# Structural Causal Bottleneck Models

**Simon Bing**[*]
Technische Universität Berlin
and
Department of Computer Science
University of Potsdam
bing@campus.tu-berlin.de

**Jonas Wahl**[*]
German Research Center for
Artificial Intelligence (DFKI)
and
Department of Philosophy
University of Bergen

**Jakob Runge**
Department of Computer Science
University of Potsdam

## Abstract

We introduce structural causal bottleneck models (SCBMs), a novel class of structural causal models in which causal effects between high-dimensional variables are mediated by low-dimensional summary statistics, or *bottlenecks*. SCBMs provide a flexible framework for mechanism-specific, target-dependent dimension reduction while remaining estimable via standard learning algorithms. We prove that bottleneck variables are identifiable up to bijection from observational data, and validate this experimentally. We then study causal effect estimation in linear SCBMs under three standard adjustment strategies. For instrumental variables, we identify a population-level non-identifiability of naïve two-stage least squares when all variables are high-dimensional, and show that bottleneck representations restore identification. For mediation and backdoor adjustment, we prove that conditioning on bottleneck variables yields valid effect estimates with explicit finite-sample efficiency gains over conditioning on full high-dimensional variables. We argue that SCBMs provide a principled alternative to existing causal dimension reduction frameworks such as causal representation learning and causal abstractions.

## 1 Introduction

A fundamental aim of scientific inquiry is to uncover and quantify causal relationships among complex phenomena, which often span large spaces, long times, or many individuals. For example, neuroscientists study how neuron clusters respond to tasks (Aoi and Pillow, 2018), while climate scientists examine interactions like the El Niño Southern Oscillation, affecting global weather patterns (Timmermann et al., 2018).

A popular type of model to formalize causal interactions of random variables is the structural causal model (SCM) (Pearl, 2009). An SCM consists of structural equations $\mathbf{X}_j := m_j(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_k}, \boldsymbol{\eta}_j)$, one per variable, specifying how each is determined by its *causal parents* and a noise term $\boldsymbol{\eta}_j$ via mechanism $m_j$. Although most SCM applications assume scalar variables, the framework does not require this. However, modeling interactions of high-dimensional vectors $\mathbf{X}_j$ may quickly become infeasible in practice without additional assumptions. Even if the mechanism functions $m_j$ are assumed to be linear and additive, i.e. $\mathbf{X}_j := \sum_{i \in \mathrm{pa}(j)} \mathbf{X}_i M_{ij} + \boldsymbol{\eta}_j$, with matrices $M_{ij}$, the associated regression tasks require large sample sizes and/or a sufficient degree of regularization to yield reliable outcomes in high dimensions. When estimating causal effects in an SCM, the curse of

[*]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

dimensionality is particularly daunting. The estimation may not only require as input the treatment and outcome variables, but also confounding covariates to remove spurious correlations, increasing the dimension of the input space further.

In this work, we investigate SCMs in which variables depend on their parents only through low-dimensional sufficient statistics, or *bottlenecks*: for each parent $\mathbf{X}_i$ of $\mathbf{X}_j$, a deterministic bottleneck function $b_{ij}$ maps $\mathbf{X}_i$ to $\mathbf{Z}_{ij} = b_{ij}(\mathbf{X}_i)$ such that

$$\mathbf{X}_j := f_j(\mathbf{Z}_{i_1 j}, \ldots, \mathbf{Z}_{i_k j}, \boldsymbol{\eta}_j)$$

depends on its parents only through their bottlenecks. For linear mechanisms, this requirement translates to the matrices $M_{ij}$ having low rank compared to the dimensions of $\mathbf{X}_i$ and $\mathbf{X}_j$.

The bottleneck assumption is natural for high-dimensional causal interactions, where a child depends not on all information in its parents but on emergent properties—a weighted average, a phase, or a system state. For instance, modelling rainfall over West Africa may require only whether the El Niño Southern Oscillation (ENSO) is in an El Niño or La Niña phase, not the full Pacific temperature distribution; such dimension reduction assumptions have been studied in spatially-aggregated vector autoregressive models of climate teleconnections (Tibau et al., 2022). Moreover, different children may depend on different aspects of the same parent, calling for *target-dependent* dimension reduction. However, reducing dimensions before estimating causal effects can discard or misidentify important information (Wahl et al., 2024; Ninad et al., 2025), and existing *sufficient dimension reduction* approaches (Izenman, 1975; Globerson and Tishby, 2003; Li, 2007, 2018) target single treatment-outcome pairs; with the notable exception of Aoi and Pillow (2018), the outcome is typically scalar.

In this work, we introduce structural causal bottleneck models (SCBMs), a class of graphical causal models that enables targeted, mechanism-specific dimension reduction. Our main contributions are:

- **SCBM framework** (Section 2): We formalize the idea that high-dimensional effects are mediated by low-dimensional summary statistics, discuss special cases, and show that bottleneck-equivalent representations are interchangeable for causal inference (Definition 2.2).
- **Identifiability** (Section 3; Lemmas 3.1–3.2): We prove that bottleneck variables are identifiable up to invertible reparametrization from observational data, validate this experimentally across linear and nonlinear mechanisms, and illustrate how the equivalence class induces meaningful partitions of high-dimensional input spaces.
- **Effect estimation** (Section 4): We study causal effect estimation under three standard adjustment strategies for linear SCBMs.
  - *Instrumental variables* (Proposition 4.1; Theorem 4.3): We identify a population-level non-identifiability of naïve 2SLS when all variables are high-dimensional, and establish a novel identification result that hinges on using bottleneck representations.
  - *Mediation and backdoor adjustment* (Lemmas 4.4, 4.6; Propositions 4.5, 4.7): Bottleneck conditioning sets are valid and yield finite-sample efficiency gains over conditioning on full high-dimensional variables, with an explicit efficiency ordering.

**Related work.** SCBMs sit at the intersection of several active research areas; we discuss connections in detail in Appendix C. Causal representation learning (Schölkopf et al., 2021) and causal abstractions (Rubenstein et al., 2017) both seek low-dimensional causal representations, but target full latent/abstracted SCM recovery rather than mechanism-specific, query-targeted maps tied to a known graph. In the single-variable setting, sufficient dimension reduction (Izenman, 1975) and reduced-rank regression share our low-rank premise but do not address multi-variable graphs or vector-valued outcomes. For high-dimensional IV estimation, prior work addresses instrument-side (Bai and Ng, 2010) or treatment-side (Lin et al., 2026) compression; we instead identify a population-level non-identifiability and show that restoring it requires bottleneck representations on both sides of the treatment. Propensity score matching (Rosenbaum and Rubin, 1983) and coarsened exact matching (Blackwell et al., 2009) can also be understood as operating in bottleneck spaces, connecting to our backdoor adjustment results.

## 2 Preliminaries & definitions

Throughout this work, we fix the following notation. $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ will refer to a directed acyclic graph (DAG) with node set $\mathcal{V}$ and edge set $\mathcal{E}$. The set of parents of a node $i \in \mathcal{V}$ will be denoted by $\mathrm{pa}(i)$, the set of its children by $\mathrm{ch}(i)$. The set $\mathcal{V}_{ex}$ is the set of exogenous (i.e. parentless) nodes, while $\mathcal{V}_{end} = \mathcal{V} \backslash \mathcal{V}_{ex}$ is the set of endogenous nodes.

We reserve the letter $d$ to refer to dimensions of vector spaces, e.g. $\mathbb{R}^d$. Unless specified otherwise, we use the word space in the sense of measurable space, i.e. a set endowed with a $\sigma$-algebra. All maps between such spaces are assumed to be measurable. We assume the causal graph $\mathcal{G}$ is known; causal discovery is outside of the scope of this work.

**Definition 2.1** (Structural causal bottleneck model)**.** A *structural causal bottleneck model (SCBM)* is a tuple $\mathfrak{C} = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}, \mathcal{H}, \boldsymbol{\eta}, \mathcal{B}, \mathcal{F}, \mathbf{X} \rangle$ consisting of

- a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a family of *node spaces* $\mathcal{X} = (\mathcal{X}_i)_{i \in \mathcal{V}}$, and a family of *noise spaces* $\mathcal{H} = (\mathcal{H}_i)_{i \in \mathcal{V}}$.
- a family of mutually independent noise vectors $\boldsymbol{\eta} = (\boldsymbol{\eta}_i)_{i \in \mathcal{V}}$, where $\boldsymbol{\eta}_i$ takes values in $\mathcal{H}_i$.
- a family of *bottleneck spaces* $\mathcal{Z} = (\mathcal{Z}_j)_{j \in \mathcal{V}_{end}}$.
- a family $\mathcal{B} = (b_j)_{j \in \mathcal{V}_{end}}$ of surjective *bottleneck functions* $b_j : \prod_{i \in \mathrm{pa}(j)} \mathcal{X}_i \to \mathcal{Z}_j$.
- a family $\mathcal{F} = (f_j)_{j \in \mathcal{V}_{end}}$ of *effect functions* $f_j : \mathcal{Z}_j \times \mathcal{H}_j \to \mathcal{X}_j$.
- a family of random vectors $\mathbf{X} = (\mathbf{X}_i)_{i \in \mathcal{V}}$ with joint distribution $P_{\mathbf{X}}$ that solves the structural assignments

$$\begin{aligned} \mathbf{X}_j &:= \boldsymbol{\eta}_j, && j \in \mathcal{V}_{ex}, \\ \mathbf{X}_j &:= f_j\left(b_j\left((\mathbf{X}_i)_{i \in \mathrm{pa}(j)}\right), \boldsymbol{\eta}_j\right), && j \in \mathcal{V}_{end}. \end{aligned}$$

Any SCBM $\mathfrak{C}$ straightforwardly induces an SCM $\mathfrak{A}(\mathfrak{C})$ in the classical sense (Pearl, 2009) by defining the mechanism functions as $m_j\left((\mathbf{X}_i)_{i \in \mathrm{pa}(j)}, \boldsymbol{\eta}_j\right) := f_j\left(b_j\left((\mathbf{X}_i)_{i \in \mathrm{pa}(j)}\right), \boldsymbol{\eta}_j\right)$. We call this SCM the induced SCM. Conversely, every SCM is the induced SCM of a trivial SCBM with $\mathcal{Z}_j = \prod_{i \in \mathrm{pa}(j)} \mathcal{X}_i$, $b_j = \mathrm{id}$ and $f_j = m_j$. Definition 2.1 by itself therefore imposes no restriction; dimension reduction enters only through the minimality condition of Lemma 3.2.

**Definition 2.2.** We call two SCBMs *structurally equivalent* if their induced SCMs coincide.

Since all interventional distributions and the observational distribution of $\mathbf{X}$ are fully determined by the induced SCM, structurally equivalent SCBMs share the same interventional distributions (interventional equivalence) and observational distribution (observational equivalence). In particular, every SCBM is structurally equivalent to the trivial SCBM of its induced SCM. The bottlenecks of interest are the minimal members of this equivalence class, which we characterize in Section 3.

For the remainder of this work, we consider so-called *factored SCBMs*, where each bottleneck can be subdivided into separate bottlenecks for each parent.

**Definition 2.3** (Factored SCBMs)**.** We call an SCBM $\mathfrak{C}$ *factored* if

(a) Each bottleneck is *factored* in the sense that $\mathcal{Z}_j$ and $b_j$ can be decomposed as $\mathcal{Z}_j = \prod_{i \in \mathrm{pa}(j)} \mathcal{Z}_{ij}$ and

$$b_j : \prod_{i \in \mathrm{pa}(j)} \mathcal{X}_i \to \prod_{i \in \mathrm{pa}(j)} \mathcal{Z}_{ij}$$
$$b_j((\mathbf{x}_i)_{i \in \mathrm{pa}(j)}) = (b_{ij}(\mathbf{x}_i))_{i \in \mathrm{pa}(j)}$$

with $\dim \mathcal{Z}_{ij} \leq d_i$. In other words, we assume that there is a separate bottleneck space for every parent of the endogenous node $j$.

(b) Each $\mathcal{X}_j$ is a vector space, and coincides with the noise space $\mathcal{H}_j = \mathcal{X}_j$. Each effect function $f_j$ is factored in the sense that there is a family of maps $f_{ij} : \mathcal{Z}_{ij} \to \mathcal{X}_j$ such that

$$\mathbf{X}_j := \sum_{i \in \mathrm{pa}(j)} f_{ij}(b_{ij}(\mathbf{X}_i)) + \boldsymbol{\eta}_j.$$

We illustrate Definition 2.3 in Figure 1; each edge in the DAG is endowed with a separate bottleneck.

**Equivalence relations induced by bottlenecks.** If $(i, j)$ is an edge in $\mathcal{G}$, then $b_{ij}$ can be considered a quotient map that induces an equivalence relation on $\mathcal{X}_i$: two states $\mathbf{x}, \mathbf{x}' \in \mathcal{X}_i$ are *bottleneck-equivalent relative to child* $j$ if $b_{ij}(\mathbf{x}) = b_{ij}(\mathbf{x}')$, which implies they are *causally equivalent* w.r.t. $\mathbf{X}_j$ in the sense of Chalupka et al. (2017):

$$P(\mathbf{X}_j \mid \mathrm{do}(\mathbf{X}_i = \mathbf{x})) = P(\mathbf{X}_j \mid \mathrm{do}(\mathbf{X}_i = \mathbf{x}')).$$

We discuss special cases of SCBMs, including interactions between random fields, temporal processes in the frequency domain, and a connection to the information bottleneck of Tishby et al. (2000), in Appendix B.

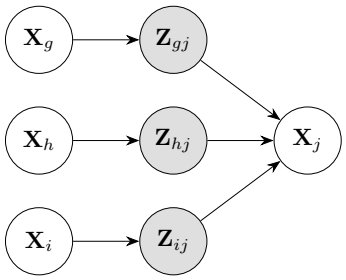


Figure 1: A *factored* SCBM.

## 3 Identifiability of SCBMs

We study identifiability of SCBMs, formalised as recovering bottleneck variables up to bijection from observational data—the degree of identifiability sufficient for causal effect estimation and intervention selection (cf. Section 2). We establish this result theoretically and validate it experimentally, illustrating how recovered bottlenecks yield meaningful partitions of the input space.

### 3.1 Identifiability theory

Inserting invertible maps on bottleneck spaces yields a structurally equivalent SCBM (Lemma 3.1); the converse holds for additive-noise SCBMs (Lemma 3.2). Proofs are in Appendix A.

**Lemma 3.1** (Bijective reparametrization preserves structural equivalence)**.** *Let* $\mathfrak{C} = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}, \mathcal{H}, \boldsymbol{\eta}, \mathcal{B}, \mathcal{F}, \mathbf{X} \rangle$ *be an SCBM. Assume that there are invertible maps* $\psi_j : \mathcal{Z}_j \to \mathcal{Z}'_j$ *for every endogeneous node* $j \in \mathcal{V}_{end}$ *to some other spaces* $\mathcal{Z}' = (\mathcal{Z}'_j)_{j \in \mathcal{V}_{end}}$*, and consider the functions* $b'_j = \psi_j \circ b_j$ *and* $f'_j(\cdot\,, \cdot) = f_j(\psi_j^{-1}(\,\cdot\,), \cdot)$*. Then* $\mathfrak{C}' = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}', \boldsymbol{\eta}, \mathcal{B}', \mathcal{F}', \mathbf{X} \rangle$ *with* $\mathcal{F}' = (f'_j)_j$ *and* $\mathcal{B}' = (b'_j)_j$ *is structurally equivalent to* $\mathfrak{C}$*.*

**Lemma 3.2** (Structural equivalence implies bijective bottlenecks)**.** *Let* $\mathfrak{C} = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}, \mathcal{H}, \boldsymbol{\eta}, \mathcal{B}, \mathcal{F}, \mathbf{X} \rangle$ *and* $\mathfrak{C}' = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}', \mathcal{H}', \boldsymbol{\eta}, \mathcal{B}', \mathcal{F}', \mathbf{X} \rangle$ *be two SCBMs with additive noises, i.e.* $f_j(\mathbf{z}_{\mathrm{pa}(j)}, \eta_j) = \tilde{f}_j(\mathbf{z}_{\mathrm{pa}(j)}) + \eta_j$*, and similarly for* $f'_j$*. Assume that the functions* $\tilde{f}_j : \mathcal{Z}_j \to \mathcal{X}_j$*,* $\tilde{f}'_j : \mathcal{Z}'_j \to \mathcal{X}_j$ *are almost surely injective. If* $\mathfrak{C}$ *and* $\mathfrak{C}'$ *are structurally equivalent, there is an invertible function* $\psi_j : \mathcal{Z}_j \to \mathcal{Z}'_j$ *such that* $b' = \psi \circ b$ *and* $f'_j = f_j \circ \psi^{-1}$*.*

Injectivity of the effect functions is a minimality assumption: together with surjectivity of $b_j$, it forces the bottleneck to retain exactly the information about the parents that the mechanism uses. In particular, it excludes the trivial SCBM with $b_j = \mathrm{id}$ unless $\tilde{f}_j$ is itself injective, in which case no compression is possible. In the linear case, it guarantees $\mathrm{rank}(f_j \circ b_j) = \dim \mathcal{Z}_j$, so the bottleneck dimension equals the rank of the mechanism. We refer to SCBMs satisfying this condition as *minimal*. The linear SCBMs in Section 4 are minimal by construction, since each $W_{ij}^f$ has full row rank. Within this class, bottlenecks are unique up to a change of basis (cf. Lemma 3.1).

### 3.2 Estimating bottlenecks

Given a graph, estimating each bottleneck function $b_{ij}$ amounts to fitting a standard regressor from $\mathbf{X}_i$ to $\mathbf{X}_j$ with a rank constraint on the bottleneck dimension, using any valid conditioning set for the effect $\mathbf{X}_i \to \mathbf{X}_j$. In the linear case this reduces to a rank-constrained matrix factorization; in the nonlinear case we use an encoder-decoder architecture. By Lemma 3.2, the recovered bottleneck satisfies $\hat{b}_{ij} = \psi_{ij} \circ b_{ij}$ for some invertible $\psi_{ij}$. Full details are in Appendix E.2; code is available at `https://github.com/simonbing/StructuralCausalBottleneckModels`.

