g_1'(Z,\hat{U}_1), g_2'(Z,\hat{U}_1,\hat{U}_2),\ldots,g_m'(Z,\hat{U}_1,\hat{U}_2,\ldots,\hat{U}_m)) \perp\!\!\!\perp_Z Y(a_1,\ldots,a_m).$$

This implies the same property holds for the left side of the equation, that is

$$(U_1,\ldots,U_m) \perp\!\!\!\perp_Z Y(a_1,\ldots,a_m).$$

□

G Proof of Lemma 2

Proof sketch. The lemma is an immediate consequence of Lemma 2.22 in Kallenberg (1997) and "no unobserved single-cause confounders". We also rely $p(\theta_{1:m})$ and $p(z_i \mid \boldsymbol{a}_i)$ are point masses, so they are *a priori* independent of the potential outcomes and the other latent variables. □

Proof. For simplicity, we consider continuous random variables A_{ij}, Z_i, θ_j. Also, we assume there are no single-cause confounders. The proof can be easily extended to accommodate discrete random variables and observed single-cause confounders.

We first state the regularity condition: The domains of the causes, $\mathscr{A}_j$, $j = 1,\ldots,m$ are Borel subsets of compact intervals. Without loss of generality, we could assume $\mathscr{A}_j = [0,1]$, $j = 1,\ldots,m$.

By Lemma 2.22 in Kallenberg (1997), there exists some measurable function $f_j : \mathscr{Z} \times [0,1] \to [0,1]$ such that $\gamma_{ij} \perp\!\!\!\perp Z_i$ and

$$A_{ij} = f_j(Z_i, \gamma_{ij}).$$

Furthermore, there exists some measurable function $h_{ij} : \Theta \times [0,1] \to [0,1]$ such that

$$\gamma_{ij} = h_{ij}(\theta_j, \omega_{ij}),$$

where $\omega_{ij} \perp\!\!\!\perp (Z_i, \theta_j)$ and $\omega_{ij} \sim \text{Uniform}[0,1]$. Lastly, we write

$$U_{ij} = F_{ij}^{-1}(\gamma_{ij}) \sim \text{Uniform}[0,1],$$

where F_{ij} is the cumulative distribution function of γ_{ij}.

Eq. 35 implies that $\omega_{ij}, i = 1,\ldots,n, j = 1,\ldots,m$ are jointly independent: if they were not, then $A_{ij} = f_j(Z_i, h_{ij}(\theta_j, \omega_{ij}))$ would not have been conditionally independent given Z_i, θ_j.

We thus have

$$A_{ij} = f_j(Z_i, U_{ij}),$$

where $U_{ij} := F_{ij}^{-1}(h_{ij}(\theta_i, \omega_{ij}))$.

Below we will prove that U_{ij} satisfies

$$(U_{i1},\ldots,U_{im}) \perp\!\!\!\perp (Z_i, Y_i(a_1,\ldots,a_m)). \tag{69}$$

We will rely on the "no single-cause confounders" assumption and the consistency of substitute confounder assumption $p(z_i \mid \boldsymbol{a}_i) = \delta_{f_\theta(\boldsymbol{a}_i)}$.

First, we notice that $\theta_{1:m}$ are point masses; they satisfy $(\theta_1,\ldots,\theta_m) \perp\!\!\!\perp (Z_i, Y_i(a_1,\ldots,a_m))$.

Next, we notice that the "no single-cause confounders" assumption implies that there exists a random variable $\tilde{Z}_i$ such that

$$p(a_{i1},\ldots,a_{im} \mid \tilde{z}_i) = \prod_{j=1}^{m} p(a_{ij} \mid \tilde{z}_i) \tag{70}$$

and

$$A_{i1},\ldots,A_{im} \perp Y_i(a_1,\ldots,a_m)|\tilde{Z}_i. \tag{71}$$

Moreover, no sigma algebra smaller than $\tilde{Z}_i$ satisfies Eq. 70. Further, the consistency of substitute confounder assumption $Z_i = f_\theta(\boldsymbol{A}_i)$ required for the factor model implies that the $\tilde{Z}_i$ that satisfies Eq. 70 is unique, i.e. $\tilde{Z}_i \overset{a.s.}{=} Z_i$. The reason is that the consistency of substitute confounder assumption implies

$$p(\boldsymbol{a}_i, z_i) = p(\boldsymbol{a}_i)p(z_i|\boldsymbol{a}_i) = p(\boldsymbol{a}_i)\cdot\delta_{f_\theta(\boldsymbol{a}_i)},$$

which is a function of $p(\boldsymbol{a}_i)$ by construction. This is a key step that illustrates how the consistency of substitute confounder assumption interacts with the no single-cause confounder assumption to provide causal identification. Hence, Z_i also satisfies the unconfoundedness condition Eq. 71, which implies Eq. 69 and also

$$(\omega_{i1},\ldots,\omega_{im}) \perp\!\!\!\perp (Y_i(a_1,\ldots,a_m), Z_i)$$

or equivalently, $(\omega_{i1},\ldots,\omega_{im}) \perp\!\!\!\perp Y_i(a_1,\ldots,a_m)|Z_i$. In particular, for $m=2$, we have

$$\begin{aligned}
&p(Y_i(a_1,\ldots,a_m),\omega_{i1},\omega_{i2}|Z_i)\\
=&p(\omega_{i1}|Z_i)\cdot p(Y_i(a_1,\ldots,a_m)|\omega_{i1},Z_i)\cdot p(\omega_{i2}|\omega_{i1},Y_i(a_1,\ldots,a_m),Z_i)\\
=&p(\omega_{i1}|Z_i)\cdot p(Y_i(a_1,\ldots,a_m)|Z_i)\cdot p(\omega_{i2}|Z_i)
\end{aligned}$$

Finally, this argument illustrates how the "no single-cause confounders" assumption interacts with the consistency of substitute confounder assumption.

If all pre-treatment single-cause confounders W_i are observed, we can simply expand Z_i; we consider $Z_i' := (Z_i, W_i)$ in the place of Z_i. The same argument applies. □

H Proof of Lemma 3

We first define multi-cause confounders. A multi-cause confounder is a confounder that confounds two or more causes. The following definition formalizes this idea. This definition stems from Definition 4 of VanderWeele and Shpitser (2013).