We sample SCBMs with varying $|\mathcal{V}|$, $d_{\mathbf{X}}$, $d_{\mathbf{Z}}$, and mechanism type (linear/nonlinear), draw $n$ samples, estimate all bottlenecks, and measure recovery via the average bidirectional $R^2$ between each $\mathbf{Z}_{ij}$

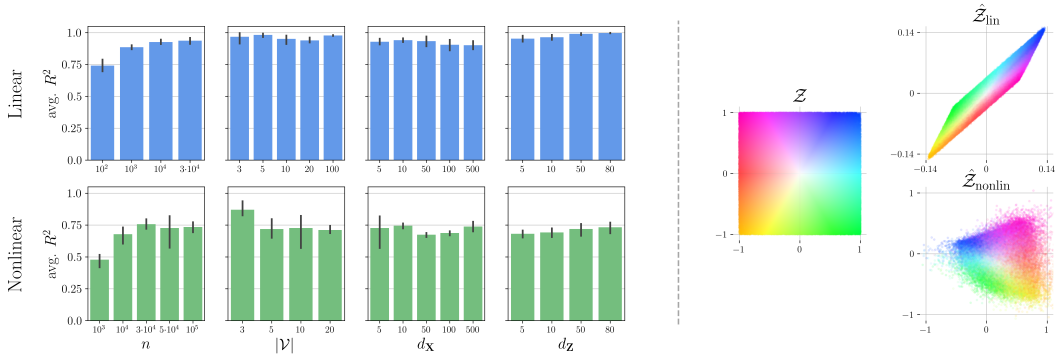


Figure 2: **Left:** Mean $\pm$ std of avg. $R^2$ for bottleneck recovery across settings; top row linear, bottom row nonlinear SCBMs. High scores confirm recovery up to bijection. **Right:** Learned bottleneck spaces $\hat{\mathcal{Z}}$ vs. ground-truth $\mathcal{Z}$, related by an affine map (linear) or smooth bijection (nonlinear).

and $\hat{\mathbf{Z}}_{ij}$; bidirectionality is required to test equivalence up to bijection rather than mere surjection. Default settings are $|\mathcal{V}| = 10$, $d_{\mathbf{X}} = 5$, $d_{\mathbf{Z}} = 2$, $n = 30000$ (linear) / 50000 (nonlinear); details are in Appendix E.1.

In linear SCBMs, recovery is successful across all settings, with performance saturating around $n = 10000$ and remaining robust as $|\mathcal{V}|$ scales, indicating minimal error propagation. For nonlinear SCBMs, scores are lower but degrade gracefully: larger $|\mathcal{V}|$ reduces performance without causing catastrophic failure,[2] and larger $n$ largely compensates. We also illustrate the learned bottleneck spaces in Figure 2: the invertible transformation $\psi$ (cf. Lemma 3.2) manifests as an affine map in the linear case and a smooth bijection in the nonlinear case. Additional visualizations are in Appendix D.1.

### 3.3 Estimation under misspecification

We now drop the assumption that $d_{\mathbf{Z}}$ is known and study how recovery degrades when the assumed bottleneck dimension $d_{\hat{\mathbf{Z}}}$ is misspecified. We sweep $d_{\hat{\mathbf{Z}}}$ from 1 to $d_{\mathbf{X}}$ on a cause-effect pair with $d_{\mathbf{Z}} = 10$ ($d_{\mathbf{X}} = 50$ linear, $d_{\mathbf{X}} = 100$ nonlinear). As shown in Figure 3, performance saturates at $d_{\hat{\mathbf{Z}}} = d_{\mathbf{Z}}$: over-compression loses information while under-compression does not hurt. This is a key practical advantage over CRL, where both under- *and* overestimating the latent dimension compromises identifiability.

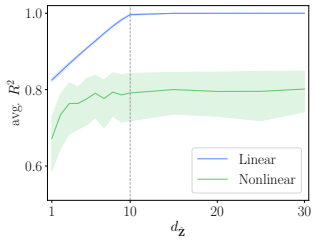


Figure 3: Mean $R^2 \pm 95\%$ CI for varying assumed bottleneck dimension $d_{\hat{\mathbf{Z}}}$; true $d_{\mathbf{Z}}$ indicated by dashed line.

### 3.4 Partitioning and intervention selection

We illustrate identifiability concretely using the three-node SCBM in Figure 4, where $\mathbf{X}_1 \in \mathbb{R}^2$ is a shared root with two children. The two bottlenecks capture fundamentally different properties of the same input: $b_{12}$ depends on the Euclidean norm, $b_{13}$ on the quadrant. Each bottleneck induces an equivalence relation on $\mathcal{X}_1$ via its level sets; since level sets are invariant to bijections, this partition is well-defined up to the identifiability class. The learned bottlenecks $\hat{b}_{12}$ and $\hat{b}_{13}$ recover the ground-truth partitions (concentric rings and quadrants, respectively), and projecting onto the output spaces reveals which input regions produce which downstream effects. Since bottleneck-equivalent inputs are causally equivalent (cf. Section 2), a practitioner need only consider one representative per class—reducing $\mathcal{X}_1$ to a one-dimensional scale for $\mathbf{X}_2$ and four classes for $\mathbf{X}_3$ without requiring interventional data.

[2] Graphs with $|\mathcal{V}| = 100$ have $\sim$3500 edges, each requiring a separate network; in practice one estimates only the bottlenecks needed for a specific query.

$$
\begin{aligned}
\mathbf{X}_1 &= \boldsymbol{\eta}_1, \\
\mathbf{X}_2 &= f_{12}(b_{12}(\mathbf{X}_1)) + \boldsymbol{\eta}_2, \\
\mathbf{X}_3 &= f_{13}(b_{13}(\mathbf{X}_1)) + \boldsymbol{\eta}_3,
\end{aligned}
$$

where

$$
b_{12} := \sqrt{{\mathbf{X}_1^{(1)}}^2 + {\mathbf{X}_1^{(2)}}^2},
$$

$$
b_{13} := \begin{cases} 0 & \text{if } \mathbf{X}_1^{(1)} \le 0, \mathbf{X}_1^{(2)} \le 0 \\ 1 & \text{if } \mathbf{X}_1^{(1)} \le 0, \mathbf{X}_1^{(2)} > 0 \\ 2 & \text{if } \mathbf{X}_1^{(1)} > 0, \mathbf{X}_1^{(2)} \le 0 \\ 3 & \text{if } \mathbf{X}_1^{(1)} > 0, \mathbf{X}_1^{(2)} > 0 \end{cases}.
$$

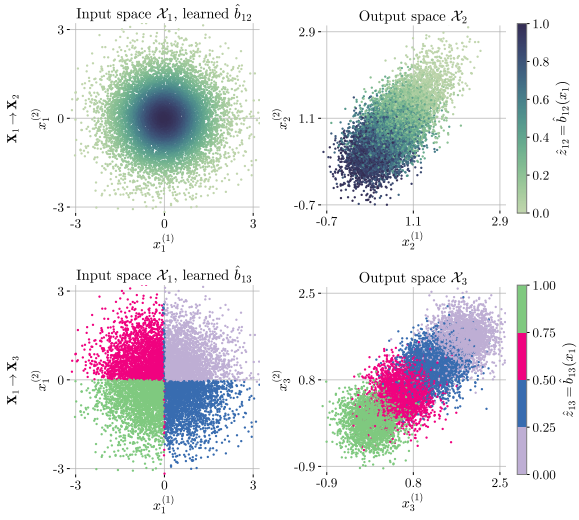


Figure 4: **Left:** Structural equations of a three-node SCBM with shared root $\mathbf{X}_1 \in \mathbb{R}^2$. **Right:** Input-space partitions (left column) and output spaces (right column) induced by $\hat{b}_{12}$ (top) and $\hat{b}_{13}$ (bottom). Colors denote bottleneck-equivalent, hence causally equivalent, states. The two bottlenecks induce fundamentally different partitions of the same $\mathbf{X}_1$—concentric rings vs. quadrants—invariant to bijective reparametrizations (Lemma 3.2). In the output spaces, only transitions across partition boundaries produce changes in the downstream distribution.

## 4 Effect estimation in SCBMs

High-dimensional variables introduce qualitatively new challenges for causal effect estimation: standard identification strategies can break down, and even when identification holds, the curse of dimensionality inflates finite-sample variance. We study three causal structures in which this occurs and show that the bottleneck structure of SCBMs provides a principled remedy in each case. In the *instrumental variable* (IV) setting, naïve 2SLS is generically non-identified when all variables are $d$-dimensional; oracle bottleneck representations restore identifiability. For *mediation* and *backdoor adjustment* without unobserved confounding, identification holds but bottleneck conditioning sets yield efficiency gains.

Throughout this section, we restrict to linear SCBMs, in which bottleneck and effect functions are linear maps, $b_{ij}(\mathbf{x}) = \mathbf{x}W_{ij}^b$ and $f_{ij}(\mathbf{z}) = \mathbf{z}W_{ij}^f$ with $W_{ij}^b \in \mathbb{R}^{d_i \times r}$ and $W_{ij}^f \in \mathbb{R}^{r \times d_j}$, where $r := \dim \mathcal{Z}_{ij}$, so that the mechanism matrix factors as $M_{ij} = W_{ij}^b W_{ij}^f$. The bottleneck assumption then reduces to a rank condition: $M_{ij}$ factors through an $r$-dimensional bottleneck space if and only if $\operatorname{rank}(M_{ij}) \le r$, and the bottleneck is minimal in the sense of Section 3 if and only if $W_{ij}^b$ and $W_{ij}^f$ both have rank $r$, i.e. $\operatorname{rank}(M_{ij}) = r$. Minimal factorizations are unique up to $W_{ij}^b \mapsto W_{ij}^b G$, $W_{ij}^f \mapsto G^{-1} W_{ij}^f$ for invertible $G \in \mathbb{R}^{r \times r}$, the linear instance of Lemmas 3.1 and 3.2. Our target is the direct effect $M_{XY}$.

It is useful to separate four ingredients of our framework. (i) The *modelling assumption* is structural: data is generated by a factored SCBM with additive noise and minimal bottlenecks (Definitions 2.1, 2.3; Lemma 3.2) over a known graph $\mathcal{G}$, here specialized to the linear case. (ii) The *identifiability target* is the family of bottleneck functions $b_{ij}$ up to invertible reparametrizations $\psi_{ij}$ (Lemma 3.2), i.e. what the observational distribution determines in principle. (iii) The *estimation procedure* (Section 3.2, Algorithm 1) consists of rank-constrained regression or an encoder–decoder with assumed bottleneck dimension $d_{\hat{\mathbf{Z}}}$ and a chosen conditioning set; these are choices of the practitioner rather than assumptions on the model, and may be misspecified (Section 3.3). (iv) The *oracle quantities* are the true bottleneck matrices $W_{ij}^b$, to which we assume access throughout this section. Since all results below are invariant under the reparametrization $W_{ij}^b \mapsto W_{ij}^b G$, the oracle only needs to supply bottlenecks up to the identifiability class in (ii). Working with oracle bottlenecks isolates their statistical benefit from the cost of learning them, which we quantify in Appendix D.5.

Accordingly, experiments in this section use oracle bottlenecks unless stated otherwise, whereas those in Section 3 use estimated ones. Full proofs are deferred to Appendix A.2; proof sketches are presented for central results.

### 4.1 Unobserved confounding

When treatment $\mathbf{X}$ and outcome $\mathbf{Y}$ are confounded by an unobserved variable $\mathbf{U}$, standard regression is biased and an instrumental variable $\mathbf{I}$ can enable identification. In the classical scalar setting, two-stage least squares (2SLS) with a valid instrument recovers the causal effect consistently. In the high-dimensional SCBM setting, however, the bottleneck structure renders $M_{IX}$ rank-deficient, so naïve 2SLS is underidentified even with a valid instrument. This is an instance of the known rank-condition failure of linear IV, under which 2SLS recovers at best a projection of the effect (Pfister and Peters, 2022; Ailer et al., 2023). Since here this failure stems from the mechanism rather than from redundant instruments, it cannot be resolved by instrument selection (Brito and Pearl, 2002). Whereas prior remedies rely on additional experiments or sparsity, we show that a bottleneck factorization of the treatment–outcome mechanism restores identification given the bottlenecks (Theorem 4.3).

**Proposition 4.1** (Non-identifiability of naïve 2SLS in linear SCBMs)**.** *Let $\mathbf{I}, \mathbf{X}, \mathbf{Y}, \mathbf{U} \in \mathbb{R}^d$ be the instrument, treatment, outcome, and unobserved confounder, respectively, related by the linear structural equations*

$$\mathbf{I} = \boldsymbol{\eta}_I, \tag{1}$$
$$\mathbf{U} = \boldsymbol{\eta}_U, \tag{2}$$
$$\mathbf{X} = \mathbf{I}\, M_{IX} + \mathbf{U}\, M_{UX} + \boldsymbol{\eta}_X, \tag{3}$$
$$\mathbf{Y} = \mathbf{X}\, M_{XY} + \mathbf{U}\, M_{UY} + \boldsymbol{\eta}_Y, \tag{4}$$

*where the mechanism matrix $M_{IX} \in \mathbb{R}^{d\times d}$ admits the SCBM factorization*

$$M_{IX} \;=\; W^b_{IX}\, W^f_{IX}, \qquad W^b_{IX} \in \mathbb{R}^{d\times r}, \quad W^f_{IX} \in \mathbb{R}^{r\times d},$$

*with $W^b_{IX}$ and $W^f_{IX}$ each of rank $r$, so that $\operatorname{rank}(M_{IX}) = r < d$. Assume $\mathbf{I} \perp (\mathbf{U}, \eta_X, \eta_Y)$ and $\Sigma_I := \mathbb{E}[\mathbf{I}^\top \mathbf{I}] \succ 0$. Then the population 2SLS estimator for $M_{XY}$ has no unique solution.*

The non-identifiability is a direct consequence of the bottleneck structure: the rank deficiency of $M_{IX}$ makes the 2SLS Gram matrix $G$ singular with a $(d-r)$-dimensional null space, so the normal equations are satisfied by infinitely many solutions. The following remark characterizes what any consistent regularized variant of 2SLS actually converges to.

*Remark* 4.2 (Limiting bias of regularized 2SLS). Any consistent regularized variant of 2SLS whose regularization is shrunk to zero—such as ridge 2SLS with penalty $\lambda \to 0^+$—converges to the minimum-norm solution of $G\,\Gamma \;=\; G\, M_{XY}$, namely

$$\Gamma^* \;=\; G^{+} G\, M_{XY} \;=\; P_{\mathrm{row}(M_{IX})}\, M_{XY},$$

where $P_{\mathrm{row}(M_{IX})}$ is the orthogonal projector onto $\mathrm{row}(M_{IX})$. In this SCBM, $\mathrm{row}(M_{IX}) = \mathrm{row}(W^f_{IX})$ and $\mathrm{col}(M_{XY}) = \mathrm{col}(W^b_{XY})$ are generically distinct $r$-dimensional subspaces of $\mathbb{R}^d$, so $\Gamma^* \neq M_{XY}$ with probability one whenever the weight matrices are drawn from any distribution that is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{d\times r}$.

This makes explicit, in the SCBM, that regularized 2SLS converges not to $M_{XY}$ but to its projection onto $\mathrm{row}(M_{IX})$—a biased target whenever the instrument–treatment and treatment–outcome subspaces differ, which holds generically. The root cause is the mismatch between the $d$-dimensional regression space and the effective $r$-dimensional signal. The minimal bottleneck factorization makes this low-rank structure explicit: projecting both the instrument and the treatment onto their respective $r$-dimensional bottleneck spaces yields a well-posed first stage and restores identifiability.

**Theorem 4.3** (Identification via oracle bottleneck 2SLS)**.** *Under the setup of Proposition 4.1, additionally let $M_{XY} = W^b_{XY} W^f_{XY}$ with $W^b_{XY} \in \mathbb{R}^{d\times r}$, $W^f_{XY} \in \mathbb{R}^{r\times d}$, each of rank $r$. Suppose the rank condition*

$$\operatorname{rank}\Big(W^f_{IX}\, W^b_{XY}\Big) = r \tag{RC}$$

*holds. Given oracle access to $W^b_{IX}$ and $W^b_{XY}$, define the $r$-dimensional bottleneck representations*

$$\mathbf{Z}_{IX} \;:=\; \mathbf{I}\, W^b_{IX}, \qquad \mathbf{Z}_{XY} \;:=\; \mathbf{X}\, W^b_{XY}.$$

*The oracle bottleneck 2SLS procedure regresses $\mathbf{Z}_{XY}$ on $\mathbf{Z}_{IX}$ (first stage) to obtain fitted values $\hat{\mathbf{Z}}_{XY}$, then regresses $\mathbf{Y}$ on $\hat{\mathbf{Z}}_{XY}$ (second stage) to obtain $\hat{\Phi}$, and forms the estimator*

$$\widehat{M}_{XY} \;:=\; W^b_{XY}\,\hat{\Phi}.$$

*Then $\widehat{M}_{XY} = M_{XY}$.*

*Proof sketch.* Since $W^b_{IX}$ has rank $r$ and $\Sigma_I \succ 0$, the first-stage Gram matrix $\Sigma_{Z_{IX}} \succ 0$ gives a unique first stage; rank condition (RC) then ensures the second-stage Gram matrix is positive definite, yielding $\hat{\Phi} = W^f_{XY}$. Composing gives $\widehat{M}_{XY} = W^b_{XY}\hat{\Phi} = M_{XY}$. □

The two oracle bottlenecks in Theorem 4.3 differ in how they can be obtained. Since the instrument is unconfounded, the first-stage regression of $\mathbf{X}$ on $\mathbf{I}$ identifies $M_{IX}$, and a rank-$r$ factorization yields $W^b_{IX}$ up to the reparametrization $G$. The bottleneck $W^b_{XY}$, by contrast, belongs to the confounded mechanism $\mathbf{X} \to \mathbf{Y}$. The IV moment conditions only determine its projection $P_{\mathrm{row}(M_{IX})}M_{XY}$ (Remark 4.2), so it cannot be learned from the observational data alone. In practice, $W^b_{XY}$ must therefore come from a transfer environment in which the $\mathbf{X} \to \mathbf{Y}$ mechanism is invariant, such as a pilot RCT intervening on $\mathbf{X}$ (Appendix D.3).

Related approaches are Bai and Ng (2010), who address finite-sample many-instruments bias by compressing many instruments via PCA, and Lin et al. (2026), who identify a backdoor bias from omitting instrument-relevant treatment directions and recover these via SVD. Both works face a bias rather than a population-level non-identifiability, and neither assumes a bottleneck factorization of the treatment–outcome mechanism $M_{XY}$; see Appendix C for a detailed comparison.

### 4.2 No unobserved confounding

When all confounders are observed, the causal effect is identifiable by standard regression. The question is one of *efficiency*: conditioning on high-dimensional observed variables inflates finite-sample variance, and lower-dimensional bottleneck representations offer a more parsimonious alternative. The results below establish which bottleneck representations are valid adjustment variables in each setting, and show that replacing the observed variable with its bottleneck reduces asymptotic variance. We consider two settings: mediator adjustment and backdoor adjustment; see Figure 5.

#### 4.2.1 Mediator adjustment

Let $\mathbf{X} \in \mathbb{R}^d$ be the treatment, $\mathbf{M} \in \mathbb{R}^d$ an observed mediator, and $\mathbf{Y} \in \mathbb{R}^d$ the outcome. The linear SCBM structural equations are

$$\mathbf{X} = \boldsymbol{\eta}_X, \tag{5}$$
$$\mathbf{M} = \mathbf{X}M_{XM} + \boldsymbol{\eta}_M, \tag{6}$$
$$\mathbf{Y} = \mathbf{X}M_{XY} + \mathbf{M}M_{MY} + \boldsymbol{\eta}_Y, \tag{7}$$

where $M_{ij} = W^b_{ij}W^f_{ij} \in \mathbb{R}^{d\times d}$ with $\mathrm{rank}(M_{ij}) = r \ll d$, and $\Sigma_{\boldsymbol{\eta}_X}, \Sigma_{\boldsymbol{\eta}_M}, \Sigma_{\boldsymbol{\eta}_Y} \succ 0$ are the noise covariances, all mutually independent. The target is the *direct effect* $M_{XY}$. Define the *outcome-side mediator bottleneck* $\mathbf{Z}_{MY} := \mathbf{M}W^b_{MY} \in \mathbb{R}^{n\times r}$.

The graph contains two bottleneck representations on the indirect path $\mathbf{X} \to \mathbf{M} \to \mathbf{Y}$: one on the $\mathbf{X} \to \mathbf{M}$ edge and one on the $\mathbf{M} \to \mathbf{Y}$ edge. The following lemma resolves which of these bottlenecks can replace the full mediator $\mathbf{M}$ as a conditioning variable.

**Lemma 4.4** (Validity of mediator bottleneck adjustment)**.** *Let $\hat{M}^S_{XY}$ be the OLS estimator of $M_{XY}$ from the regression of $\mathbf{Y}$ on $(\mathbf{X}, S)$. Both $S = \mathbf{M}$ and $S = \mathbf{Z}_{MY}$ yield consistent estimators:*

$$\hat{M}^M_{XY} \xrightarrow{p} M_{XY}, \qquad \hat{M}^{Z_{MY}}_{XY} \xrightarrow{p} M_{XY}.$$

Both $\mathbf{M}$ and $\mathbf{Z}_{MY}$ are therefore consistent conditioning variables. Since $\mathbf{Z}_{MY}$ is $r$-dimensional while $\mathbf{M}$ is $d$-dimensional with $r \ll d$, conditioning on $\mathbf{Z}_{MY}$ requires far fewer effective degrees of freedom. This translates into a reduction in asymptotic variance.

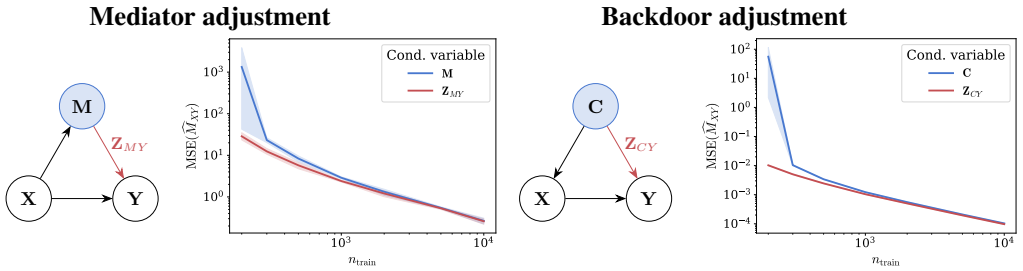


Figure 5: **Left:** Mediator adjustment; **Right:** backdoor adjustment. MSE of $\widehat{M}_{XY}$ vs. $n_{\text{train}}$ (log-log) for conditioning on the full high-dimensional variable (blue, $d = 100$) vs. the bottleneck (red, $d_{\mathbf{Z}} = 5$). Bottleneck conditioning yields a consistent finite-sample advantage, with both estimators converging for large $n_{\text{train}}$. Results of additional parameter configurations provided in Appendix D.2.

**Proposition 4.5** (Efficiency gain of mediator bottleneck adjustment)**.** *Under the linear SCBM, the asymptotic covariance of* $\hat{M}^S_{XY}$ *is* $\Phi_S := Q_S^{-1}\Sigma_{\boldsymbol{\eta}_Y}Q_S^{-1}$*, where* $Q_S := \mathrm{Cov}(\tilde{\mathbf{X}}^S)$ *denotes the covariance of the Frisch–Waugh–Lovell (Frisch and Waugh, 1933; Lovell, 1963) (FWL) residual of* $\mathbf{X}$ *after projecting out* $S$*. Then*

$$\Phi_{Z_{MY}} \preceq \Phi_M.$$

We provide experimental evidence that underlines the efficiency gains of $\mathbf{Z}_{MY}$ in Figure 5 (left).

The efficiency gain of $\mathbf{Z}_{MY}$ over $\mathbf{M}$ may suggest that any low-dimensional bottleneck along the indirect path is beneficial; in particular, one might ask whether the treatment-side bottleneck $\mathbf{Z}_{XM} := \mathbf{X}W^b_{XM}$ is also a valid adjustment variable. It is not: since $\mathbf{Z}_{XM}$ is a linear function of $\mathbf{X}$, regressing $\mathbf{Y}$ on $(\mathbf{X}, \mathbf{Z}_{XM})$ is equivalent to regressing on $\mathbf{X}$ alone, so the estimator converges to the total effect $M_{XY} + M_{XM}M_{MY}$, instead of the direct effect $M_{XY}$. A valid conditioning variable must lie on the $\mathbf{M} \to \mathbf{Y}$ side of the indirect path—as $\mathbf{Z}_{MY}$ does.

### 4.2.2 Backdoor adjustment

Let $\mathbf{C}, \mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$ be confounder, treatment, and outcome satisfying the linear structural equations

$$\mathbf{C} = \boldsymbol{\eta}_C, \tag{8}$$
$$\mathbf{X} = \mathbf{C}\, M_{CX} + \boldsymbol{\eta}_X, \tag{9}$$
$$\mathbf{Y} = \mathbf{X}\, M_{XY} + \mathbf{C}\, M_{CY} + \boldsymbol{\eta}_Y, \tag{10}$$

where $M_{CX} = W^b_{CX}W^f_{CX}$ and $M_{CY} = W^b_{CY}W^f_{CY}$ with $W^b_{CX}, W^b_{CY} \in \mathbb{R}^{d\times r}$ each of rank $r < d$, mutually independent noise terms, and $\Sigma_{\eta_X}, \Sigma_{\eta_Y} \succ 0$. The target is $M_{XY}$. Define the bottleneck representations $\mathbf{Z}_{CX} := \mathbf{C}\, W^b_{CX} \in \mathbb{R}^{n\times r}$ and $\mathbf{Z}_{CY} := \mathbf{C}\, W^b_{CY} \in \mathbb{R}^{n\times r}$.

Unlike the mediator setting, the confounder $\mathbf{C}$ has two outgoing bottlenecks: $\mathbf{Z}_{CX}$ captures the part of $\mathbf{C}$ that drives treatment, and $\mathbf{Z}_{CY}$ captures the part that drives outcome. Both turn out to be valid adjustment variables for $M_{XY}$, though for different reasons.

**Lemma 4.6** (Bottleneck adjustment validity)**.** *Let* $\mathbf{Z}$ *be an* $r$*-dimensional bottleneck representation of* $\mathbf{C}$*. The population OLS estimator of* $\mathbf{Y}$ *on* $(\mathbf{X}, \mathbf{Z})$ *consistently identifies* $M_{XY}$ *if* $\mathbf{Z}$ *satisfies either of the following conditions:*

*(a)* ***Outcome sufficiency:*** $\mathbf{C}\, M_{CY}$ *is a measurable function of* $\mathbf{Z}$*, i.e. the confounding path* $\mathbf{C} \to \mathbf{Y}$ *factors through* $\mathbf{Z}$*;*

*(b)* ***Treatment sufficiency:*** $\mathbf{X} \perp \mathbf{C} \mid \mathbf{Z}$*, i.e.* $\mathbf{Z}$ *screens off* $\mathbf{C}$ *from* $\mathbf{X}$*.*

*In the linear SCBM,* $\mathbf{Z}_{CY}$ *satisfies* (a) *and* $\mathbf{Z}_{CX}$ *satisfies* (b)*.*

Both bottleneck representations are valid, but they need not be equally efficient. Since $\mathbf{Z}_{CY}$ directly explains the confounding path $\mathbf{C} \to \mathbf{Y}$, while $\mathbf{Z}_{CX}$ only screens off $\mathbf{C}$ from $\mathbf{X}$ without fully capturing the outcome-side variation, we might expect $\mathbf{Z}_{CY}$ to be the more efficient conditioning variable. The following proposition confirms this and establishes the complete efficiency ordering.

**Proposition 4.7** (Efficiency ordering of bottleneck adjustment)**.** *Let $\hat{M}^{C}_{XY}$, $\hat{M}^{CX}_{XY}$, and $\hat{M}^{CY}_{XY}$ denote the OLS estimators of $M_{XY}$ obtained by conditioning on $\mathbf{C}$, $\mathbf{Z}_{CX}$, and $\mathbf{Z}_{CY}$, respectively. All three are consistent. Defining $\tilde{\mathbf{V}}^{S} := \mathbf{V} - \mathbb{E}[\mathbf{V} \mid S]$ as the population residual and $\delta_S := \tilde{\mathbf{Y}}^{S} - \tilde{\mathbf{X}}^{S} M_{XY}$ as the effective regression noise, the asymptotic variance of each estimator is characterized by*

$$\Phi_S \;:=\; Q_S^{-1}\,\Sigma_{\delta_S}\,Q_S^{-1}, \qquad Q_S \;:=\; \mathrm{Cov}(\tilde{\mathbf{X}}^{S}), \quad \Sigma_{\delta_S} \;:=\; \mathrm{Cov}(\delta_S),$$

*and satisfies $\Phi_{CY} \prec \Phi_{C} \prec \Phi_{CX}$, i.e.*

$$\mathrm{Avar}\left(\hat{M}^{CY}_{XY}\right) \;\prec\; \mathrm{Avar}\left(\hat{M}^{C}_{XY}\right) \;\prec\; \mathrm{Avar}\left(\hat{M}^{CX}_{XY}\right),$$

*with both strict inequalities holding whenever $\ker(W^{b,\top}_{CY}) \cap \mathrm{col}(M^{\top}_{CX}) \neq \{0\}$ and $\ker(W^{b,\top}_{CX}) \cap \mathrm{col}(W^{b}_{CY}) \neq \{0\}$, respectively.*

Figure 5 (right) shows the gains of the outcome-side bottleneck over $C$.

The efficiency ordering $\Phi_{CY} \prec \Phi_{C} \prec \Phi_{CX}$ reveals an important asymmetry. $\mathbf{Z}_{CX}$ plays the role of a *propensity score* (Rosenbaum and Rubin, 1983): it screens off $\mathbf{C}$ from $\mathbf{X}$ but leaves residual confounding in the outcome regression, mirroring the efficiency loss of propensity-score adjustment relative to conditioning on the full confounder (Hahn, 1998; Hirano et al., 2003). $\mathbf{Z}_{CY}$, by directly explaining the path $\mathbf{C} \to \mathbf{Y}$, is both valid and strictly more efficient than conditioning on $\mathbf{C}$ itself.

More generally, condition (a) of Lemma 4.6 only requires the confounding path to factor through $\mathbf{Z}$, so any representation from which $\mathbf{Z}_{CY}$ can be recovered, including non-minimal ones (cf. Section 3.3), remains a valid adjustment variable. Such representations pay an efficiency cost relative to $\mathbf{Z}_{CY}$ (cf. proof of Proposition 4.5), with $\mathbf{C}$ itself as the extreme case of Proposition 4.7. Finally, validity for a given task does not imply equivalence: $\mathbf{Z}_{CX}$ and $\mathbf{Z}_{CY}$ are both valid but generically capture different subspaces of $\mathbf{C}$.

## 5 Discussion

We introduce structural causal bottleneck models, a novel family of graphical causal models for vector-valued variables, in which effects between high-dimensional variables are mediated by low-dimensional bottlenecks. We prove that bottlenecks are identifiable up to bijection from observational data, show that the bottleneck structure provides a principled remedy for three qualitatively different challenges in high-dimensional causal effect estimation, and validate these results experimentally. Our estimation method is relatively simple compared to most CRL and causal abstraction approaches, which we hope provides robustness to assumption violations in real data (Gamella* et al., 2025).

**Limitations.** The primary limitation of our effect estimation results is the oracle assumption of access to the true bottleneck matrices $W^{b}_{ij}$. Two remedies arise naturally. First, *transfer environments*—pilot RCTs, pre-intervention baselines, or domain shifts—can serve as held-out data from which bottlenecks are estimated before use in the target environment, as we demonstrate for the IV setting in Appendix D.3. Second, pre-trained models may provide approximate bottlenecks without additional interventional data: a representation that contains the bottleneck remains a valid adjustment variable at an efficiency cost (Section 4.2.2), whereas a predictive representation that misses part of it does not, and whether encoders trained for prediction achieve this containment is an open question. Beyond the oracle assumption, our effect estimation theory is restricted to linear SCBMs and, like virtually all causal effect estimation methods, requires knowledge of the causal graph. Finally, the efficiency gains of bottleneck adjustment vanish under even small full-rank perturbations of the mechanisms (Appendix D.4); if low rank cannot be justified, one should adjust for the full variables, which remain valid.

**Outlook.** This is an introductory work that establishes a novel modeling framework, leaving several directions for future work. A formal finite-sample characterization of the efficiency gains from using bottleneck conditioning sets would complement our asymptotic consistency results. Extending the IV non-identifiability result and its bottleneck-based remedy to nonlinear SCBMs is another open direction. A natural extension lies in developing application specific estimators, as our methodology generally permits fitting any model compatible with in- and output data modalities, without requiring a bespoke causal loss. Finally, investigating whether the bottleneck assumption can be exploited to assist causal discovery—partially or fully learning the graph we currently assume known—is a promising direction for future work.

## Acknowledgements

S.B. and J.R. received funding for this project from the European Research Council (ERC) Starting Grant CausalEarth under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 948112). S.B. received support from the German Academic Scholarship Foundation. J.W. received support from the German Federal Ministry of Research, Technology, and Space (BMFTR) as part of the project MAC-MERLin (Grant Agreement No. 01IW24007) and from the European Regional Development Fund (ERDF) and the German Federal State of Saarland within the scope of (To)CERTAIN (Project ID EFRE-AuF-0000942). This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID 1083.

## References

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional Causal Representation Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 372–407, 2023.

Elisabeth Ailer, Jason Hartford, and Niki Kilbertus. Sequential underspecified instrument selection for cause-effect estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 408–420. PMLR, 2023.

Mikio Aoi and Jonathan W Pillow. Model-based targeted dimensionality reduction for neuronal population data. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jushan Bai and Serena Ng. Instrumental variable estimation in a data-rich environment. *Econometric Theory*, 26(6):1577–1606, 2010. doi: 10.1017/S0266466609990727.

Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying Linearly-Mixed Causal Representations from Multi-Node Interventions. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, pages 843–867, 2024.

Arne Bjerhammar. Application of calculus of matrices to method of least squares: with special reference to geodetic calculations. *Kungl. Tekniska Högskolans handlingar*, 49, 1951.

Matthew Blackwell, Stefano Iacus, Gary King, and Giuseppe Porro. cem: Coarsened exact matching in stata. *The stata journal*, 9(4):524–546, 2009.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, 35, pages 38319–38331, 2022.

Carlos Brito and Judea Pearl. Generalized instrumental variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 85–93, 2002.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems*, volume 36, pages 45419–45462, 2023.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44:137–164, 2017.

Gabriele D'Acunto, Fabio Massimo Zennaro, Yorgos Felekis, and Paolo Di Lorenzo. Causal abstraction learning based on the semantic embedding principle. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 11776–11812. PMLR, 2025.

Paul Erdős and Alfréd Rényi. On Random Graphs I. *Publicationes Mathematicae Debrecen*, 6: 290–297, 1959.

Yorgos Felekis, Fabio Massimo Zennaro, Nicola Branchini, and Theodoros Damoulas. Causal optimal transport of abstractions. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, pages 462–498, 2024.

Ragnar Frisch and Frederick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933. doi: 10.2307/1907330.

Juan L. Gamella*, Simon Bing*, and Jakob Runge. Sanity checking causal representation learning on a simple real-world system. In *International Conference on Machine Learning*, pages 18143–18169. PMLR, 2025.

Amir Globerson and Naftali Tishby. Sufficient dimensionality reduction. *Journal of Machine Learning Research*, 3:1307–1331, 2003.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–332, 1998. doi: 10.2307/2998560.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003. doi: 10.1111/1468-0262.00442.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

Frederik Hytting Jørgensen, Luigi Gresele, and Sebastian Weichwald. What is causal about causal models and representations? *arXiv preprint arXiv:2501.19335*, 2025.

Paul Kroeger, Kevin Xia, and Elias Bareinboim. Learning invariances for causal abstraction inference. Technical Report R-129, Causal Artificial Intelligence Laboratory, Columbia University, May 2025. URL https://causalai.net/r129.pdf.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *Journal of Machine Learning Research*, 27(71):1–90, 2026.

Steffen L Lauritzen and Thomas S Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):321–348, 2002.

Bing Li. *Sufficient dimension reduction: Methods and applications with R*. Chapman and Hall/CRC, 2018.

Lexin Li. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shiangyi Lin, Hui Lan, and Vasilis Syrgkanis. Learning treatment representations for downstream instrumental variable regression. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 13557–13603, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *International Conference on Learning Representations*, 2023a.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. BISCUIT: Causal Representation Learning from Binary Interactions. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, pages 1263–1273, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Michael C. Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963. doi: 10.1080/01621459.1963.10480682.