Definition 6. *(Multi-cause confounder) A pretreatment covariate C_i is a multi-cause confounder if there exists a set of pre-treatment covariates V_i (possibly empty) and a set $J \subset \{1,\ldots,m\}$ with $|J| \geq 2$ such that $(A_{ij})_{j\in J} \perp\!\!\!\perp Y_i(a_{i1},\ldots,a_{im})|\sigma(V_i,C_i)$. Moreover, there is no proper subset S_i of $\sigma(V_i,C_i)$ and no proper subset J' of J such that $(A_{ij})_{j\in J'} \perp\!\!\!\perp Y_i(a_{i1},\ldots,a_{im})|S_i$.*

Proof sketch. This proposition is a consequence of Lemma 1, Lemma 2, and a proof by contradiction. The intuition is that if a confounder affects two or more causes then the substitute confounder Z_i must have captured it. Why? Obtain the substitute confounder Z_i from a factor model; Lemma 1 ensures that it satisfies unconfoundedness. Now suppose we omitted a multi-cause confounder C_i. Then the substitute confounder Z_i could not have satisfied unconfoundedness: the omitted confounder C_i renders the causes and potential outcomes conditionally dependent, even given Z_i. Figure 1 gives the intuition with a graphical model and Appendix H gives a detailed proof. □

Proof. Without loss of generality, we work with two-cause confounders. The proof is directly applicable to general multi-cause confounders.

We prove the proposition by contradiction. Suppose there exists such a multi-cause confounder $W_{i,bad}$ that is not measurable with respect to $\sigma(Z_i)$; we show that Z_i could not have satisfied the factor model Eq. 36.

By Lemma 2.22 in Kallenberg (1997), there exist some function f_j such that $A_{ij} = f_j(Z_i, U_{ij})$, where $U_{ij} \perp\!\!\!\perp Z_i$. ($f_j$ is non-constant in Z_i.)

Then $W_{i,bad}$ being a multi-cause confounder has two implications:

1. There exist j_1, j_2 and nontrivial functions g_1, g_2 such that $U_{ij_1} = g_1(W_{i,bad}, \gamma_{ij_1})$ and $U_{ij_2} = g_2(W_{i,bad}, \gamma_{ij_2})$, where $(\gamma_{ij_1}, \gamma_{ij_2}) \perp\!\!\!\perp W_{i,bad}$;
2. There exists a nontrivial function h such that $Y_i(a_{i1}, \ldots, a_{im}) = h(W_{i,bad}, \epsilon)$, where $\epsilon \perp\!\!\!\perp W_{i,bad}$.

These two statements implies that

$$(U_{ij_1}, U_{ij_2}) \not\perp\!\!\!\perp Y_i(a_{i1}, \ldots, a_{im}) | Z_i,$$

because $W_{i,bad}$ is not measurable with respect to $\sigma(Z_i)$. This implies

$$(U_{i1}, \ldots, U_{im}) \not\perp\!\!\!\perp Y_i(a_{i1}, \ldots, a_{im}) | Z_i.$$

It contradicts the fact that Z_i comes from the factor model (Eq. 35) with $(U_{i1}, \ldots, U_{im}) \perp\!\!\!\perp Y_i(a_{i1}, \ldots, a_{im}) | Z_i$. Therefore, there does not exist such a multi-cause confounder. □

Corollary 9. *Under "no unobserved single-cause confounders", any confounder must be measurable with respect to the σ-algebra generated by the substitute confounder Z_i and the observed covariates X_i.*

Proof. Because of "no unobserved single-cause confounders", a single-cause confounder must be measurable with respect to the observed covariates X_i. Because of Lemma 3, a multi-cause confounder must be measurable with respect to the substitute confounder Z_i. Thus all confounders must be measurable with respect to the union of the substitute confounders and the observed covariates (Z_i, X_i). □

Corollary 9 shows how the "no unobserved single-cause confounder" assumption is necessary for the deconfounder; the substitute confounder Z_i can only handle multi-cause confounders.

I Proof of Lemma 4

Proof sketch. The deconfounder separates inference of the substitute confounder from estimation of causal effects; see Algorithm 1. This two-stage procedure guarantees that the substitute confounder is "pre-treatment"; it does not contain a mediator. The reason is that a mediator is, by

definition, a post-treatment variable that affects the potential outcome. Thus it (almost surely) cannot be identified with only the assigned causes and it is not measurable with respect to the observed (pre-treatment) covariates X_i. Appendix I provides a detailed proof. □

Proof. We prove the proposition by contradiction.

Consider a mediator M. We denote $M_i(a)$ as the potential value of the mediator M for unit i when the assigned cause is a. We show that $M_i(\boldsymbol{a}_i)$ is almost surely not measurable with respect to $\boldsymbol{Z}_i$.

The deconfounder operating in two stages. Inferring the substitute confounder $\boldsymbol{Z}_i$ is separated from estimating the potential outcome. It implies that the substitute confounder is independent of the outcomes conditional on the causes $\boldsymbol{A}_i$: $\boldsymbol{Z}_i \perp\!\!\!\perp Y_i(\boldsymbol{A}_i) | \boldsymbol{A}_i$. The intuition is that, without looking at $Y_i(\cdot)$, the only dependence between $\boldsymbol{Z}_i$ and Y_i must come from $\boldsymbol{A}_i$.

However, a mediator must satisfy $M_i(\boldsymbol{A}_i) \not\perp\!\!\!\perp Y_i(\boldsymbol{A}_i) | \boldsymbol{A}_i$; otherwise, it has no mediation effect (Imai et al., 2010). If a mediator is measurable with $\boldsymbol{Z}_i$, then $\boldsymbol{Z}_i \not\perp\!\!\!\perp Y_i(\boldsymbol{A}_i) | \boldsymbol{A}_i$. This contradicts the conditional independence of $\boldsymbol{Z}_i$ and $Y_i(\boldsymbol{A}_i)$ given $\boldsymbol{A}_i$. We ensured this conditional independence by inferring the substitute confounder $\boldsymbol{Z}_i$ based only on the causes $\boldsymbol{A}_i$.

□

As a consequence of "no unobserved single-cause confounders", the substitute confounder, together with the observed covariates, captures all confounders.

J Proof of Proposition 5

The first part is a direct consequence of Lemmas 1 and 2.