Francisco Madaleno, Pratik Misra, and Alex Markham. Coarsening causal DAG models. In *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pages 1318–1344. PMLR, 2026.

Riccardo Massidda, Sara Magliacane, and Davide Bacciu. Learning causal abstractions of linear structural causal models. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Eliakim H Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the american mathematical society*, 26:294–295, 1920.

Urmi Ninad, Jonas Wahl, Andreas Gerhardus, and Jakob Runge. Causal discovery on vector-valued variables and consistency-guided aggregation. *arXiv preprint arXiv:2505.10476*, 2025.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Roger Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

Niklas Pfister and Jonas Peters. Identifiability of sparse causal effects using instrumental variables. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1613–1622. PMLR, 2022.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Weidmantel, and Bernhard Schölkopf. Causal consistency of structural equation models. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, volume 73 of *Proceedings of Machine Learning Research*, pages 563–572. AUAI Press, 2017.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Felix Schur, Pio Blieske, and Jonas Peters. DecoR: Deconfounding time series with robust regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(3):799–818, 2026. doi: 10.1093/jrsssb/qkaf067.

Francisco N. F. Q. Simoes, Mehdi Dastani, and Thijs van Ommen. The causal information bottleneck and optimal causal variable abstractions. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 3878–3897. PMLR, 2025.

Chandler Squires, Anna Seigal, Salil S. Bhate, and Caroline Uhler. Linear Causal Disentanglement via Interventions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 32540–32560, 2023.

Xavier-Andoni Tibau, Christian Reimers, Andreas Gerhardus, Joachim Denzler, Veronika Eyring, and Jakob Runge. A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections. *Environmental Data Science*, 1:e12, 2022.

Axel Timmermann, Soon-Il An, Jong-Seong Kug, Fei-Fei Jin, Wenju Cai, Antonietta Capotondi, Kim M Cobb, Matthieu Lengaigne, Michael J McPhaden, Malte F Stuecker, et al. El niño–southern oscillation complexity. *Nature*, 559(7715):535–545, 2018.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Jonas Wahl, Urmi Ninad, and Jakob Runge. Foundations of causal discovery on groups of variables. *Journal of Causal Inference*, 12(1):20230041, 2024.

Kevin Xia and Elias Bareinboim. Neural causal abstractions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20585–20595, 2024.

Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, 2024.

Fabio Massimo Zennaro, Máté Drávucz, Geanina Apachitei, W Dhammika Widanage, and Theodoros Damoulas. Jointly learning consistent causal abstractions over multiple interventional distributions. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, pages 88–121, 2023.

# Appendix

## Table of Contents


**A Proofs** **16**

A.1 Proofs for Section 3 (Identifiability of SCBMs) . . . . . 16

A.2 Proofs for Section 4 (Effect estimation in SCBMs) . . . . . 16

A.2.1 Proofs for Section 4.1 (Unobserved confounding) . . . . . 16

A.2.2 Proofs for Section 4.2.1 (Mediator adjustment) . . . . . 17

A.2.3 Proofs for Section 4.2.2 (Backdoor adjustment) . . . . . 18

**B Special cases and interpretations** **19**

**C Detailed related work** **21**

**D Additional experimental results** **23**

D.1 Identifiability . . . . . 23

D.2 Effect estimation . . . . . 23

D.3 Transfer environments for IV estimation . . . . . 26

D.4 Approximately low-rank mechanisms . . . . . 26

D.5 Sample cost of estimating bottlenecks . . . . . 28

**E Experimental details** **29**

E.1 General data generation process . . . . . 29

E.2 Detailed estimation procedure . . . . . 29

E.3 Data generation details & hyperparameters . . . . . 31

E.3.1 Sampling nonlinear bottleneck and effect functions . . . . . 31

E.3.2 Estimation under misspecification . . . . . 31

E.3.3 Partitioning and intervention selection . . . . . 31

E.3.4 Adjustment experiments . . . . . 32

E.3.5 Approximately Low-Rank Mechanisms . . . . . 32

E.4 Compute resources . . . . . 32

**F On randomly sampling nonlinear functions** **33**

**G Declaration of LLM usage** **33**

# A Proofs

## A.1 Proofs for Section 3 (Identifiability of SCBMs)

**Lemma 3.1.** *Let $\mathfrak{C} = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}, \mathcal{H}, \boldsymbol{\eta}, \mathcal{B}, \mathcal{F}, \mathbf{X} \rangle$ be an SCBM. Assume that there are invertible maps $\psi_j : \mathcal{Z}_j \to \mathcal{Z}'_j$ for every endogeneous node $j \in \mathcal{V}_{end}$ to some other spaces $\mathcal{Z}' = (\mathcal{Z}'_j)_{j \in \mathcal{V}_{end}}$, and consider the functions $b'_j = \psi_j \circ b_j$ and $f'_j(\cdot\,,\,\cdot) = f_j(\psi_j^{-1}(\,\cdot\,),\,\cdot)$. Then $\mathfrak{C}' = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}', \boldsymbol{\eta}, \mathcal{B}', \mathcal{F}', \mathbf{X} \rangle$ with $\mathcal{F}' = (f'_j)_j$ and $\mathcal{B}' = (b'_j)_j$ is structurally equivalent to $\mathfrak{C}$.*

*Proof of Lemma 3.1.* The result follows directly from the fact that the structural equations in both models coincide:

$$\begin{aligned} m'_j\left((\mathbf{X}_i)_{i \in \mathrm{pa}(j)}, \boldsymbol{\eta}_j\right) &= f'_j\left(b'_j\left((\mathbf{X}_i)_{i \in \mathrm{pa}(j)}\right), \boldsymbol{\eta}_j\right) \\ &= f_j(\psi_j^{-1}(\psi_j\left(b_j\left((\mathbf{X}_i)_{i \in \mathrm{pa}(j)}\right)\right) \cdot\,), \eta_j) \\ &= f_j\left(b_j\left((\mathbf{X}_i)_{i \in \mathrm{pa}(j)}\right), \boldsymbol{\eta}_j\right) \\ &= m_j\left((\mathbf{X}_i)_{i \in \mathrm{pa}(j)}, \boldsymbol{\eta}_j\right) \end{aligned}$$

□

**Lemma 3.2.** *Let $\mathfrak{C} = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}, \mathcal{H}, \boldsymbol{\eta}, \mathcal{B}, \mathcal{F}, \mathbf{X} \rangle$ and $\mathfrak{C}' = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}', \mathcal{H}', \boldsymbol{\eta}, \mathcal{B}', \mathcal{F}', \mathbf{X} \rangle$ be two SCBMs with additive noises, i.e. $f_j(\mathbf{z}_{\mathrm{pa}(j)}, \eta_j) = \tilde{f}_j(\mathbf{z}_{\mathrm{pa}(j)}) + \eta_j$, and similarly for $f'_j$. Assume that the functions $\tilde{f}_j : \mathcal{Z}_j \to \mathcal{X}_j$, $\tilde{f}'_j : \mathcal{Z}'_j \to \mathcal{X}_j$ are almost surely injective. If $\mathfrak{C}$ and $\mathfrak{C}'$ are structurally equivalent, there is an invertible function $\psi_j : \mathcal{Z}_j \to \mathcal{Z}'_j$ such that $b' = \psi \circ b$ and $f'_j = f_j \circ \psi^{-1}$.*

*Proof of Lemma 3.2.* Since the models are structurally equivalent and the noise terms are the same in both models, it follows directly that $f_j \circ b_j = f'_j \circ b'_j$ $P_{\mathbf{X}}$-almost surely. In particular the maps $f_j, f'_j$ have the same range, so that the map $\psi_j = f_j'^{-1} \circ F_j$ is well-defined and has the desired properties. □

## A.2 Proofs for Section 4 (Effect estimation in SCBMs)

### A.2.1 Proofs for Section 4.1 (Unobserved confounding)

**Proposition 4.1.** *Let $\mathbf{I}, \mathbf{X}, \mathbf{Y}, \mathbf{U} \in \mathbb{R}^d$ be the instrument, treatment, outcome, and unobserved confounder, respectively, related by the linear structural equations*

$$\mathbf{I} = \boldsymbol{\eta}_I, \tag{11}$$
$$\mathbf{U} = \boldsymbol{\eta}_U, \tag{12}$$
$$\mathbf{X} = \mathbf{I}\, M_{IX} + \mathbf{U}\, M_{UX} + \boldsymbol{\eta}_X, \tag{13}$$
$$\mathbf{Y} = \mathbf{X}\, M_{XY} + \mathbf{U}\, M_{UY} + \boldsymbol{\eta}_Y, \tag{14}$$

*where the mechanism matrix $M_{IX} \in \mathbb{R}^{d \times d}$ admits the SCBM factorization*

$$M_{IX} \;=\; W^b_{IX}\, W^f_{IX}, \qquad W^b_{IX} \in \mathbb{R}^{d \times r}, \quad W^f_{IX} \in \mathbb{R}^{r \times d},$$

*with $W^b_{IX}$ and $W^f_{IX}$ each of rank $r$, so that $\mathrm{rank}(M_{IX}) = r < d$. Assume $\mathbf{I} \perp (\mathbf{U},\, \eta_X,\, \eta_Y)$ and $\Sigma_I := \mathbb{E}[\mathbf{I}^\top \mathbf{I}] \succ 0$. Then the population 2SLS estimator for $M_{XY}$ has no unique solution.*

*Proof of Proposition 4.1.* **First stage.** The population OLS of $\mathbf{X}$ on $\mathbf{I}$ solves $\widehat{M}_{\mathrm{FS}} = \arg\min_M \mathbb{E}[\|\mathbf{X} - \mathbf{I}M\|^2]$. Substituting (3) and using the exogeneity conditions $\mathbf{I} \perp \mathbf{U}$, $\mathbf{I} \perp \boldsymbol{\eta}_X$ gives

$$\mathbb{E}[\mathbf{I}^\top \mathbf{X}] \;=\; \Sigma_I\, M_{IX},$$

so $\widehat{M}_{\mathrm{FS}} = \Sigma_I^{-1} \Sigma_I M_{IX} = M_{IX}$. The population first-stage fitted values are therefore $\hat{\mathbf{X}} = \mathbf{I}\, M_{IX}$.

**Second stage.** 2SLS regresses $\mathbf{Y}$ on $\hat{\mathbf{X}}$, leading to the population normal equations

$$G\,\Gamma \;=\; M_{IX}^\top\, \mathbb{E}[\mathbf{I}^\top \mathbf{Y}], \qquad G \;:=\; M_{IX}^\top\, \Sigma_I\, M_{IX}. \tag{15}$$

Using exogeneity ($\mathbf{I} \perp \mathbf{U}, \mathbf{I} \perp \boldsymbol{\eta}_Y$) and (4), we obtain $\mathbb{E}[\mathbf{I}^\top \mathbf{Y}] = \Sigma_I \, M_{IX} \, M_{XY}$, so the right-hand side of (15) equals $G \, M_{XY}$ and the normal equations reduce to

$$G \, \Gamma \;=\; G \, M_{XY}. \tag{16}$$

Since $\Sigma_I \succ 0$ and $\operatorname{rank}(M_{IX}) = r < d$, we have $\operatorname{rank}(G) = r < d$, so $G$ is singular with $\ker(G) = \ker(M_{IX})$, a subspace of dimension $d - r > 0$. Consequently (16) is satisfied by $\Gamma = M_{XY} + N$ for any $N \in \ker(M_{IX})$, yielding infinitely many solutions. The 2SLS estimator for $M_{XY}$ is therefore not uniquely defined. □

**Theorem 4.3.** *Under the setup of Proposition 4.1, additionally let $M_{XY} = W^b_{XY} W^f_{XY}$ with $W^b_{XY} \in \mathbb{R}^{d \times r}$, $W^f_{XY} \in \mathbb{R}^{r \times d}$, each of rank $r$. Suppose the rank condition*

$$\operatorname{rank}\Big(W^f_{IX} \, W^b_{XY}\Big) = r \tag{RC}$$

*holds. Given oracle access to $W^b_{IX}$ and $W^b_{XY}$, define the $r$-dimensional bottleneck representations*

$$\mathbf{Z}_{IX} \;:=\; \mathbf{I} \, W^b_{IX}, \qquad \mathbf{Z}_{XY} \;:=\; \mathbf{X} \, W^b_{XY}.$$

*The oracle bottleneck 2SLS procedure regresses $\mathbf{Z}_{XY}$ on $\mathbf{Z}_{IX}$ (first stage) to obtain fitted values $\hat{\mathbf{Z}}_{XY}$, then regresses $\mathbf{Y}$ on $\hat{\mathbf{Z}}_{XY}$ (second stage) to obtain $\hat{\Phi}$, and forms the estimator*

$$\widehat{M}_{XY} \;:=\; W^b_{XY} \, \hat{\Phi}.$$

*Then $\widehat{M}_{XY} = M_{XY}$.*

*Proof of Theorem 4.3.* Let $A := W^f_{IX} \, W^b_{XY} \in \mathbb{R}^{r \times r}$ and $\Sigma_{b_{IX}} := (W^b_{IX})^\top \Sigma_I \, W^b_{IX} \in \mathbb{R}^{r \times r}$.

**First stage.** The population Gram matrix for the OLS of $\mathbf{Z}_{XY}$ on $\mathbf{Z}_{IX}$ is $\Sigma_{Z_{IX}}$. Since $\Sigma_I \succ 0$ and $W^b_{IX}$ has rank $r$, we have $\Sigma_{Z_{IX}} \succ 0$, so the first stage has a unique solution. Using exogeneity ($\mathbf{I} \perp \mathbf{U}, \mathbf{I} \perp \boldsymbol{\eta}_X$) and (3):

$$\mathbb{E}[\mathbf{Z}_{IX}^\top \mathbf{Z}_{XY}] \;=\; (W^b_{IX})^\top \mathbb{E}[\mathbf{I}^\top \mathbf{X}] \, W^b_{XY} \;=\; (W^b_{IX})^\top \Sigma_I \, M_{IX} \, W^b_{XY} \;=\; \Sigma_{Z_{IX}} \, A,$$

so the population first-stage coefficient is $A^* = \Sigma_{Z_{IX}}^{-1} \Sigma_{Z_{IX}} \, A = A$, and the population fitted values are $\hat{\mathbf{Z}}_{XY} = \mathbf{Z}_{IX} \, A = \mathbf{I} \, M_{IX} \, W^b_{XY}$.

**Second stage.** The population Gram matrix for the OLS of $\mathbf{Y}$ on $\hat{\mathbf{Z}}_{XY}$ is

$$H \;:=\; \mathbb{E}[\hat{\mathbf{Z}}_{XY}^\top \hat{\mathbf{Z}}_{XY}] \;=\; (W^b_{XY})^\top M_{IX}^\top \Sigma_I \, M_{IX} \, W^b_{XY} \;=\; A^\top \Sigma_{Z_{IX}} \, A.$$

Condition (RC) gives $\operatorname{rank}(A) = r$, so $H \succ 0$ and the second stage has a unique solution. Using exogeneity ($\mathbf{I} \perp \mathbf{U}, \mathbf{I} \perp \boldsymbol{\eta}_Y$) and (4), the same calculation as in the proof of Proposition 4.1 gives $\mathbb{E}[\mathbf{I}^\top \mathbf{Y}] = \Sigma_I \, M_{IX} \, M_{XY}$, and hence

$$\mathbb{E}[\hat{\mathbf{Z}}_{XY}^\top \mathbf{Y}] \;=\; (W^b_{XY})^\top M_{IX}^\top \Sigma_I \, M_{IX} \, M_{XY} \;=\; (W^b_{XY})^\top M_{IX}^\top \Sigma_I \, M_{IX} \, W^b_{XY} \, W^f_{XY} \;=\; H \, W^f_{XY}.$$

Since $H \succ 0$, the normal equations yield uniquely $\hat{\Phi} = H^{-1} H \, W^f_{XY} = W^f_{XY}$.

**Composition.**

$$\widehat{M}_{XY} \;=\; W^b_{XY} \, \hat{\Phi} \;=\; W^b_{XY} \, W^f_{XY} \;=\; M_{XY}.$$ □

### A.2.2 Proofs for Section 4.2.1 (Mediator adjustment)

**Lemma 4.4.** *Let $\hat{M}^S_{XY}$ be the OLS estimator of $M_{XY}$ from the regression of $\mathbf{Y}$ on $(\mathbf{X}, S)$. Both $S = \mathbf{M}$ and $S = \mathbf{Z}_{MY}$ yield consistent estimators:*

$$\hat{M}^{M}_{XY} \xrightarrow{p} M_{XY}, \qquad \hat{M}^{Z_{MY}}_{XY} \xrightarrow{p} M_{XY}.$$

*Proof of Lemma 4.4.* Substituting $M_{MY} = W^b_{MY} W^f_{MY}$ into (7) gives

$$\mathbf{Y} = \mathbf{X} M_{XY} + \mathbf{Z}_{MY} W^f_{MY} + \boldsymbol{\eta}_Y.$$

For $S = \mathbf{Z}_{MY}$, this is a valid linear regression: the regressor matrix $(\mathbf{X}, \mathbf{Z}_{MY})$ is uncorrelated with $\boldsymbol{\eta}_Y$ because $\mathbf{Z}_{MY} = \mathbf{M} W^b_{MY}$ depends only on $(\mathbf{X}, \boldsymbol{\eta}_M)$, and all noise terms are mutually independent. OLS is therefore consistent for $M_{XY}$. The case $S = \mathbf{M}$ follows identically, since $\mathbf{M}$ also determines $\mathbf{Z}_{MY}$ and contains strictly more information. □

**Proposition 4.5.** *Under the linear SCBM, the asymptotic covariance of $\hat{M}^S_{XY}$ is $\Phi_S := Q_S^{-1}\Sigma_{\boldsymbol{\eta}_Y}Q_S^{-1}$, where $Q_S := \mathrm{Cov}(\tilde{\mathbf{X}}^S)$ denotes the covariance of the Frisch–Waugh–Lovell (Frisch and Waugh, 1933; Lovell, 1963) (FWL) residual of $\mathbf{X}$ after projecting out $S$. Then*

$$\Phi_{Z_{MY}} \preceq \Phi_M.$$

*Proof of Proposition 4.5.* By the Frisch–Waugh–Lovell theorem, $\hat{M}^S_{XY}$ equals OLS of $\tilde{\mathbf{Y}}^S$ on $\tilde{\mathbf{X}}^S$, with residual $\delta_S := \tilde{\mathbf{Y}}^S - \tilde{\mathbf{X}}^S M_{XY}$. Substituting the structural equation and using the independence $\boldsymbol{\eta}_Y \perp (\mathbf{X}, \mathbf{M})$, one obtains $\delta_S \xrightarrow{d} \boldsymbol{\eta}_Y$ for both $S \in \{\mathbf{M}, \mathbf{Z}_{MY}\}$, so $\Sigma_{\delta_S} = \Sigma_{\boldsymbol{\eta}_Y}$ in both cases.

Since $\mathbf{Z}_{MY} = \mathbf{M}W^b_{MY}$ is a function of $\mathbf{M}$, the $\sigma$-algebra generated by $\mathbf{Z}_{MY}$ is coarser than that of $\mathbf{M}$. By the law of total variance, conditioning on the finer variable removes more variance:

$$Q_M = \mathrm{Cov}\big(\mathbf{X} - \mathbb{E}[\mathbf{X} \mid \mathbf{M}]\big) \preceq \mathrm{Cov}\big(\mathbf{X} - \mathbb{E}[\mathbf{X} \mid \mathbf{Z}_{MY}]\big) = Q_{Z_{MY}}.$$

Hence $Q^{-1}_{Z_{MY}} \preceq Q_M^{-1}$, and since $\Phi_S = Q_S^{-1}\Sigma_{\boldsymbol{\eta}_Y}Q_S^{-1}$:

$$\Phi_{Z_{MY}} = Q^{-1}_{Z_{MY}}\Sigma_{\boldsymbol{\eta}_Y}Q^{-1}_{Z_{MY}} \preceq Q_M^{-1}\Sigma_{\boldsymbol{\eta}_Y}Q_M^{-1} = \Phi_M. \qquad \square$$

$\square$

### A.2.3 Proofs for Section 4.2.2 (Backdoor adjustment)

**Lemma 4.6.** *Let $\mathbf{Z}$ be an $r$-dimensional bottleneck representation of $\mathbf{C}$. The population OLS estimator of $\mathbf{Y}$ on $(\mathbf{X}, \mathbf{Z})$ consistently identifies $M_{XY}$ if $\mathbf{Z}$ satisfies either of the following conditions:*

- *(a)* ***Outcome sufficiency:*** *$\mathbf{C}\,M_{CY}$ is a measurable function of $\mathbf{Z}$, i.e. the confounding path $\mathbf{C} \to \mathbf{Y}$ factors through $\mathbf{Z}$;*
- *(b)* ***Treatment sufficiency:*** *$\mathbf{X} \perp \mathbf{C} \mid \mathbf{Z}$, i.e. $\mathbf{Z}$ screens off $\mathbf{C}$ from $\mathbf{X}$.*

*In the linear SCBM, $\mathbf{Z}_{CY}$ satisfies* (a) *and $\mathbf{Z}_{CX}$ satisfies* (b).

*Proof of Lemma 4.6.* Identification requires $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}] = \mathbf{X}\,M_{XY} + g(\mathbf{Z})$ for some function $g$. Substituting (10):

$$\mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}] \;=\; \mathbf{X}\,M_{XY} \;+\; \mathbb{E}[\mathbf{C}\,M_{CY} \mid \mathbf{X}, \mathbf{Z}].$$

It suffices to show the second term depends only on $\mathbf{Z}$.

**Case (a), $\mathbf{Z} = \mathbf{Z}_{CY}$.** Since $M_{CY} = W^b_{CY}W^f_{CY}$, we have $\mathbf{C}\,M_{CY} = \mathbf{Z}_{CY}\,W^f_{CY}$, so $\mathbb{E}[\mathbf{C}\,M_{CY} \mid \mathbf{X}, \mathbf{Z}_{CY}] = \mathbf{Z}_{CY}\,W^f_{CY}$ is a function of $\mathbf{Z}_{CY}$ alone. Hence $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}_{CY}] = \mathbf{X}\,M_{XY} + \mathbf{Z}_{CY}\,W^f_{CY}$.

**Case (b), $\mathbf{Z} = \mathbf{Z}_{CX}$.** From (9), $\mathbf{X} = \mathbf{Z}_{CX}\,W^f_{CX} + \boldsymbol{\eta}_X$ with $\boldsymbol{\eta}_X \perp \mathbf{C}$, so the variation in $\mathbf{X}$ given $\mathbf{Z}_{CX}$ is $\boldsymbol{\eta}_X$, which is independent of $\mathbf{C}$. Therefore $\mathbf{X} \perp \mathbf{C} \mid \mathbf{Z}_{CX}$, giving $\mathbb{E}[\mathbf{C} \mid \mathbf{X}, \mathbf{Z}_{CX}] = \mathbb{E}[\mathbf{C} \mid \mathbf{Z}_{CX}] =: g(\mathbf{Z}_{CX})$. Hence $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}_{CX}] = \mathbf{X}\,M_{XY} + g(\mathbf{Z}_{CX})\,M_{CY}$.

In both cases the coefficient on $\mathbf{X}$ in the population OLS is $M_{XY}$. $\square$

**Proposition 4.7.** *Let $\hat{M}^C_{XY}$, $\hat{M}^{CX}_{XY}$, and $\hat{M}^{CY}_{XY}$ denote the OLS estimators of $M_{XY}$ obtained by conditioning on $\mathbf{C}$, $\mathbf{Z}_{CX}$, and $\mathbf{Z}_{CY}$, respectively. All three are consistent. Defining $\tilde{\mathbf{V}}^S := \mathbf{V} - \mathbb{E}[\mathbf{V} \mid S]$ as the population residual and $\delta_S := \tilde{\mathbf{Y}}^S - \tilde{\mathbf{X}}^S M_{XY}$ as the effective regression noise, the asymptotic variance of each estimator is characterized by*

$$\Phi_S \;:=\; Q_S^{-1}\,\Sigma_{\delta_S}\,Q_S^{-1}, \qquad Q_S \;:=\; \mathrm{Cov}(\tilde{\mathbf{X}}^S), \quad \Sigma_{\delta_S} \;:=\; \mathrm{Cov}(\delta_S),$$

*and satisfies $\Phi_{CY} \prec \Phi_C \prec \Phi_{CX}$, i.e.*

$$\mathrm{Avar}\Big(\hat{M}^{CY}_{XY}\Big) \;\prec\; \mathrm{Avar}\Big(\hat{M}^{C}_{XY}\Big) \;\prec\; \mathrm{Avar}\Big(\hat{M}^{CX}_{XY}\Big),$$

*with both strict inequalities holding whenever $\ker(W^{b,\top}_{CY}) \cap \mathrm{col}(M^\top_{CX}) \neq \{0\}$ and $\ker(W^{b,\top}_{CX}) \cap \mathrm{col}(W^b_{CY}) \neq \{0\}$, respectively.*

*Proof of Proposition 4.7.* We compute $Q_S$ and $\Sigma_{\delta_S}$ for each conditioning variable.

**Conditioning on C.** From (9)–(10) and $\boldsymbol{\eta}_X \perp \mathbf{C}$:

$$\tilde{\mathbf{X}}^C = \boldsymbol{\eta}_X, \qquad \tilde{\mathbf{Y}}^C = \tilde{\mathbf{X}}^C M_{XY} + \boldsymbol{\eta}_Y, \qquad \delta_C = \boldsymbol{\eta}_Y.$$

Hence $Q_C = \Sigma_{\boldsymbol{\eta}_X}$ and $\Sigma_{\delta_C} = \Sigma_{\boldsymbol{\eta}_Y}$.

**Conditioning on $\mathbf{Z}_{CY}$.** Since $\mathbf{C}\, M_{CY} = \mathbf{Z}_{CY}\, W^f_{CY}$ is a function of $\mathbf{Z}_{CY}$:

$$\tilde{\mathbf{Y}}^{CY} \;=\; \tilde{\mathbf{X}}^{CY} M_{XY} + \boldsymbol{\eta}_Y, \qquad \delta_{CY} = \boldsymbol{\eta}_Y, \qquad \Sigma_{\delta_{CY}} = \Sigma_{\boldsymbol{\eta}_Y}.$$

The treatment residual retains the component of $\mathbf{C}$ outside $\mathrm{col}(W^b_{CY})$:

$$Q_{CY} \;=\; \mathrm{Cov}(\tilde{\mathbf{X}}^{CY}) \;=\; M_{CX}^\top \,\mathrm{Cov}(\mathbf{C} \mid \mathbf{Z}_{CY})\, M_{CX} + \Sigma_{\boldsymbol{\eta}_X} \;\succeq\; \Sigma_{\boldsymbol{\eta}_X} \;=\; Q_C.$$

Since $\Sigma_{\delta_{CY}} = \Sigma_{\delta_C} = \Sigma_{\boldsymbol{\eta}_Y}$ but $Q_{CY} \succeq Q_C$, we have $Q_{CY}^{-1} \preceq Q_C^{-1}$ and therefore $\Phi_{CY} \preceq \Phi_C$.

**Conditioning on $\mathbf{Z}_{CX}$.** Since $\mathbf{X} = \mathbf{Z}_{CX}\, W^f_{CX} + \boldsymbol{\eta}_X$ with $\boldsymbol{\eta}_X \perp \mathbf{C}$:

$$\tilde{\mathbf{X}}^{CX} = \boldsymbol{\eta}_X, \qquad Q_{CX} = \Sigma_{\boldsymbol{\eta}_X} = Q_C.$$

However, $\mathbf{Z}_{CX}$ does not fully explain $\mathbf{C}\, M_{CY}$, so the effective noise retains residual confounding:

$$\delta_{CX} \;=\; \big(\mathbf{C} - \mathbb{E}[\mathbf{C} \mid \mathbf{Z}_{CX}]\big)\, M_{CY} + \boldsymbol{\eta}_Y,$$

giving

$$\Sigma_{\delta_{CX}} \;=\; M_{CY}^\top \,\mathrm{Cov}(\mathbf{C} \mid \mathbf{Z}_{CX})\, M_{CY} + \Sigma_{\boldsymbol{\eta}_Y} \;\succeq\; \Sigma_{\boldsymbol{\eta}_Y} \;=\; \Sigma_{\delta_C}.$$

Since $Q_{CX} = Q_C$ but $\Sigma_{\delta_{CX}} \succeq \Sigma_{\delta_C}$, we have $\Phi_{CX} \succeq \Phi_C$.

**Strict inequalities.** $\Phi_{CY} \prec \Phi_C$ iff $Q_{CY} \succ Q_C$, which holds iff $\mathrm{Cov}(\mathbf{C} \mid \mathbf{Z}_{CY})$ is nonzero in the direction of $M_{CX}$, i.e. $\ker(W^{b,\top}_{CY}) \cap \mathrm{col}(M^\top_{CX}) \neq \{0\}$. $\Phi_{CX} \succ \Phi_C$ iff $\Sigma_{\delta_{CX}} \succ \Sigma_{\boldsymbol{\eta}_Y}$, which holds iff $\mathrm{Cov}(\mathbf{C} \mid \mathbf{Z}_{CX})$ is nonzero in the direction of $M_{CY}$, i.e. $\ker(W^{b,\top}_{CX}) \cap \mathrm{col}(W^b_{CY}) \neq \{0\}$. Both conditions hold with probability one whenever the weight matrices are drawn from any distribution absolutely continuous with respect to Lebesgue measure. □

## B Special cases and interpretations

The class of SCBMs is large and allows to model interactions among spatial as well as temporal phenomena as we illustrate below. We also show that SCBMs can be interpreted through the lens of Tishby's information bottlenecks (Tishby et al., 2000).

**Modeling interactions between random fields.** The quantity of interest in geo-spatial modelling are often represented by random fields over some spatial area $\mathcal{D}$. This spatial area can be a discrete grid or a continuous domain. The random field is a random function over $\mathcal{D}$, which in the discrete case can be identified with a vector of dimension $N$, where $N$ is the number of grid points. The spatial structure in such a model is reflected in the correlation structure of the field. For instance, if $\mathcal{D}$ is a discrete grid, a typical assumption is the spatial Markov property, which states that every node $X_i,\ i \in \mathcal{D}$, is independent of non-adjacent nodes given its neighbors. In continuous random fields, a standard assumption would be that the correlation between two nodes $X(y), X(z),\ y, z \in \mathcal{D}$, decreases when the spatial distance of $y$ and $z$ increases. An SCBM could for example consist of three random fields of noises $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3$ over bounded spatial domains $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ where fields affect each other only through weighted spatial means, e.g.

$$\begin{aligned}
\mathbf{X}_1 &:= \boldsymbol{\eta}_1 \\
\mathbf{X}_2 &:= A \cdot \int_{\mathcal{D}_1} \alpha(y)\mathbf{X}_1(y)\, dy + \boldsymbol{\eta}_2 \\
\mathbf{X}_3 &:= B \cdot \int_{\mathcal{D}_1} \beta(y)\mathbf{X}_1(y)\, dy + C \cdot \int_{\mathcal{D}_2} \gamma(z)\mathbf{X}_2(z)\, dz + \boldsymbol{\eta}_3,
\end{aligned}$$

where $\alpha, \beta, \gamma$ are normalized to one, e.g. $\int_{\mathcal{D}_1} \alpha(y)\, dy = 1$. In this case, the spaces $\mathcal{X}_i$ are function spaces over the respective domains, e.g. $\mathcal{X}_i = L^2(\mathcal{D}_i)$, the bottleneck spaces are one dimensional $\mathcal{Z}_i \cong \mathbb{R}$ and the bottleneck functions are given by the weighted integrals. The effect functions are simple embeddings by multiplication with a constant, e.g. $f_{12} : \mathcal{Z}_{12} \to L^2(\mathcal{D}_2) : z \mapsto Az \cdot \mathbb{1}_{\mathcal{D}_2}$, where $\mathbb{1}_{\mathcal{D}_2}$ denotes the constant function on $\mathcal{D}_2$.

**Temporal processes in the frequency domain.** SCBMs can also model interactions of temporal processes in the frequency domain. For instance, $\mathbf{X}_1$ and $\mathbf{X}_2$ may be continuous stochastic processes viewed as random variables with values in $L^2(\mathbb{R}_{\geq 0})$. Their child process $\mathbf{Y}$ may be affected by $\mathbf{X}_1, \mathbf{X}_2$ within at most $K$ frequency components with frequencies $\omega_1, \dots \omega_K$, see (Schur et al., 2026) for an example of this. Thus, the bottleneck maps $b_i : L^2(\mathbb{R}_{\geq 0}) \to \mathbb{R}^K$ map the processes to the $K$ Fourier coefficients at $\omega_1, \dots \omega_K$, and the effect function expresses $\mathbf{Y}$ as $\mathbf{Y}_t = \sum_{k=1,\dots K}(\lambda_1 b_1(\mathbf{x}_1) + \lambda_2 b_2(\mathbf{x}_2))e^{-it\omega_k} + \boldsymbol{\eta}$ where $\boldsymbol{\eta}$ is a noise process, for instance a Brownian motion.

**SCBMs as information bottlenecks.** Consider two high-dimensional multivariate random variables $\mathbf{X}, \mathbf{Y}$. The idea of the information bottleneck framework of Tishby et al. (2000) is to find a minimal sufficient statistic for $\mathbf{T} = g(\mathbf{X})$ that captures as much information about the target variable $\mathbf{Y}$ as possible. This goal is captured by the *minimal compression* optimization objective

$$\min_{\mathbf{T}\,:\, I(\mathbf{X},\mathbf{Y}|\mathbf{T})=0} I(\mathbf{X}, \mathbf{T}).$$

The *independence constraint* $I(\mathbf{X}, \mathbf{Y}|\mathbf{T}) = 0$ is equivalent to $I(\mathbf{X}, \mathbf{Y}) = I(\mathbf{T}, \mathbf{Y})$. Note also that the data processing inequality enforces

$$I(g(\mathbf{X}), \mathbf{Y}) \leq I(\mathbf{X}, \mathbf{Y})$$

for any abstraction function $g$, an information bottleneck is thus nothing but an abstraction of $\mathbf{X}$ that maximizes $I(g(\mathbf{X}), \mathbf{Y})$. This property is used to incorporate the independence constraint as a soft constraint in the joint objective

$$\min_{\mathbf{T}} \left(I(\mathbf{X}, \mathbf{T}) - \beta I(\mathbf{Y}, \mathbf{T})\right)$$

with regularization parameter $\beta > 0$. Consider now an SCBM with a shared bottleneck for all children of a node, i.e., $\mathbf{Z}_{ij} = \mathbf{Z}_i$ for all $j \in \mathrm{ch}(i)$, over variables $\mathbf{X}_i$ with bottleneck variables $\mathbf{Z}_i = b_i(\mathbf{X}_i)$ for every non-sink node $i$, i.e.,

$$\mathbf{X}_j = f_j(\mathbf{Z}_{i_1}, \dots, \mathbf{Z}_{i_m}) + \boldsymbol{\eta}_j \qquad \mathrm{pa}(j) = \{i_1, \dots, i_m\}.$$

Our goal is to produce an optimization objective and constraints to learn the bottleneck variables $\mathbf{Z}_i$. The minimal compression requirement now means that $\mathbf{Z}_i$ contains the minimal necessary information about $\mathbf{X}$ once $\mathbf{X}$'s parents are known, i.e. $\mathbf{Z}_i \in \operatorname{argmin} I(\mathbf{X}_i, \mathbf{Z}_i|\mathbf{Z}_{\mathrm{pa}(i)})$ for all $i$.