We now prove the second part. We provide two constructions.

We start with the first trivial one. For any assigned causes $\boldsymbol{A}_i$, we consider a special case when $\boldsymbol{A}_i \overset{a.s.}{=} \boldsymbol{Z}_i$. We have

$$p(a_{i1}, \ldots, a_{im} | z_i) = \delta_{z_i} = \prod_{j=1}^{m} \delta_{z_{ij}} = \prod_{j=1}^{m} p(a_{ij} | z_i) \tag{72}$$

This step is due to point masses are factorizable. Therefore, we can write the distribution of $\boldsymbol{A}_i$ in the form of a factor model; we set $\theta_j \overset{a.s.}{=} 0, j = 1, \ldots, m$ and $\boldsymbol{Z}_i \overset{a.s.}{=} \boldsymbol{A}_i$:

$$\begin{aligned} p(\theta_{1:m}, z_{1:n}, \boldsymbol{a}_{1:n}) &= p(\theta_{1:m}) p(z_{1:n} | \theta_{1:m}) p(\boldsymbol{a}_{1:n} | z_{1:n}, \theta_{1:m}) && (73) \\ &= p(\theta_{1:m}) p(z_{1:n}) p(\boldsymbol{a}_{1:n} | z_{1:n}) && (74) \\ &= p(\theta_{1:m}) p(z_{1:n}) \prod_{i=1}^{n} \prod_{j=1}^{m} p(a_{ij} | z_i) && (75) \end{aligned}$$

The second equality is due to $\boldsymbol{Z}_i \perp\!\!\!\perp \theta_{1:m}$ and $\boldsymbol{A}_i \perp\!\!\!\perp \theta_{1:m} | \boldsymbol{Z}_i$. They are because θ_j's are point masses. The third equality is due to the SUTVA assumption and Eq. 72.

Choosing $Z_i \overset{a.s.}{=} \boldsymbol{A}_i$, that is letting the substitute confounder Z_i be the same as the assigned causes $\boldsymbol{A}_i$, does not help with causal inference; see a related discussion on overlap around Eq. 6.

This result is only meant to exemplify the large capacity of factor models. Finally, this $Z_i \overset{a.s.}{=} \boldsymbol{A}_i$ example also illustrates the fact that a factor model capturing $p(\boldsymbol{a}_i)$ is not necessarily the true assignment model.

We now present the second construction. It relies on copulas and the Sklar's theorem. We follow the modified distribution function from Rüschendorf (2009). Let X be a real random variable with distribution function F and let $V \sim U(0,1)$ be uniformly distributed on $(0,1)$ and independent of X. The modified distribution function $F(x,\lambda)$ is defined by

$$F(x,\lambda) := P(X < x) + \lambda P(X = x). \tag{76}$$

Then if we construct U variables as

$$U := F(X,V), \tag{77}$$

then we have

$$U = F(X-) + V(F(X) - F(X-)), \tag{78}$$

$$U \overset{d}{=} Uniform(0,1), \tag{79}$$

$$X \overset{a.s.}{=} F^{-1}(U). \tag{80}$$

Now we set $Z_{ij} = F_{ij}^{-1}(A_{ij})$, where F_{ij} is the modified distribution function of A_{ij}. We also set $\theta_j, j = 1,\ldots,m$ as point masses. The Sklar's theorem then implies

$$\begin{aligned} p(\theta_{1:m}, z_{1:n}, \boldsymbol{a}_{1:n}) &= p(\theta_{1:m}) p(z_{1:n} \mid \theta_{1:m}) p(\boldsymbol{a}_{1:n} \mid z_{1:n}, \theta_{1:m}) && (81)\\ &= p(\theta_{1:m}) p(z_{1:n}) p(\boldsymbol{a}_{1:n} \mid z_{1:n}, \theta_{1:m}) && (82)\\ &= p(\theta_{1:m}) p(z_{1:n}) \prod_{i=1}^{n} \prod_{j=1}^{m} p(a_{ij} \mid z_i, \theta_j) && (83) \end{aligned}$$

The second equality is due to $\theta_{1:m}$ being point masses; $\theta_j, j = 1,\ldots,m$ can be considered as parameters of the marginal distribution of A_{ij}. The third equality is due to the SUTVA assumption and the Sklar's theorem.

This construction aligns more closely with the idea of the deconfounder; it aims to capture multi-causes confounders that induces the dependence structure, i.e. the copula. However, the deconfounder is different from directly estimating the copula; the latter is a more general (and harder) problem.

K Proof of Theorem 6

Proof sketch. Theorem 6 rely on two results: (1) "no unobserved single-cause confounders" and Lemma 3 ensure (X_i, Z_i) capture all confounders; (2) the pre-treatment nature of X_i and Lemma 4

ensure (X_i, Z_i) capture no mediators. These results assert unconfoundedness given the substitute confounders Z_i and the observed covariates X_i. They greenlight us for causal inference if the factor model admits consistent estimates of Z_i, i.e. the substitute confounder has a degenerate distribution $P(Z_i \mid \boldsymbol{A}_i) = \delta_{f(\boldsymbol{A}_i)}$.

Given these results, Theorem 6 identifies the average causal effect of all the causes by assuming $\nabla_{\boldsymbol{a}} f(a_1, \ldots, a_m) = 0$ almost everywhere and a separable outcome model. These two assumptions let us identify the average causal effect without assuming overlap.

More specifically, $\nabla_{\boldsymbol{a}} f(a_1, \ldots, a_m) = 0$ roughly requires that the substitute confounder is a step function of the all causes. In other words, we can partition all possible values of $(a_1, \ldots, a_m)$ into countably many regions. In each region, the value of the substitute confounder must be a constant. But the substitute confounder can take different values in different regions. This condition ensures that the average causal effect $\mathbb{E}_Y[Y_i(\boldsymbol{a})] - \mathbb{E}_Y\left[Y_i(\boldsymbol{a}')\right]$ is identifiable if $\boldsymbol{a}$ and $\boldsymbol{a}'$ belong to the same region.