The bottleneck $\mathbf{Z}_i$ is intended to be maximally informative about the children of $\mathbf{X}_i$ provided that all backdoor paths to these children are closed which can be achieved by conditioning on the parental bottlenecks $\mathbf{Z}_{\mathrm{pa}(i)} = \{\mathbf{Z}_k, k \in \mathrm{pa}(i)\}$. Thus, the conditional independence constraint is

$$I(\mathbf{X}_{\mathrm{ch}(i)}, \mathbf{X}_i|\mathbf{Z}_i, \mathbf{Z}_{\mathrm{pa}(i)}) = 0.$$

for non-sink nodes $i$. By the chain rule for conditional mutual information, this is equivalent to

$$I(\mathbf{X}_{\mathrm{ch}(i)}, \mathbf{Z}_i|\mathbf{Z}_{\mathrm{pa}(i)}) = I(\mathbf{X}_{\mathrm{ch}(i)}, \mathbf{X}_i|\mathbf{Z}_{\mathrm{pa}(i)}).$$

Thus, writing $\mathbf{Z} = (\mathbf{Z}_i)_{i \in \mathcal{V}_{ns}}$, we want to solve the family of optimization objectives

$$\min_{\mathbf{Z}\,:\, I(\mathbf{X}_{\mathrm{ch}(i)}, \mathbf{X}_i|\mathbf{Z}_i, \mathbf{Z}_{\mathrm{pa}(i)})=0} I(\mathbf{X}_i, \mathbf{Z}_i|\mathbf{Z}_{\mathrm{pa}(i)}), \quad i \in \mathcal{V}_{ns}.$$

These objectives are linked by the fact that the i-th objective involves not only the bottleneck $\mathbf{Z}_i$ but also the bottlenecks of the parent variables which may be part of other objectives as well. Instead of solving for $\mathbf{Z}$ globally, it is possible to consider these objectives sequentially along a causal order. To formalize this, define the *causal grading* $\mathcal{V} = \bigsqcup_s \mathcal{V}_s$ where $\mathcal{V}_0 = \mathcal{V}_{ex}$ and $i \in \mathcal{V}_s$ for $s > 0$ if and only if $i \notin \mathcal{V}_q, q < s$ and $\mathrm{pa}(i) \subset \bigsqcup_{q<s} \mathcal{V}_q$. Fixing a causal order $(i_1, \dots, i_n)$ that respects the causal grading allows us to solve the objectives

$$\min_{\mathbf{Z}_i} \left(I(\mathbf{X}_i, \mathbf{Z}_i|\mathbf{Z}_{\mathrm{pa}(i)}) - \beta_i I(\mathbf{X}_{\mathrm{ch}(i)}, \mathbf{Z}_i|\mathbf{Z}_{\mathrm{pa}(i)})\right)$$

sequentially along this order since all parental bottlenecks appearing in the $i$-th equation have been computed as solutions of earlier objectives, see also Appendix E.2.

## C Detailed related work

**Causal representation learning.** Although we also learn representations of high-dimensional observations, our approach differs from canonical causal representation learning (CRL) approaches (Schölkopf et al., 2021; Brehmer et al., 2022; Ahuja et al., 2023; Lippe et al., 2022, 2023a,b; Squires et al., 2023; Buchholz et al., 2023; Liang et al., 2023; Varıcı et al., 2025; Bing et al., 2024; Lachapelle et al., 2026; von Kügelgen et al., 2024; Yao et al., 2024) in key ways. Instead of mapping observations to SCM nodes, we learn *multiple* maps tied to mechanisms between known (vector-valued) SCM nodes. Unlike most CRL methods that seek latent variables up to permutation and rescaling, we allow a broader class of invertible transformations for identifiability. Additionally, while CRL assumes invertible maps, we focus on surjective maps to enable true dimension reduction by discarding irrelevant observational details.

CRL aims to recover a latent low-dimensional SCM including its constituting variables. In contrast, learning bottlenecks in an SCBM is focused on causal effect estimation. We learn representations assuming a known graph and target a specific downstream causal query. Defining causal representations by their usefulness for downstream tasks, rather than recovering a single postulated ground-truth latent model, has been proposed as a step forward for CRL (Jørgensen et al., 2025), and our targeted approach aligns with this view.

**Causal abstractions.** Our work is related to causal abstraction learning, but *what* we abstract is different from existing approaches. Common approaches abstract causal models as a whole (Zennaro et al., 2023; Felekis et al., 2024; Xia and Bareinboim, 2024; Massidda et al., 2024; D'Acunto et al., 2025; Madaleno et al., 2026), while our notion of abstraction is a within-model operation that reduces a random vector to the essentials needed for downstream effect estimations.

When a bottleneck is shared across all children of a node and discrete, the inverse image of the bottleneck function recovers exactly the intra-variable partitions of Xia and Bareinboim (2024). A factored discrete bottleneck can then be understood as a target-specific variant of their setting. More precisely, the level sets of each bottleneck function $b_{ij}$ define which states of $\mathbf{X}_i$ are causally equivalent for child $\mathbf{X}_j$—a per-child instance of the structural invariances of Kroeger et al. (2025) (Definition 7). In the factored SCBM setting, distinct children carry distinct invariances, which generalizes beyond what a single structural invariance—required by Definition 7 to hold for all children simultaneously—can express. Two distinctions persist across all cases: SCBMs require no causal model over the micro-level variables—only a joint distribution, e.g. a Markov random field—and bottleneck spaces need not be discrete.

An approach closely related to ours is the causal feature learning method proposed by Chalupka et al. (2017), which learns to partition the spaces of a cause-effect pair to extract high-level features from low-level observations. The SCBM framework extends this idea to settings with more than two variables and allows for continuous high-level variables (bottlenecks), as opposed to only permitting discrete variables as a result of a clustering operation.

The Causal Information Bottleneck (CIB) (Simoes et al., 2025) seeks a representation for a specific causal query, but differs from our approach in its problem setting and method. The CIB framework focuses on a single cause-effect pair which must be identifiable via the backdoor criterion, while we handle arbitrary acyclic graphs. Simoes et al. (2025) extend the Information Bottleneck Lagrangian of Tishby et al. (2000) to trade off compression and what they call "causal control" via a constrained optimization, whereas we assume a data generating process that conceptualizes optimal low-dimensional representations directly and enables bottleneck estimation through regression or likelihood-based losses. The CIB method requires access to all conditional probability distributions, as it must compute $p(\mathbf{Y} \mid \mathrm{do}(\mathbf{X} = \mathbf{x}))$, while we use only observational data and graph structure. Additionally, CIB assumes discrete variables; we allow both discrete and continuous.

**Dimension reduction.** Principal component analysis (PCA) (Pearson, 1901) and similar dimension reduction techniques can be understood to represent the opposing side of a trade-off between training effort required before application and guarantees over retained relevant information for a downstream task. PCA does not need to be fit to a specific setting, but may also discard data relevant to a downstream task. In our framework, we do require additional data to fit an estimator of a given bottleneck, but we gain guarantees of the compressed representation being optimal for describing the specific mechanism of interest.

In a setting with only one regressor $\mathbf{X}$ and one regressand $\mathbf{Y}$, the task of finding an appropriate bottleneck variable is known as sufficient dimension(ality) reduction (Globerson and Tishby, 2003; Li, 2007, 2018). If in addition the relationship of $\mathbf{X}$ and $\mathbf{Y}$ is assumed linear, *reduced-rank regression* (Izenman, 1975) implicitly assumes the existence of a low-dimensional bottleneck variable by restricting the rank of the effect matrix.

**High-dimensional instrumental variables.** Bai and Ng (2010) assume $N$ noisy scalar instruments each dependent on $r$ latent factors, and use PCA to extract those factors as the effective instruments. Their bottleneck is on the instrument side (many noisy proxies $\to r$ factors), and their core problem is a many-instruments finite-sample bias of order $N/T$ in 2SLS; population-level non-identifiability does not arise because the number of instruments grows. Lin et al. (2026) consider the complementary regime: a low-dimensional instrument $\mathbf{Z} \in \mathbb{R}^k$ and a high-dimensional treatment $\mathbf{X} \in \mathbb{R}^m$ with $m \gg k$. They show that naive unsupervised compression of $\mathbf{X}$ creates an unblocked backdoor path $\mathbf{Z} \to \mathbf{X} \setminus \tilde{D} \to Y$, biasing the IV estimate, and propose recovering the instrument-relevant treatment directions via SVD of the reduced-form regression of $\mathbf{X}$ on $\mathbf{Z}$.

Our IV result in Section 4 is related to both but differs in three key respects. First, we identify a *population-level* non-identifiability: when $\dim(\mathbf{I}) = \dim(\mathbf{X}) = \dim(\mathbf{Y}) = d$ and $\operatorname{rank}(M_{IX}) = r < d$, the 2SLS problem has infinitely many solutions—a failure that arises regardless of sample size. Second, the oracle bottleneck 2SLS procedure uses the *outcome-side* bottleneck $W^b_{XY}$ in the second stage, in addition to the instrument-side $W^b_{IX}$; these two subspaces are generically distinct, and conflating them (as a naïve analogue of Lin et al. (2026) would do) leaves a residual bias characterized by Remark 4.2. Third, all mechanism matrices in the SCBM—including $M_{XY}$—admit bottleneck factorizations, which is what makes our rank condition (RC) the checkable criterion for identification.

**Connections to matching.** Matching is a prominent technique for estimating the effect of an (often binary) treatment variable $T$ on an outcome $Y$ in the presence of confounding covariates $X_1, \dots, X_k$. Roughly speaking, the idea behind matching is to pair treated and untreated units that are as similar as possible in terms of their covariate values. *Exact matching*, that is, matching directly in the covariate space $\mathcal{X}_1 \times \cdots \times \mathcal{X}_k$, for instance by measuring the Euclidean distance of the covariate values of the units, is often unfeasible if the sample size is small compared to the number of covariates. In this case, matching is commonly done by measuring the similarity of units in a bottleneck space. For instance, *propensity score matching* declares units similar that are close in propensity score $e(\mathbf{x}) = P(T = 1|\mathbf{X} = \mathbf{x})$ (Rosenbaum and Rubin, 1983) which in our terminology is a $[0, 1]$-valued bottleneck between the covariates and the treatment. *Coarsened exact matching* (Blackwell et al., 2009) interpolates between exact matching and the propensity score by operating in a coarser bottleneck space.

# D Additional experimental results

## D.1 Identifiability

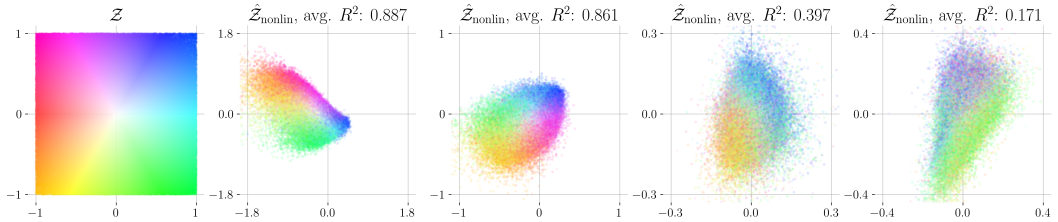


Figure 6: Visualizations of maps from ground truth bottleneck space $\mathcal{Z}$ (first plot) to different learned bottleneck spaces $\hat{\mathcal{Z}}_{\text{nonlin}}$. The second and third plots show examples of successfully identified bottlenecks, as visualized by the smooth, structure preserving bijections from ground-truth to learned bottleneck spaces. We additionally show the identifiability metric, which in both cases is close to one. The third and fourth plots show examples where identifiability is not achieved (by not training to convergence). For such non-identified cases, the learned map is not a bijection and the local structure of the ground-truth bottleneck space is not preserved. The low $R^2$ metric in these failure cases further reflects non-identification of the ground-truth bottleneck.

## D.2 Effect estimation

We repeat the mediator and backdoor adjustment experiments from Section 4.2 with varying values of $d_{\mathbf{X}}$ and $d_{\mathbf{Z}}$. As a slight abuse of notation, we refer to the dimension of *all* high-dimensional variables with $d_{\mathbf{X}}$, not just of the treatment $\mathbf{X}$. Results for the mediator case are plotted in Figure 7, those for the backdoor case in Figure 8. Across settings, conditioning on the bottleneck variable provides a finite sample benefit over conditioning on the high-dimensional mediator or confounder. This advantage is most pronounced when the bottleneck dimension is smaller and diminishes as the ratio $d_{\mathbf{X}}/d_{\mathbf{Z}}$ approaches one.

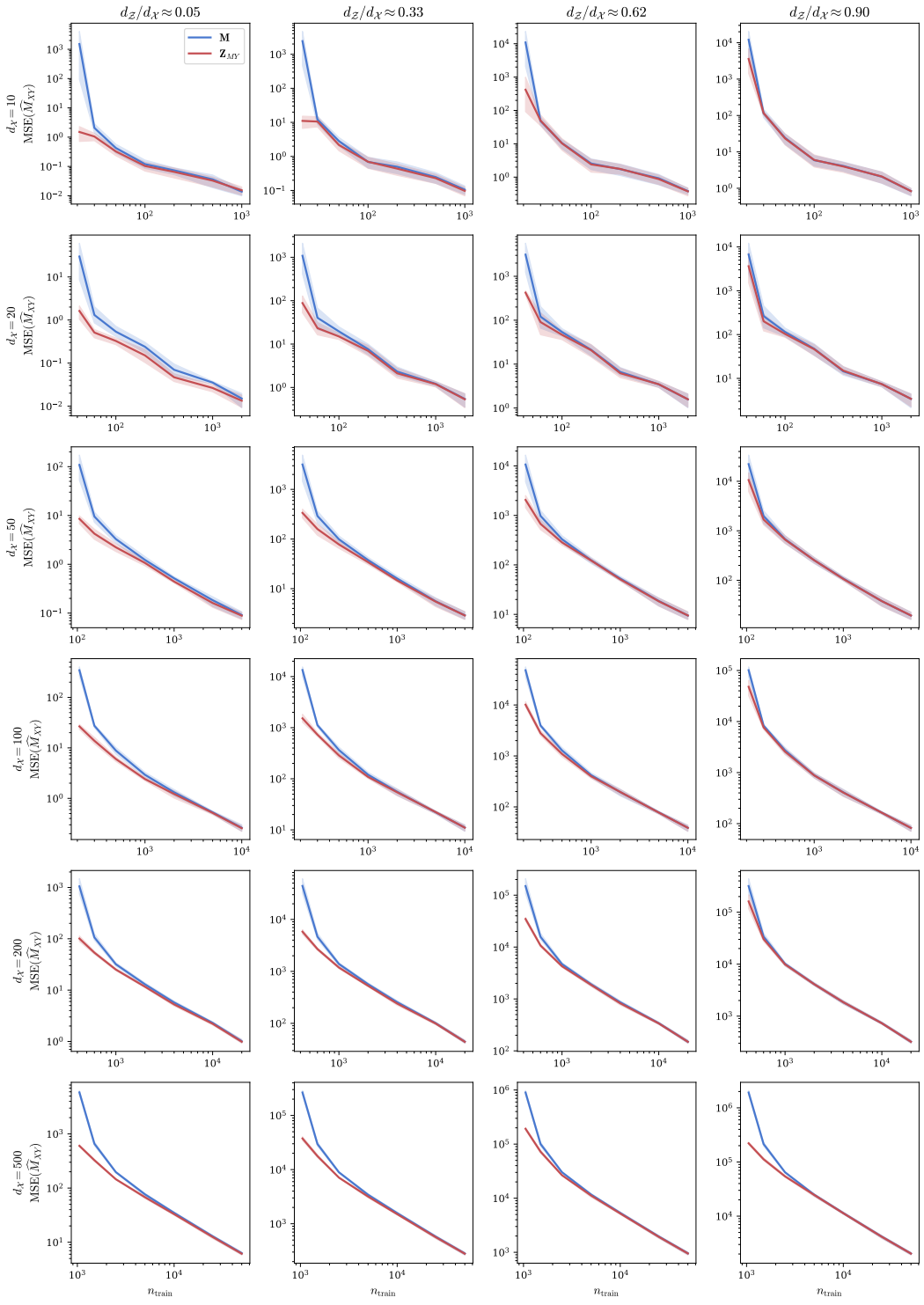


Figure 7: Results of the mediator adjustment experiment across various parameter configurations. Bottlenecks are preferable as conditioning variables over high-dimensional variables across settings. As $d_{\mathbf{X}}/d_{\mathbf{Z}} \to 1$, the finite sample benefits of bottlenecks diminish, with the notable exception of $d_{\mathbf{X}} = 500$.

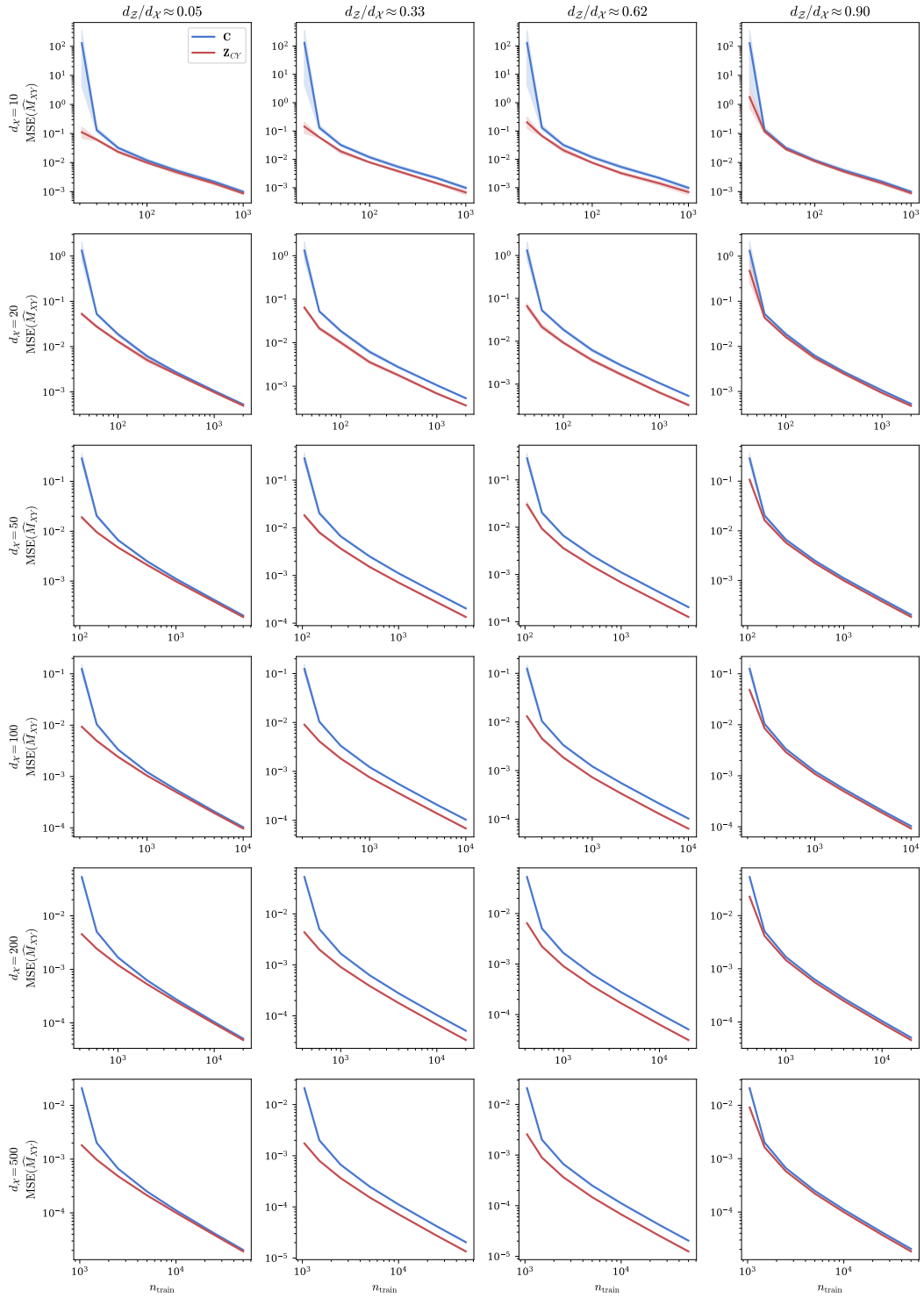


Figure 8: Results of the backdoor adjustment experiment across various parameter configurations. Bottlenecks are preferable as conditioning variables over high-dimensional variables across settings. As $d_{\mathbf{X}}/d_{\mathbf{Z}} \to 1$, the finite sample benefits of bottlenecks diminish. Interestingly, for intermediate values of $d_{\mathbf{X}}/d_{\mathbf{Z}}$ (middle two columns) the gap in estimation error of $M_{XY}$ when using the bottleneck vs. the high-dimensional confounder persists, even for large $n_{\text{train}}$.