Further, we assume the outcome model be separable in the substitute confounder and the causes. It roughly requires that there is no interaction between the substitute confounder and the causes. This separability condition lets us identify the average causal effect for all values of $\boldsymbol{a}$ and $\boldsymbol{a}'$. The full proof is in Appendix K. □

Proof. For notational simplicity, denote $\boldsymbol{a} = (a_1, \ldots, a_m)$, $\boldsymbol{a}' = (a'_1, \ldots, a'_m)$, and $\boldsymbol{A}_i = (A_{i1}, \ldots, A_{im})$. We also write $f_\theta(\cdot) = f(\cdot)$.

We start with rewriting $\mathbb{E}_Y[Y_i(\boldsymbol{a})] - \mathbb{E}_Y\left[Y_i(\boldsymbol{a}')\right]$ using the unconfoundedness assumption and the separability assumption.

First notice that

$$\begin{aligned}
\mathbb{E}_Y[Y_i(\boldsymbol{a})] &= \mathbb{E}_{Z,X}[\mathbb{E}_Y[Y_i(\boldsymbol{a}) \mid X_i, Z_i]] && (84)\\
&= \mathbb{E}_X[f_1(\boldsymbol{a}, X_i)] + \mathbb{E}_Z[f_2(Z_i)]. && (85)
\end{aligned}$$

The first equality is due to the tower property. The second equality is due to the separability assumption. The third equality is due to linearity of expectations.

Hence we have

$$\begin{aligned}
\mathbb{E}_Y[Y_i(\boldsymbol{a})] - \mathbb{E}_Y\left[Y_i(\boldsymbol{a}')\right] &= \mathbb{E}_X[f_1(\boldsymbol{a}, X_i)] - \mathbb{E}_X\left[f_1(\boldsymbol{a}', X_i)\right] && (86)\\
&= \int_{C(\boldsymbol{a}, \boldsymbol{a}')} \nabla_{\boldsymbol{a}} \mathbb{E}_X[f_1(\boldsymbol{a}, X_i)] \, \mathrm{d}\boldsymbol{a}, && (87)
\end{aligned}$$

where $C(\boldsymbol{a}, \boldsymbol{a}')$ is a line where $\boldsymbol{a}$ and $\boldsymbol{a}'$ are the end points. The second equality is due to the fundamental theorem of calculus.

Next we see how the gradient of the potential outcome function $\nabla_{\boldsymbol{a}} \mathbb{E}_X[f_1(\boldsymbol{a}, X_i)]$ relates to the gradient of the outcome model we fit. The key idea here is that the two gradients are equal in regions $\{\boldsymbol{a} : f(\boldsymbol{a}) = c\}$ for each c.

We will rely on the consistent substitute confounder assumption. Notice that, for almost all $\boldsymbol{a}$, we have

$$\nabla_{\boldsymbol{a}}\mathbb{E}_X[f_1(\boldsymbol{a})] = \nabla_{\boldsymbol{a}}\mathbb{E}_X[f_3(\boldsymbol{a})] \tag{88}$$

It is due to two observations. The first observation is that

$$\begin{aligned}
&\nabla_{\boldsymbol{a}}\mathbb{E}_X[\mathbb{E}_Y[Y_i \mid Z_i = f(\boldsymbol{a}), A_i = \boldsymbol{a}, X_i]] && (89)\\
=&\nabla_{\boldsymbol{a}}\mathbb{E}_X[\mathbb{E}_Y[Y_i(\boldsymbol{a}) \mid Z_i = f(\boldsymbol{a}), A_i = \boldsymbol{a}, X_i]] && (90)\\
=&\nabla_{\boldsymbol{a}}\mathbb{E}_X[\mathbb{E}_Y[Y_i(\boldsymbol{a}) \mid Z_i = f(\boldsymbol{a}), X_i]] && (91)\\
=&\nabla_{\boldsymbol{a}}\mathbb{E}_X[f_1(\boldsymbol{a}, X_i)] + \nabla_{\boldsymbol{a}} f_2(f(\boldsymbol{a})) && (92)\\
=&\nabla_{\boldsymbol{a}}\mathbb{E}_X[f_1(\boldsymbol{a}, X_i)] + \nabla_{f(\boldsymbol{a})} f_2 \cdot \nabla_{\boldsymbol{a}} f(\boldsymbol{a}) && (93)\\
=&\nabla_{\boldsymbol{a}}\mathbb{E}_X[f_1(\boldsymbol{a}, X_i)] && (94)
\end{aligned}$$

The first equality is due to SUTVA. The second equality is due to is due to Proposition 5.1: $Y_i(\boldsymbol{a}) \perp \boldsymbol{A}_i \mid X_i, Z_i$. The third equality is due to the separability condition. The fourth equality is due to the chain rule. The fifth equality is due to $\nabla_{\boldsymbol{a}} f(\boldsymbol{a}) = 0$ up to a set of Lebesgue measure zero.

The second observation is that

$$\begin{aligned}
&\nabla_{\boldsymbol{a}}\mathbb{E}_X[\mathbb{E}_Y[Y_i \mid Z_i = f(\boldsymbol{a}), \boldsymbol{A}_i = \boldsymbol{a}, X_i]] && (95)\\
=&\nabla_{\boldsymbol{a}}\mathbb{E}_X[f_3(\boldsymbol{a}, X_i)] + \nabla_{\boldsymbol{a}} f_4(f(\boldsymbol{a})) && (96)\\
=&\nabla_{\boldsymbol{a}}\mathbb{E}_X[f_3(\boldsymbol{a}, X_i)] && (97)
\end{aligned}$$

Hence Eq. 88 is true because f_1 and f_3 are continuously differentiable.