### D.3 Transfer environments for IV estimation

Theorem 4.3 assumes oracle access to $W^b_{IX}$ and $W^b_{XY}$. $W^b_{IX}$ can be estimated from the observational data, since the instrument is unconfounded. $W^b_{XY}$, by contrast, belongs to the confounded mechanism $\mathbf{X} \to \mathbf{Y}$. We therefore estimate it from a transfer environment that shares this mechanism; here a pilot RCT in which $\mathbf{X}$ is intervened on.

**Setup.** We sample a linear SCBM as in Proposition 4.1 with $d = 50$ and $r = 2$, and draw $n_{\text{obs}} = 30000$ observational samples. In contrast to the other experiments, we sample all weights from $\text{Uniform}(-1.5, 1.5)$: strictly positive weights make the rank-$r$ mechanisms nearly rank one (median $\sigma_2/\sigma_1 = 0.04$ for $M_{XY}$, vs. 0.84 with centered weights), which leaves the weak bottleneck direction poorly determined by the pilot and makes the estimated 2SLS unstable in $n_{\text{pilot}}$. For the pilot, we draw $n_{\text{pilot}}$ samples from the same model under the intervention $\text{do}(\mathbf{X} \sim \mathcal{N}(0, I_d))$, which removes the confounding through $\mathbf{U}$. We estimate $\widehat{W}^b_{IX}$ by rank-$r$ factorization of the first-stage regression of $\mathbf{X}$ on $\mathbf{I}$ in the observational data, and $\widehat{W}^b_{XY}$ by rank-$r$ factorization of the regression of $\mathbf{Y}$ on $\mathbf{X}$ in the pilot data. We compare four estimators of $M_{XY}$: (i) *oracle bottleneck 2SLS*, i.e. the procedure of Theorem 4.3 with the true $W^b_{IX}, W^b_{XY}$; (ii) *estimated bottleneck 2SLS*, the same procedure with $\widehat{W}^b_{IX}, \widehat{W}^b_{XY}$; (iii) *ridge 2SLS* using the full $\mathbf{I}$ as instrument (penalty $\lambda = 10^{-3}$); (iv) *pilot-only OLS*, regressing $\mathbf{Y}$ on $\mathbf{X}$ in the pilot data alone. We report the median MSE of $\widehat{M}_{XY}$ over 10 seeds in Table 1.

Table 1: Median MSE of $\widehat{M}_{XY}$ over 10 seeds for varying pilot sizes ($n_{\text{obs}} = 30000$, $d = 50$, $r = 2$).

| $n_{\text{pilot}}$ | Oracle b-2SLS | Estimated b-2SLS | Ridge 2SLS | Pilot-only OLS |
|---|---|---|---|---|
| 100 | $1.2 \times 10^{-5}$ | 8.0 | $1 \times 10^{3}$ | 1.1 |
| 500 | $1.2 \times 10^{-5}$ | $7.9 \times 10^{-2}$ | $1 \times 10^{3}$ | $1.3 \times 10^{-1}$ |
| 2000 | $1.2 \times 10^{-5}$ | $4.9 \times 10^{-3}$ | $1 \times 10^{3}$ | $2.6 \times 10^{-2}$ |

**Results.** Ridge 2SLS stays at the biased projection target of Remark 4.2 regardless of sample size, confirming that the failure in Proposition 4.1 is a population-level one. Estimated bottleneck 2SLS outperforms pilot-only OLS from $n_{\text{pilot}} = 500$ onward and approaches the oracle as the pilot grows. The reason is that the pilot only needs to determine the $r$-dimensional subspace $\text{col}(W^b_{XY})$, since all results are invariant under $W^b_{XY} \mapsto W^b_{XY} G$ (Section 4). This amounts to $r(d - r)$ parameters, whereas pilot-only OLS must estimate all $d^2$ entries of $M_{XY}$. For very small pilots ($n_{\text{pilot}} = 100$), however, the subspace estimate is too noisy, and pilot-only OLS performs better. Transfer environments are therefore most useful when a moderately sized pilot is available but too small to estimate the full effect matrix reliably.

### D.4 Approximately low-rank mechanisms

Our efficiency results for bottleneck adjustment assume that mechanisms factor exactly through a rank-$r$ bottleneck. We study robustness to violations of this assumption in the backdoor setting of Section 4.2.2 by replacing the confounder–outcome mechanism with

$$M_{CY} = W^b_{CY} W^f_{CY} + E, \qquad \rho := \frac{\|E\|_F^2}{\|W^b_{CY} W^f_{CY}\|_F^2},$$

where $E$ is a full-rank perturbation. Since $\text{Cov}(\mathbf{C}) = I$, $\rho$ is the variance contributed by $\mathbf{C}E$ relative to that of $\mathbf{C}\, W^b_{CY} W^f_{CY}$, and $\rho = 0$ recovers the original SCBM. We condition on the rank-$r$ bottleneck $\mathbf{Z}_{CY} = \mathbf{C}\, W^b_{CY}$ of the low-rank part, which violates outcome sufficiency (Lemma 4.6(a)) for any $\rho > 0$, and compare against conditioning on the full confounder $\mathbf{C}$, which remains valid for all $\rho$. Details are given in Appendix E.3.5.

**Results.** The efficiency advantage of the bottleneck at $\rho = 0$ is modest (ratio of full to bottleneck MSE between 1.02 and 1.36) and is lost at very small perturbation levels (Figure 9). Across all configurations, the crossover $\rho^*$, i.e. the smallest $\rho$ at which the bottleneck MSE exceeds that of full

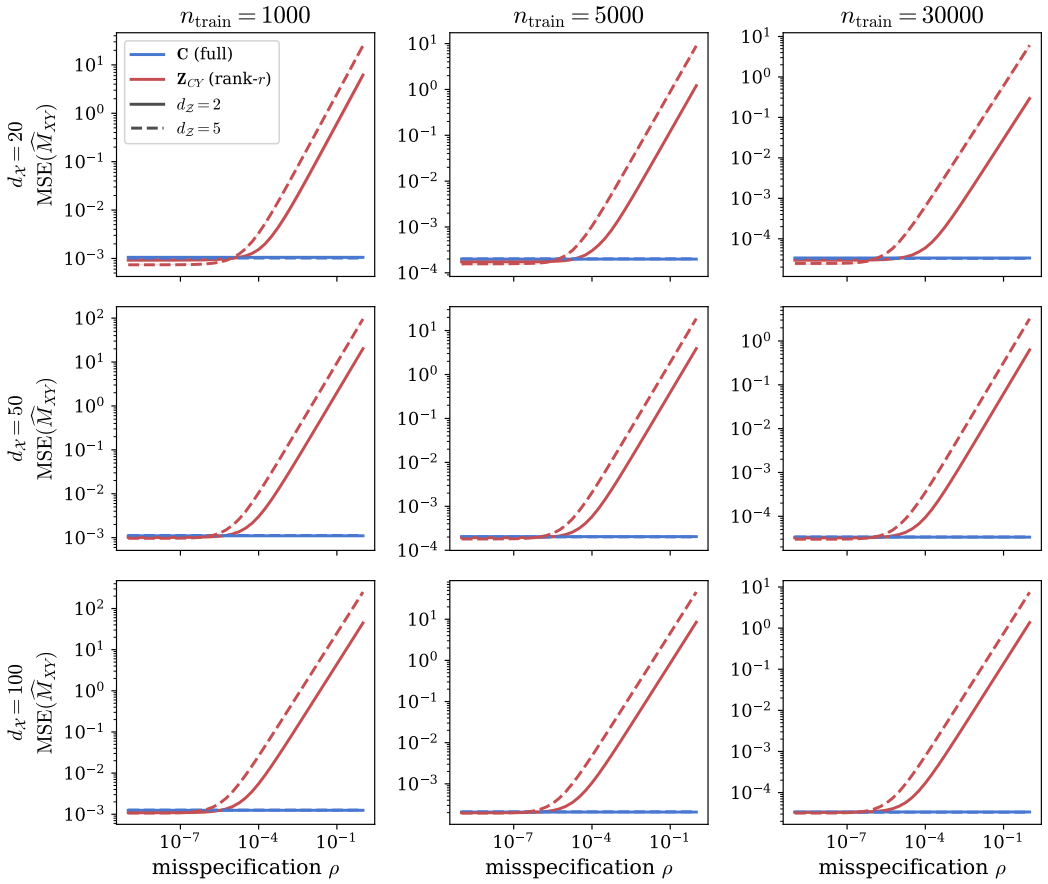


Figure 9: Backdoor adjustment under approximately low-rank $M_{CY}$. MSE of $\widehat{M}_{XY}$ vs. misspecification $\rho$ (log-log; $\rho > 0$ shown) for conditioning on the full confounder $\mathbf{C}$ (blue) vs. the rank-$d_{\mathbf{Z}}$ bottleneck $\mathbf{Z}_{CY}$ (red), for $d_{\mathbf{Z}} = 2$ (solid) and $d_{\mathbf{Z}} = 5$ (dashed). Rows: $d_{\mathbf{X}}$; columns: $n_{\text{train}}$. Lines show medians and bands interquartile ranges over 10 seeds. Full conditioning is unaffected by $\rho$, while the bottleneck loses its advantage at $\rho$ of order $10^{-7}$–$10^{-5}$.

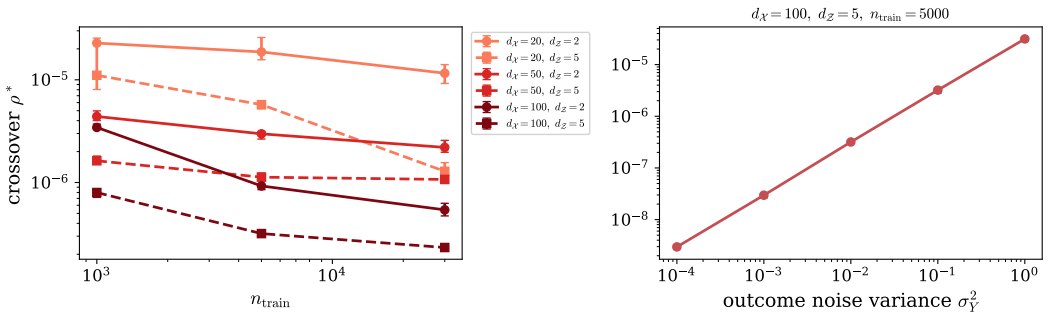


Figure 10: Crossover $\rho^*$ above which conditioning on the bottleneck is worse than conditioning on the full confounder. **Left:** $\rho^*$ vs. $n_{\text{train}}$; $\rho^*$ decreases with sample size because the misspecification bias does not vanish. **Right:** $\rho^*$ vs. outcome noise variance $\sigma_Y^2$ ($d_{\mathbf{X}} = 100$, $d_{\mathbf{Z}} = 5$, $n_{\text{train}} = 5000$); $\rho^*$ scales linearly with $\sigma_Y^2$. Markers show medians and error bars interquartile ranges of the per-seed crossover over 10 seeds.

conditioning, lies between $2.3 \cdot 10^{-7}$ and $2.3 \cdot 10^{-5}$. For the configuration of Figure 5 ($d_{\mathbf{X}} = 100$, $d_{\mathbf{Z}} = 5$, $n_{\text{train}} = 5000$), $\rho^* = 3.2 \cdot 10^{-7}$, and at $\rho = 0.02$ the MSE of $\widehat{M}_{XY}$ is 0.92 with the bottleneck versus $2.1 \cdot 10^{-4}$ with full conditioning. The unmodelled component $\mathbf{C}E$ affects the bottleneck estimator in two ways: it enters the residual of $\mathbf{Y}$ as additional outcome noise, inflating the variance, and its part that is correlated with $\mathbf{X}$ given $\mathbf{Z}_{CY}$ biases $\widehat{M}_{XY}$. Unlike the variance, the bias does not vanish with $n_{\text{train}}$, so $\rho^*$ decreases with sample size (by a factor of 1.5 to 8.7 from $n_{\text{train}} = 10^3$ to $3 \cdot 10^4$; Figure 10, left). The decrease is strongest where the bias dominates: for $d_{\mathbf{X}} = 20$, $d_{\mathbf{Z}} = 5$, it accounts for 89% of the bottleneck MSE at $n_{\text{train}} = 3 \cdot 10^4$ and $\rho = 0.01$, compared with less than 1% for $d_{\mathbf{X}} = 100$. Finally, $\rho^*$ scales linearly with the outcome noise variance, from $3.0 \cdot 10^{-9}$ at $\sigma_Y^2 = 10^{-4}$ to $3.1 \cdot 10^{-5}$ at $\sigma_Y^2 = 1$ (Figure 10, right): what matters is the size of the perturbation relative to the noise, not relative to the low-rank part. The efficiency gains of bottleneck adjustment therefore hinge on the low-rank assumption holding essentially exactly. If it cannot be justified, one should condition on the full variables, which remain valid.

### D.5 Sample cost of estimating bottlenecks

The efficiency results of Section 4 assume oracle bottlenecks. To quantify the cost of estimating them, we repeat the mediator and backdoor experiments of Figure 5 ($d = 100$, $d_{\mathbf{Z}} = 5$), using 10 seeds under a fixed budget of $N$ samples and compare four strategies: (i) *full* conditioning on the high-dimensional mediator or confounder, using all $N$ samples; (ii) a naive *split*, estimating the bottleneck on $N/2$ samples and the effect on the remaining $N/2$; (iii) *cross-fitting*, which additionally swaps the roles of the two halves and averages, as is standard in double machine learning; (iv) the *oracle* bottleneck, using all $N$ samples. Table 2 reports mean $\pm$ std of the MSE of $\widehat{M}_{XY}$. Each seed draws a new mechanism, and the std mainly reflects variation across these draws. Since all strategies use the same data within a seed, we compare them seed by seed below.

Table 2: MSE of $\widehat{M}_{XY}$ under a fixed sample budget $N$ (mean $\pm$ std over 10 seeds).

| $N$ | Full | Split | Cross-fit | Oracle |
|---|---|---|---|---|
| *Mediator* | | | | |
| $10^3$ | $3.3 \pm 0.64$ | $6.2 \pm 1.4$ | $3.1 \pm 0.52$ | $2.8 \pm 0.47$ |
| $5 \cdot 10^3$ | $0.52 \pm 0.046$ | $0.98 \pm 0.060$ | $0.50 \pm 0.053$ | $0.49 \pm 0.049$ |
| $3 \cdot 10^4$ | $0.087 \pm 0.014$ | $0.17 \pm 0.027$ | $0.085 \pm 0.013$ | $0.084 \pm 0.013$ |
| $10^5$ | $0.025 \pm 0.0046$ | $0.048 \pm 0.0075$ | $0.025 \pm 0.0039$ | $0.025 \pm 0.0038$ |
| $3 \cdot 10^5$ | $0.0090 \pm 0.0020$ | $0.017 \pm 0.0030$ | $0.0087 \pm 0.0020$ | $0.0087 \pm 0.0020$ |
| *Backdoor* ($\times 10^{-5}$) | | | | |
| $10^3$ | $130 \pm 1.3$ | $4700 \pm 4800$ | $1900 \pm 1600$ | $110 \pm 1.4$ |
| $5 \cdot 10^3$ | $21 \pm 0.42$ | $590 \pm 1100$ | $650 \pm 900$ | $19 \pm 0.41$ |
| $3 \cdot 10^4$ | $3.4 \pm 0.064$ | $9.9 \pm 8.0$ | $4.6 \pm 3.6$ | $3.2 \pm 0.065$ |
| $10^5$ | $1.0 \pm 0.0094$ | $1.9 \pm 0.021$ | $0.95 \pm 0.0093$ | $0.95 \pm 0.0095$ |
| $3 \cdot 10^5$ | $0.33 \pm 0.0043$ | $0.63 \pm 0.0067$ | $0.31 \pm 0.0042$ | $0.31 \pm 0.0042$ |

The naive split is consistently worst, as it halves the data available for effect estimation. In the mediator setting, cross-fitting is close to the oracle at all budgets and indistinguishable from it from $N = 10^4$ onward. Its advantage over full conditioning is small and within seed variability, mirroring the modest gap between oracle and full conditioning in this configuration. In the backdoor setting, estimating the rank-5 bottleneck from $N/2$ samples is unreliable at small budgets, driven by a few mechanism draws for which the bottleneck subspace is hard to estimate. Cross-fitting is therefore worse than full conditioning for $N \leq 5 \cdot 10^3$ and comparable at $N = 3 \cdot 10^4$. From $N = 10^5$ onward, it is within 1% of the oracle and improves on full conditioning for all 10 seeds.

Estimated bottlenecks thus retain the benefit of oracle bottlenecks once the budget suffices to estimate them, but can be costly below that. This cost is paid once per mechanism: since bottlenecks are properties of mechanisms, an estimated bottleneck can be reused across all queries involving the same mechanism, or estimated from a larger auxiliary dataset, whereas full conditioning incurs the high-dimensional cost on every query. Moreover, only the bottlenecks relevant to a given query need to be estimated, so this cost does not scale with the total number of edges in the graph.