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_Y[Y_i(\boldsymbol{a})] - \mathbb{E}_Y\left[Y_i(\boldsymbol{a}')\right] && (98)\\
=&\int_{C(\boldsymbol{a},\boldsymbol{a}')} \nabla_{\boldsymbol{a}}\mathbb{E}_X[f_1(\boldsymbol{a}, X_i)]\,\mathrm{d}\boldsymbol{a} && (99)\\
=&\int_{C(\boldsymbol{a},\boldsymbol{a}')} \nabla_{\boldsymbol{a}}\mathbb{E}_X[f_3(\boldsymbol{a}, X_i)]\,\mathrm{d}\boldsymbol{a} && (100)\\
=&\mathbb{E}_X[f_3(\boldsymbol{a}, X_i)] - \mathbb{E}_X\left[f_3(\boldsymbol{a}', X_i)\right] && (101)\\
=&(\mathbb{E}_X[f_3(\boldsymbol{a}, X_i)] + \mathbb{E}[f_4(Z_i)]) - (\mathbb{E}_X\left[f_3(\boldsymbol{a}', X_i)\right] + \mathbb{E}[f_4(Z_i)]) && (102)\\
=&\int \mathbb{E}_Y\left[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}', X_i, Z_i\right] P(Z_i, X_i)\,\mathrm{d}Z_i\,\mathrm{d}X_i\\
&-\int \mathbb{E}_Y[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, X_i, Z_i] P(Z_i, X_i)\,\mathrm{d}Z_i\,\mathrm{d}X_i && (103)\\
=&\mathbb{E}_{Z,X}[\mathbb{E}_Y[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i, X_i]] - \mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}', Z_i, X_i\right]\right]. && (104)
\end{aligned}$$

The first equality is due to Eq. 87. The second equality is due to Eq. 88. The third equality is due to the fundamental theorem of calculus. The fourth equality is due to simple algebra. The fifth equality is due to the separability condition.

□

L Proof of Theorem 7

Proof. Lemma 1 and Lemma 2, together with "no unobserved single-cause confounders", ensures that the substitute confounder Z_i and the observed covariate X_i satisfies

$$(A_{i1},\ldots,A_{im}) \perp\!\!\!\perp Y_i(a_{i1},\ldots,a_{im}) \,|\, Z_i, X_i. \tag{105}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_{1:k}, A_{i,(k+1):m})\right]\right] && (106)\\
=&\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_1,\ldots,a_k,A_{i,k+1},\ldots,A_{im})\right]\right] && (107)\\
=&\mathbb{E}_{Z,X}\left[\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_1,\ldots,a_k,A_{i,k+1},\ldots,A_{im}) \,|\, Z_i, X_i\right]\right]\right] && (108)\\
=&\mathbb{E}_{Z,X}\left[\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_1,\ldots,a_k,A_{i,k+1},\ldots,A_{im}) \,|\, Z_i, X_i, A_{i1}=a_1,\ldots,A_{ik}=a_k\right]\right]\right] && (109)\\
=&\mathbb{E}_{Z,X}\left[\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(A_{i1},\ldots,A_{ik},A_{i,k+1},\ldots,A_{im}) \,|\, Z_i, X_i, A_{i1}=a_1,\ldots,A_{ik}=a_k\right]\right]\right] && (110)\\
=&\mathbb{E}_{Z,X}\left[\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i \,|\, Z_i, X_i, A_{i1}=a_1,\ldots,A_{ik}=a_k\right]\right]\right] && (111)\\
=&\mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \,|\, Z_i, X_i, A_{i1}=a_1,\ldots,A_{ik}=a_k\right]\right] && (112)\\
=&\mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \,|\, Z_i, X_i, A_{i,1:k}=a_{1:k}\right]\right] && (113)
\end{aligned}$$

The first equality is an expansion of the notations. The second equality is due to the tower property. The third equality is due to Eq. 105. The fourth equality is due to $A_{i1}=a_1,\ldots,A_{ik}=a_k$. The fifth equality is due to SUTVA. The sixth equality is due to the inner expectation does not depend on $A_{(k+1):m}$.

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_{1:k}, A_{i,(k+1):m})\right]\right] - \mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a'_{1:k}, A_{i,(k+1):m})\right]\right]\\
=&\mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \,|\, Z_i, X_i, A_{i,1:k}=a_{1:k}\right]\right] - \mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \,|\, Z_i, X_i, A_{i,1:k}=a'_{1:k}\right]\right]
\end{aligned}$$

by the linearity of expectation.

Finally, $\mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \,|\, Z_i, X_i, A_{i,1:k}=a_{1:k}\right]\right]$ can be estimated from the observed data because (1) $A_{i,1:k}$ satisfy overlap with respect to (Z_i, X_i) and (2) the substitute confounder Z can be consistently estimated. □

M Proof of Theorem 8

Proof. As with Theorem 6 and Theorem 7, Theorem 8 relies on the unconfoundedness given the substitute confounders Z_i and the observed covariates X_i due to Lemma 3 and Lemma 4.

Given this unconfoundedness, Theorem 8 identifies the mean potential outcome of an individual given its current cause assignment $A_i = (a_1,\ldots,a_m)$; it only requires that the new cause assignment of interest $(a'_1,\ldots,a'_m)$ lead to the same substitute confounder estimate: $f(a_1,\ldots,a_m) = f(a'_1,\ldots,a'_m)$.

To prove identification, we rewrite this conditional mean potential outcome

$$\begin{aligned}
&\mathbb{E}_Y\left[Y_i(a'_1,\ldots,a'_m)\,|\,A_{i1}=a_1,\ldots,A_{im}=a_m\right] && (114)\\
=&\mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i(a'_1,\ldots,a'_m)\,|\,A_{i1}=a_1,\ldots,A_{im}=a_m,Z_i,X_i\right]\right] && (115)\\
=&\mathbb{E}_X\left[\mathbb{E}_Y\left[Y_i(a'_1,\ldots,a'_m)\,|\,A_{i1}=a_1,\ldots,A_{im}=a_m,Z_i=f(a_1,\ldots,a_m),X_i\right]\right] && (116)\\
=&\mathbb{E}_X\left[\mathbb{E}_Y\left[Y_i(a'_1,\ldots,a'_m)\,|\,A_{i1}=a'_1,\ldots,A_{im}=a'_m,Z_i=f(a_1,\ldots,a_m),X_i\right]\right] && (117)\\
=&\mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i\,|\,A_{i1}=a'_1,\ldots,A_{im}=a'_m,Z_i,X_i\right]\right] && (118)
\end{aligned}$$

The first equality is due to the tower property. The second equality is due to the consistency requirement on the substitute confounder $P(Z_i\,|\,\boldsymbol{A}_i)=\delta_{f(\boldsymbol{A}_i)}$. The third equality is due to unconfoundedness given Z_i,X_i. The fourth equality is estimable from the data because $f(a_1,\ldots,a_m)=f(a'_1,\ldots,a'_m)$. Hence the nonparametric identification of $\mathbb{E}_Y\left[Y_i(a'_1,\ldots,a'_m)\,|\,A_{i1}=a_1,\ldots,A_{im}=a_m\right]$ is established. We note that this identification result does not require overlap. □

N Details of Section 6.2

We follow Song et al. (2015) in simulating the allele frequencies. We present the full details here.

We simulate the $n\times m$ matrix of genotypes A from $A_{ij}\sim\text{Binomial}(2,F_{ij})$, where F is the $n\times m$ matrix of allele frequencies. Let $F=\Gamma S$, where Γ is $n\times d$ and S is $d\times m$ with $d\le m$. The $d\times m$ matrix S encodes the genetic population structure. The $n\times d$ matrix Γ maps how the structure affects the allele frequencies of each SNP. Table 19 details how we generate Γ and S for each simulation setup.

For each simulation scenarios, we generate 100 independent studies. We then simulate a trait; we consider two types: one continuous and one binary. For each trait, three components contributing to the trait: causal signals $\sum_{j=1}^m\beta_j a_{ij}$, confounder λ_i, and random effects ϵ_i.

Notice that the SNPs are affected by some latent population structure. We simulate the confounder λ_i and the random effects ϵ_i so that they depend on the latent population structure as well.

For the confounder λ_i, we first perform K-means clustering on the columns of S with $K=3$ using Euclidean distance. This assigns each individual i to one of three mutually exclusive cluster sets $\mathcal{S}_1,\mathcal{S}_2,\mathcal{S}_3$, where $\mathcal{S}_k\subset\{1,2,\ldots,n\}$. Set $\lambda_j=k$ if $j\in\mathcal{S}_k,k=1,2,3$.

We then simulate the random effects ϵ_i. Let $\tau_1^2,\tau_2^2,\tau_3^2\overset{iid}{\sim}\text{InvGamma}(3,1)$, and set $\sigma_i^2=\tau_k^2$ for all $j\in\mathcal{S}_i,k=1,2,3$. Draw $\epsilon_i\sim\mathcal{N}(0,\sigma_i^2)$.

We control the SNR to mimic the highly noisy nature of GWAS data sets. In the low SNR setting, we simulate datasets of $n=5000$ individuals and $m=100{,}000$ SNPs; we let the causal signals $\sum_{j=1}^m\beta_j a_{ij}$ contribute to $\nu_{\text{gene}}=0.1$ of the variance, the confounder λ_i contribute $\nu_{\text{conf}}=0.2$, and the random effects ϵ_i contribute $\nu_{\text{noise}}=0.7$. We set the first 10% of the m SNPs to be the true causal SNPs ($\beta_j\neq 0,\beta_j\overset{iid}{\sim}\mathcal{N}(0,1)$; $\beta_j=0$ for the rest of the SNPs. In the high SNR setting, we simulate datasets of $n=5{,}000$ individuals and $m=5{,}000$ SNPs; we have $\nu_{\text{gene}}=0.4$, $\nu_{\text{conf}}=0.4$, and $\nu_{\text{noise}}=0.2$.

We set

$$\lambda_i \leftarrow \left[\frac{s.d.\{\sum_{j=1}^{m}\beta_j a_{ij}\}_{i=1}^{n}}{\sqrt{v_{\text{gene}}}}\right]\left[\frac{\sqrt{v_{\text{conf}}}}{s.d.\{\lambda_i\}_{i=1}^{n}}\right]\lambda_i, \tag{119}$$

$$\epsilon_i \leftarrow \left[\frac{s.d.\{\sum_{j=1}^{m}\beta_j a_{ij}\}_{i=1}^{n}}{\sqrt{v_{\text{gene}}}}\right]\left[\frac{\sqrt{v_{\text{noise}}}}{s.d.\{\epsilon_i\}_{i=1}^{n}}\right]\epsilon_i. \tag{120}$$

We finally generate a real-valued outcome from a linear model and a binary outcome from a logistic model:

$$y_{i,\text{real}} = \sum_{j=1}^{m}\beta_j a_{ij} + \lambda_i + \epsilon_i, \tag{121}$$

$$y_{i,\text{binary}} \sim \text{Bernoulli}\left(\frac{1}{1+\exp(\sum_{j=1}^{m}\beta_j a_{ij} + \lambda_i + \epsilon_i)}\right). \tag{122}$$

		Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
$\alpha = 0.01$	No control	—	40.68	30.37
	Control for confounders*	—	34.35	28.21
	(G)LMM	—	39.09	28.36
	PPCA	0.15	38.14	28.97
	PF	0.16	**34.77**	28.67
	LFA	0.16	35.87	28.33
	GMM	0.02	38.15	29.69
	DEF	0.18	34.84	**28.04**
$\alpha = 0.1$	No control	—	43.87	36.77
	Control for confounders*	—	37.62	33.89
	(G)LMM	—	39.97	35.76
	PPCA	0.21	39.60	35.61
	PF	0.19	38.95	**34.28**
	LFA	0.18	39.28	34.73
	GMM	0.00	44.38	36.44
	DEF	0.20	**38.75**	34.85
$\alpha = 0.5$	No control	—	47.38	41.84
	Control for confounders*	—	43.63	39.86
	(G)LMM	—	47.28	42.91
	PPCA	0.14	46.90	41.41
	PF	0.16	43.29	40.69
	LFA	0.17	43.60	40.77
	GMM	0.02	46.95	42.47
	DEF	0.18	**43.09**	**40.03**
$\alpha = 1.0$	No control	—	53.94	49.32
	Control for confounders*	—	47.12	45.96
	(G)LMM	—	49.21	48.96
	PPCA	0.21	50.57	47.58
	PF	0.19	48.07	46.16
	LFA	0.17	49.27	46.16
	GMM	0.02	52.28	50.31
	DEF	0.23	**47.82**	**45.62**

Table 9: GWAS high-SNR simulation IV: Pritchard-Stephens-Donnelly (PSD). ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms (G)LMM; DEF often performs the best among the five factor models. Predictive checking offers a good indication of when the deconfounder fails.

		Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
$\tau = 0.1$	No control	—	47.47	45.16
	Control for confounders*	—	44.22	43.85
	(G)LMM	—	47.35	44.15
	PPCA	0.08	47.61	44.36
	PF	0.09	47.13	43.69
	LFA	0.09	47.16	43.87
	GMM	0.01	47.55	45.95
	DEF	0.10	**46.95**	**43.62**
$\tau = 0.25$	No control	—	44.68	41.10
	Control for confounders*	—	41.23	39.65
	(G)LMM	—	43.42	**40.67**
	PPCA	0.11	**43.26**	41.28
	PF	0.12	43.30	41.10
	LFA	0.13	43.62	41.65
	GMM	0.01	44.81	41.02
	DEF	0.13	43.35	40.97
$\tau = 0.5$	No control	—	45.18	40.92
	Control for confounders*	—	41.33	37.35
	(G)LMM	—	44.83	40.59
	PPCA	0.10	43.78	**39.99**
	PF	0.09	43.65	40.23
	LFA	0.10	43.88	40.04
	GMM	0.01	46.08	40.76
	DEF	0.12	**43.57**	40.02
$\tau = 1.0$	No control	—	56.57	57.70
	Control for confounders*	—	52.98	55.46
	(G)LMM	—	56.44	56.33
	PPCA	0.14	55.18	57.36
	PF	0.12	55.29	56.31
	LFA	0.13	**54.75**	56.66
	GMM	0.01	57.15	57.55
	DEF	0.12	55.07	**56.22**

Table 10: GWAS high-SNR simulation V: Spatial model. ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder often outperforms (G)LMM. Predictive checking offers a good indication of when the deconfounder fails: GMM poorly captures the SNPs; it can amplify the error in causal estimates.

	Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
No control	—	6.55	5.75
Control for confounders*	—	6.54	5.75
(G)LMM	—	6.54	**5.74**
PPCA	0.14	6.52	**5.74**
PF	0.16	6.53	**5.74**
LFA	0.14	6.54	**5.74**
GMM	0.01	6.54	**5.74**
DEF	0.19	**6.47**	**5.74**

Table 11: GWAS low-SNR simulation I: Balding-Nichols Model. ("Control for all confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms LMM; DEF performs the best among the five factor models; it also outperforms using the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

	Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
No control	—	8.31	4.85
Control for confounders*	—	8.28	4.85
(G)LMM	—	8.29	4.85
PPCA	0.14	8.29	4.85
PF	0.15	8.29	4.85
LFA	0.17	8.26	4.85
GMM	0.02	8.30	4.85
DEF	0.20	**8.11**	**4.84**

Table 12: GWAS low-SNR simulation II: 1000 Genomes Project (TGP). ("Control for all confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms LMM; DEF performs the best among the five factor models; it also outperforms using the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

	Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
No control	—	9.59	5.84
Control for confounders*	—	9.52	5.84
(G)LMM	—	9.57	5.84
PPCA	0.14	9.55	5.84
PF	0.13	9.56	5.84
LFA	0.14	9.54	5.84
GMM	0.03	9.59	5.84
DEF	0.16	**9.47**	**5.83**

Table 13: GWAS low-SNR simulation III: Human Genome Diversity Project (HGDP). ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms LMM; DEF performs the best among the five factor models; it also outperforms using the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

		Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
$\alpha = 0.01$	No control	—	3.73	3.23
	Control for confounders*	—	3.71	3.23
	(G)LMM	—	3.71	3.23
	PPCA	0.13	3.64	3.23
	PF	0.16	3.67	3.23
	LFA	0.16	3.66	3.23
	GMM	0.02	3.72	3.23
	DEF	0.18	**3.59**	**3.22**
$\alpha = 0.1$	No control	—	4.09	3.84
	Control for confounders*	—	4.09	3.84
	(G)LMM	—	4.09	3.84
	PPCA	0.20	4.08	3.84
	PF	0.18	4.08	3.84
	LFA	0.18	4.07	3.84
	GMM	0.00	4.09	3.84
	DEF	0.20	**4.05**	**3.83**
$\alpha = 0.5$	No control	—	4.82	4.14
	Control for confounders*	—	4.81	4.14
	(G)LMM	—	4.82	4.14
	PPCA	0.14	4.81	**4.13**
	PF	0.17	**4.80**	**4.13**
	LFA	0.16	4.81	4.14
	GMM	0.03	4.82	4.14
	DEF	0.19	**4.80**	**4.13**
$\alpha = 1.0$	No control	—	5.43	4.58
	Control for confounders*	—	5.38	4.57
	(G)LMM	—	5.40	4.58
	PPCA	0.21	5.38	**4.57**
	PF	0.16	5.41	**4.57**
	LFA	0.19	5.40	**4.57**
	GMM	0.02	5.43	4.58
	DEF	0.24	**5.37**	**4.57**

Table 14: GWAS low-SNR simulation IV: Pritchard-Stephens-Donnelly (PSD). ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms LMM; DEF performs the best among the five factor models; it also outperforms using the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

		Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
$\tau = 0.1$	No control	—	4.66	4.74
	Control for confounders*	—	4.63	4.73
	(G)LMM	—	4.57	**4.73**
	PPCA	0.09	4.62	4.74
	PF	0.08	4.58	4.74
	LFA	0.09	4.54	**4.73**
	GMM	0.02	4.70	4.74
	DEF	0.10	**4.53**	**4.73**
$\tau = 0.25$	No control	—	4.30	3.81
	Control for confounders*	—	3.81	3.79
	(G)LMM	—	4.28	**3.80**
	PPCA	0.10	4.26	**3.80**
	PF	0.12	4.26	**3.80**
	LFA	0.12	4.27	**3.80**
	GMM	0.01	4.30	3.81
	DEF	0.13	**4.25**	**3.80**
$\tau = 0.5$	No control	—	4.30	3.85
	Control for confounders*	—	3.82	3.83
	(G)LMM	—	4.28	**3.83**
	PPCA	0.11	4.27	**3.83**
	PF	0.09	4.28	3.84
	LFA	0.11	4.27	3.84
	GMM	0.01	4.29	3.84
	DEF	0.13	**4.25**	3.84
$\tau = 1.0$	No control	—	6.71	5.52
	Control for confounders*	—	5.43	5.51
	(G)LMM	—	6.70	5.52
	PPCA	0.14	6.70	5.52
	PF	0.12	6.70	5.52
	LFA	0.12	6.69	5.52
	GMM	0.01	6.72	5.53
	DEF	0.13	**6.62**	**5.51**

Table 15: GWAS low-SNR simulation V: Spatial model. ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder often outperforms LMM; DEF often performs the best among the five factor models. Yet, the deconfounder does not outperform using the (unobserved) confounder information. Spatially-induced SNPs challenge many latent variable models to capture its patterns and fully deconfound causal inference. Predictive checking offers a good indication of when the deconfounder fails: GMM poorly captures the SNPs; it can amplify the error in causal estimates.

Control	Average predictive log-likelihood
No Control	**-1.1**
Control for X	**-1.1**
Control for $\hat{a}_{\text{PPCA}}$	-1.2
Control for $\hat{a}_{\text{PF}}$	-1.2
Control for $\hat{a}_{\text{DEF}}$	-1.2
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.3
Control for $(\hat{a}_{\text{PF}}, X)$	-1.2
Control for $(\hat{a}_{\text{DEF}}, X)$	-1.2

Table 16: Average predictive log-likelihood on a holdout set of all movies. (X represents the observed covariates.) Causal models (the deconfounder) predicts slightly worse than prediction models.

Control	Average predictive log-likelihood
No Control	-2.5
Control for X	-2.1
Control for $\hat{a}_{\text{PPCA}}$	-1.6
Control for $\hat{a}_{\text{PF}}$	**-1.5**
Control for $\hat{a}_{\text{DEF}}$	**-1.5**
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.7
Control for $(\hat{a}_{\text{PF}}, X)$	**-1.5**
Control for $(\hat{a}_{\text{DEF}}, X)$	-1.6

Table 17: Average predictive log-likelihood on the holdout set of non-English movies. (X represents the observed covariates.) On a test set of uncommon movies, causal models with the deconfounder predict better than prediction models.

Control	Average predictive log-likelihood
No Control	-2.1
Control for X	-1.9
Control for $\hat{a}_{\text{PPCA}}$	-1.4
Control for $\hat{a}_{\text{PF}}$	**-1.2**
Control for $\hat{a}_{\text{DEF}}$	-1.3
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.4
Control for $(\hat{a}_{\text{PF}}, X)$	-1.3
Control for $(\hat{a}_{\text{DEF}}, X)$	**-1.2**

Table 18: Average predictive log-likelihood on the holdout set of non-drama/comedy/action movies. (X represents the observed covariates.) On a test set of uncommon movies, causal models with the deconfounder predict better than prediction models.

Model	Simulation details
Balding-Nichols Model (Balding-Nichols)	Each row i of Γ has i.i.d. three independent and identically distributed draws from the Balding- Nichols model: $\gamma_{ik} \overset{iid}{\sim} \text{BN}(p_i, F_i)$, where $k \in \{1,2,3\}$. The pairs (p_i, F_i) are computed by randomly selecting a SNP in the HapMap data set, calculating its observed allele frequency and estimating its F_{ST} value using the Weir & Cockerham estimator (Weir and Cockerham, 1984). The columns of S were Multinomial$(60/210, 60/210, 90/210)$, which reflect the subpopulation proportions in the HapMap data set.
1000 Genomes Project (TGP)	The matrix Γ was generated by sampling $\gamma_{ik} \overset{iid}{\sim} 0.9 \times \text{Uniform}(0, 0.5)$, for $k = 1,2$ and setting $\gamma_{i3} = 0.05$. In order to generate S, we compute the first two principal components of the TGP genotype matrix after mean centering each SNP. We then transformed each principal component to be between $(0,1)$ and set the first two rows of S to be the transformed principal components. The third row of S was set to 1, i.e. an intercept.
Human Genome Diversity Project (HGDP)	Same as TGP but generating S with the HGDP genotype matrix.
Pritchard-Stephens-Donnelly (PSD)	Each row i of Γ has i.i.d. three independent and identically distributed draws from the Balding- Nichols model: $\gamma_{ik} \overset{iid}{\sim} \text{BN}(p_i, F_i)$, where $k \in \{1,2,3\}$. The pairs (p_i, F_i) are computed by randomly selecting a SNP in the HGPD data set, calculating its observed allele frequency and estimating its F_{ST} value using the Weir & Cockerham estimator (Weir and Cockerham, 1984). The estimator requires each individual to be assigned to a subpopulation, which were made according to the $K = 5$ subpopulations from the analysis in Rosenberg et al. (2002). The columns of S were sampled $(s_{1j}, s_{2j}, s_{3j} \overset{iid}{\sim} \text{Dirichlet}(\alpha, \alpha, \alpha)$ for $j = 1, \ldots, m, \alpha = 0.01, 0.1, 0.5, 1$.
Spatial	The matrix Γ was generated by sampling $\gamma_{ik} \overset{iid}{\sim} 0.9 \times \text{Uniform}(0, 0.5)$, for $k = 1,2$ and setting $\gamma_{i3} = 0.05$. The first two rows of S correspond to coordinates for each individual on the unit square and were set to be independent and identically distributed samples from $\text{Beta}(\tau, \tau), \tau = 0.1, 0.25, 0.5, 1$, while the third row of S was set to be 1, i.e. an intercept. As $\tau \to 0$, the individuals are placed closer to the corners of the unit square, while when $\tau = 1$, the individuals are distributed uniformly.

Table 19: Simulating allele frequencies.

References

Airoldi, E., Blei, D., Erosheva, E., and Fienberg, S., editors (2014). *Handbook of Mixed Membership Models and Their Applications*. CRC Press.

Darnell, G., Georgiev, S., Mukherjee, S., and Engelhardt, B. E. (2017). Adaptive randomized dimension reduction on massive data. *Journal of Machine Learning Research*, 18(140):1–30.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Resources*, 7(4):574–578.

Hao, W., Song, M., and Storey, J. D. (2015). Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, pages 51–71.

Kallenberg, O. (1997). Foundations of modern probability. *Collection: Probability and Its Applications, Springer*.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284.

Pearl, J. (2009). *Causality*. Cambridge University Press, 2nd edition.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000a). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000b). Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602):2381–2385.

Rüschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927.

Song, M., Hao, W., and Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. *Nature Genetics*, 47(5):550–554.

Tran, D. and Blei, D. M. (2017). Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*.

VanderWeele, T. J. and Shpitser, I. (2013). On the definition of a confounder. *The Annals of Statistics*, pages 196–220.

Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203.