# E Experimental details

## E.1 General data generation process

We generate data by first randomly sampling an SCBM $\mathfrak{C}$ and then drawing $n$ samples from the joint distribution induced by $\mathfrak{C}$. To sample an SCBM, we set the number of vertices $\mathcal{V}$, the internal dimension of the nodes $d_{\mathbf{X}}$ and the dimension of the bottleneck spaces $d_{\mathbf{Z}}$. With these fixed parameters, first the graph $\mathcal{G}$ is sampled from an Erdős–Rényi model (Erdős and Rényi, 1959) with edge probability $p = 0.7$. Then, for each node $\mathbf{X}_j$ we sample the distribution of its respective noise term from a Markov Random Field whose joint distribution is a Gaussian and internal dynamics are described by the Langevin diffusion (Lauritzen and Richardson, 2002) with dimension $d_{\mathbf{X}}$. For each edge in $\mathcal{G}$, we randomly sample both a bottleneck function $b_{ij}$, as well as an effect function $f_{ij}$. For linear models, we sample $b_{ij}$ by sampling a random matrix $W_{ij}^b \in [0,1]^{d_{\mathbf{X}} \times d_{\mathbf{Z}}}$ and $f_{ij}$ by sampling a random matrix $W_{ij}^f \in [0,1]^{d_{\mathbf{Z}} \times d_{\mathbf{X}}}$, both with rank $= d_{\mathbf{Z}}$. For nonlinear models, $b_{ij}$ and $f_{ij}$ are implemented by randomly initialized, multilayer perceptrons (MLPs). Details for each experimental setup are provided in Appendix E.3. See also Appendix F for a detailed discussion on sampling nonlinear functions. Unless specified otherwise, by default we use the parameters $|\mathcal{V}| = 10$, $d_{\mathbf{X}} = 5$, $d_{\mathbf{Z}} = 2$ and $n = 30000$ for linear models and $n = 50000$ for nonlinear models, respectively. We conduct experiments where we vary one of these parameters while keeping all others fixed.

## E.2 Detailed estimation procedure

As per Definition 2.3, we consider the case where there is a separate bottleneck for each parent of an endogenous node. Graphically, this means there is a bottleneck space for each edge in the graph. Since each edge $\mathbf{X}_i \to \mathbf{X}_j$ can be decomposed into a bottleneck function $b_{ij}$, mapping to the bottleneck space $\mathcal{Z}_{ij}$, and an effect function $f_{ij}$ mapping from $\mathcal{Z}_{ij}$ to $\mathcal{X}_j$, the targets of the estimation procedure are $b_{ij}$ and $f_{ij}$. We recover the joint map $m_{ij} := f_{ij} \circ b_{ij} : \mathcal{X}_i \times \mathcal{H}_j \to \mathcal{X}_j$ by fitting an estimator from $\mathbf{X}_i$ to $\mathbf{X}_j$. By Lemma 3.2, from the estimated map $\hat{m}_{ij}$ we recover both $b_{ij}$ and $f_{ij}$ up to an invertible map $\psi_{ij}$, i.e., $\hat{b}_{ij} := \psi_{ij} \circ b_{ij}$ and $\hat{f}_{ij} := f_{ij} \circ \psi_{ij}^{-1}$. How $\hat{b}_{ij}$ and $\hat{f}_{ij}$ are recovered from $\hat{m}_{ij}$ is described in the following. Code to reproduce all experimental results is available at `https://github.com/simonbing/StructuralCausalBottleneckModels`. We implement our estimation pipeline using the JAX machine learning library (Bradbury et al., 2018).

**Conditioning sets.** Estimating $b_{ij}$ for edge $\mathbf{X}_i \to \mathbf{X}_j$ reduces to estimating the direct effect $\mathbf{X}_i \to \mathbf{X}_j$, so any conditioning set valid for that problem is valid here. The parent variables $\{\mathbf{X}_k\}_{k \in \mathrm{pa}(i)} \cup \{\mathbf{X}_\ell\}_{\ell \in \mathrm{pa}(j)}$ always constitute a valid, if high-dimensional, conditioning set. When bottleneck variables from prior estimation steps are available, they can be used instead. The results of Sections 4.2.1 and 4.2.2 govern which bottlenecks are valid substitutes: for a backdoor path through a confounder $\mathbf{X}_k$, either the treatment-side bottleneck $\mathbf{Z}_{ki}$ or the outcome-side bottleneck $\mathbf{Z}_{kj}$ is valid (Lemma 4.6); for a frontdoor path $\mathbf{X}_i \to \mathbf{X}_m \to \mathbf{X}_j$ through a mediator, only the outcome-side bottleneck $\mathbf{Z}_{mj}$ is valid—the treatment-side $\mathbf{Z}_{im}$ is a function of $\mathbf{X}_i$ and cannot block the indirect path (cf. Lemma 4.4).

**Linear vs. nonlinear estimators.** Let $\mathbf{X}_i \in \mathbb{R}^{d_i}, \mathbf{X}_j \in \mathbb{R}^{d_j}$ and $\mathbf{Z}_{ij} \in \mathbb{R}^{d_{ij}}$. Fitting an estimator between $\mathbf{X}_i$ and $\mathbf{X}_j$ results in a map $\hat{m}_{ij} : \mathbb{R}^{d_i} \mapsto \mathbb{R}^{d_j}$ from which we must extract $\hat{b}_{ij} : \mathbb{R}^{d_i} \mapsto \mathbb{R}^{d_{ij}}$ and $\hat{f}_{ij} : \mathbb{R}^{d_{ij}} \mapsto \mathbb{R}^{d_j}$. In the linear case, fitting an estimator returns a weight matrix $\widehat{M}_{ij} \in \mathbb{R}^{d_i \times d_j}$, which by assumption has rank $d_{ij} \ll d_i, d_j$. Extracting $\hat{b}_{ij}$ and $\hat{f}_{ij}$ amounts to finding a matrix factorization $\widehat{M}_{ij} = \widehat{W}_{ij}^b \widehat{W}_{ij}^f$, where $\widehat{W}_{ij}^b \in \mathbb{R}^{d_i \times d_{ij}}$ and $\widehat{W}_{ij}^f \in \mathbb{R}^{d_{ij} \times d_j}$. We recover $\widehat{W}_{ij}^b$ by selecting $d_{ij}$ linearly independent columns from $\widehat{M}_{ij}$ and set $\widehat{W}_{ij}^f = \widehat{W}_{ij}^{b+} \widehat{M}_{ij}$, where $(\cdot)^+$ denotes the Moore-Penrose inverse (Moore, 1920; Bjerhammar, 1951; Penrose, 1955). In the nonlinear case, we employ an encoder-decoder architecture: an encoder network $\hat{b}_\theta$ maps from $\mathbf{X}_i$ to $\mathbf{Z}_{ij}$ and a decoder network $\hat{f}_\phi$ maps from $\mathbf{Z}_{ij}$ to $\mathbf{X}_j$.

**Estimating all bottlenecks.** Note that for a specific causal query, one need only estimate the bottlenecks on paths relevant to that query. The procedure below is needed only when the goal is to estimate *all* bottlenecks in a given SCBM.

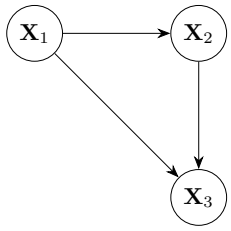


Figure 11: Three-node, fully connected graph.

If our goal is to estimate *all* bottlenecks in a given SCBM, the order of estimation matters: any bottleneck used as a conditioning variable must already have been estimated. We propose to loop over target nodes in causal order and, for each target, over its parents in *reverse* causal order. To illustrate, consider the three-node graph in Figure 11 with bottlenecks $\mathbf{Z}_{12}, \mathbf{Z}_{13}, \mathbf{Z}_{23}$. We first estimate $\mathbf{Z}_{12}$ (no conditioning needed), then estimate $\mathbf{Z}_{23}$ conditioning on $\hat{\mathbf{Z}}_{12}$ to block the backdoor path $\mathbf{X}_2 \leftarrow \mathbf{X}_1 \rightarrow \mathbf{X}_3$ (valid by treatment-side sufficiency, Lemma 4.6(b)), and finally estimate $\mathbf{Z}_{13}$ conditioning on $\hat{\mathbf{Z}}_{23}$ to block the frontdoor path $\mathbf{X}_1 \rightarrow \mathbf{X}_2 \rightarrow \mathbf{X}_3$ (valid by outcome-side sufficiency, Lemma 4.6(a) applied to the mediator $\mathbf{X}_2$). At each step, all required conditioning bottlenecks have already been estimated.

This generalizes to arbitrary DAGs: backdoor paths through $\mathbf{X}_k \in \mathrm{pa}(\mathbf{X}_i)$ are blocked by $\mathbf{Z}_{ki}$, available because $\mathbf{X}_i$ was the target in a prior outer-loop iteration; frontdoor paths through $\mathbf{X}_\ell \in \mathrm{pa}(\mathbf{X}_j)$ are blocked by $\mathbf{Z}_{\ell j}$, available because $\mathbf{X}_\ell$ is later in causal order and was therefore processed first in the inner reverse loop. Pseudocode is given in Algorithm 1; the bottleneck estimate is $\hat{\mathbf{Z}}_{ij} := \hat{b}_{ij}(\mathbf{X}_i)$.

**Algorithm 1** Estimate bottleneck & effect functions

**Input:** SCBM $\mathfrak{C} = \langle \mathcal{G} = (\mathcal{V}, \mathcal{E}), \mathcal{X}, \mathcal{Z}, \mathcal{H}, \boldsymbol{\eta}, \mathcal{B}, \mathcal{F}, \mathbf{X} \rangle$, causal order $C[1..|\mathcal{V}|]$, assumed bottleneck dimension $d_{\hat{\mathbf{Z}}}$

**for** $j$ **in** $C$ **do** ▷ Outer loop over target nodes.
  $target \leftarrow \mathbf{X}_j$
  **if** $target$ is root node **then**
    pass ▷ Root nodes have no bottlenecks.
  **else**
    **for** $i$ **in** reversed$(C[:j])$ **do** ▷ Inner loop over source nodes.
      $source \leftarrow \mathbf{X}_i$
      $\hat{\mathbf{Z}}_{cond} \leftarrow GetCondSet(\mathfrak{C}, i, j)$ ▷ According to procedure described in Appendix E.2.
      $\hat{m}_{ij} \leftarrow fit(source, target, \hat{\mathbf{Z}}_{cond})$ ▷ Using linear or nonlinear regressor.
      $\hat{b}_{ij}, \hat{f}_{ij} \leftarrow factorize(\hat{m}_{ij}, d_{\hat{\mathbf{Z}}})$
      Cache $\hat{b}_{ij}, \hat{f}_{ij}$
    **end for**
  **end if**
**end for**

**Proposition E.1** (Correctness of Algorithm 1)**.** *Let $\mathfrak{C}$ be a factored SCBM (Definition 2.3) with additive noise and almost-surely injective effect functions $\tilde{f}_{ij}$ for all $(i,j) \in \mathcal{E}$. If the assumed bottleneck dimensions equal the true ones, then at the population level Algorithm 1 returns, for each $(i,j) \in \mathcal{E}$,*

$$\hat{b}_{ij} = \psi_{ij} \circ b_{ij}, \qquad \hat{f}_{ij} = f_{ij} \circ \psi_{ij}^{-1},$$

*for some invertible $\psi_{ij}$.*

*Proof.* We proceed by induction on the iteration count of Algorithm 1. Index iterations by pairs $(j,i)$ where $j$ is the outer-loop index and $i$ the inner-loop index; order them lexicographically with $j$ increasing and, within fixed $j$, $i$ decreasing (matching the reverse causal order of the inner loop).

**Induction hypothesis.** All bottleneck estimates $\hat{b}_{k\ell}$ cached in iterations preceding $(j, i)$ satisfy $\hat{b}_{k\ell} = \psi_{k\ell} \circ b_{k\ell}$ for some invertible $\psi_{k\ell}$.

**Induction step.** Consider estimating $b_{ij}$ in iteration $(j, i)$. The conditioning set $\hat{\mathbf{Z}}_{\text{cond}}$ returned by `GetCondSet` consists of two types of previously estimated bottlenecks.

(i) *Backdoor conditioning*: for each $\mathbf{X}_k \in \text{pa}(\mathbf{X}_i)$, `GetCondSet` returns $\hat{\mathbf{Z}}_{ki}$, estimated during outer-loop iteration $i$, which precedes $j$ in causal order so $(i, \cdot) < (j, i)$ in our ordering. By the induction hypothesis, $\hat{b}_{ki} = \psi_{ki} \circ b_{ki}$ for some invertible $\psi_{ki}$, so $\hat{\mathbf{Z}}_{ki}$ is an invertible reparametrization of $\mathbf{Z}_{ki}$. Since $\mathbf{Z}_{ki}$ satisfies treatment sufficiency ($\mathbf{X}_i \perp \mathbf{X}_k \mid \mathbf{Z}_{ki}$; Lemma 4.6(b)), and invertible reparametrization preserves conditional independence, $\hat{\mathbf{Z}}_{ki}$ is a valid conditioning variable for blocking the backdoor path through $\mathbf{X}_k$.

(ii) *Frontdoor conditioning*: for each $\mathbf{X}_\ell \in \text{pa}(\mathbf{X}_j)$ with $\ell \neq i$ such that $\mathbf{X}_i \to \cdots \to \mathbf{X}_\ell$ in $\mathcal{G}$, `GetCondSet` returns $\hat{\mathbf{Z}}_{\ell j}$. Since $\mathbf{X}_\ell$ is a descendant of $\mathbf{X}_i$, we have $i < \ell < j$ in causal order. The inner loop for target $j$ processes sources in reverse causal order, so $\mathbf{X}_\ell$ (with $\ell > i$) is processed in iteration $(j, \ell)$ before $(j, i)$. By the induction hypothesis, $\hat{b}_{\ell j} = \psi_{\ell j} \circ b_{\ell j}$ for some invertible $\psi_{\ell j}$, so $\hat{\mathbf{Z}}_{\ell j}$ is an invertible reparametrization of $\mathbf{Z}_{\ell j}$. Since $\mathbf{Z}_{\ell j}$ satisfies outcome sufficiency for the indirect path $\mathbf{X}_i \to \mathbf{X}_\ell \to \mathbf{X}_j$ (Lemma 4.4), and invertible reparametrization preserves this property, $\hat{\mathbf{Z}}_{\ell j}$ is a valid conditioning variable for blocking the frontdoor path through $\mathbf{X}_\ell$.

Together, (i) and (ii) imply that $\hat{\mathbf{Z}}_{\text{cond}}$ is a valid adjustment set for the direct effect $\mathbf{X}_i \to \mathbf{X}_j$. At the population level, the regression of $\mathbf{X}_j$ on $(\mathbf{X}_i, \hat{\mathbf{Z}}_{\text{cond}})$ therefore identifies the joint map $m_{ij} = f_{ij} \circ b_{ij}$. Since the effect functions are almost-surely injective and the SCBM has additive noise, Lemma 3.2 applies: factorizing the estimated map yields $\hat{b}_{ij} = \psi_{ij} \circ b_{ij}$ and $\hat{f}_{ij} = f_{ij} \circ \psi_{ij}^{-1}$ for some invertible $\psi_{ij}$.

**Base case.** The first iteration is $(j_{\min}, i_{\max})$, where $j_{\min}$ is the first non-root node in causal order and $i_{\max}$ is its highest-indexed predecessor. Since $j_{\min}$ is the first non-root, all nodes preceding it in causal order are roots, so $\text{pa}(i_{\max}) = \emptyset$ (no backdoor conditioning needed) and there are no intermediary nodes on any path from $\mathbf{X}_{i_{\max}}$ to $\mathbf{X}_{j_{\min}}$ with higher causal index than $i_{\max}$ (no frontdoor conditioning needed). No prior estimates are required; Lemma 3.2 applies directly.

By induction, all bottleneck estimates satisfy the identifiability guarantee. □

### E.3 Data generation details & hyperparameters

#### E.3.1 Sampling nonlinear bottleneck and effect functions

To sample nonlinear bottleneck and effect functions, we randomly initialize MLPs with two hidden layers with $d_{\text{hidden}} = d_{\mathbf{X}}$ and orthogonalize as described in Appendix F. We use the swish nonlinearity. Hyperparameters of the estimator are described in Table 3.

#### E.3.2 Estimation under misspecification

The data generation process and hyperparameters for the misspecification experiments follow those of the identifiability experiments described above.

#### E.3.3 Partitioning and intervention selection

The input space partitioning experiment (Figure 4) uses a three-node nonlinear SCBM with $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 \in \mathbb{R}^2$ and graph $\mathbf{X}_1 \to \mathbf{X}_2$, $\mathbf{X}_1 \to \mathbf{X}_3$. The two ground-truth bottleneck functions are $b_{12}(\mathbf{x}) = \|\mathbf{x}\|_2$ (Euclidean norm, capturing radial distance) and $b_{13}(\mathbf{x}) = 2 \cdot \mathbb{1}[x^{(1)} > 0] + \mathbb{1}[x^{(2)} > 0]$ (quadrant index, taking values in $\{0, 1, 2, 3\}$); both map to a one-dimensional bottleneck space. Effect functions are linear: $f_{12}(\mathbf{z}) = \mathbf{z}W_{12}$ and $f_{13}(\mathbf{z}) = \mathbf{z}W_{13}$ with $W_{12}, W_{13} \in [0, 1]^{1\times 2}$ sampled uniformly. The root $\mathbf{X}_1 \sim \mathcal{N}(0, I_2)$, and downstream noise terms $\eta_2, \eta_3 \sim \mathcal{N}(0, 0.09\, I_2)$. We train on $n = 10000$ samples.

Table 3: Hyperparameters for identifiability and misspecification experiments.

| HYPERPARAMETER | VALUE |
|---|---|
| ENCODER HIDDEN LAYER SIZES | $[256, 128, 64, 64, 32]$ |
| DECODER HIDDEN LAYER SIZES | $[32, 64, 64, 128, 256]$ |
| EPOCHS | 500 |
| BATCH SIZE | 1024 |
| OPTIMIZER | ADAMW (Loshchilov and Hutter, 2019) |
| LEARNING RATE SCHEDULER | COSINE DECAY W/ WARMUP |
| MAX. LEARNING RATE | 1e−5 |
| MIN. LEARNING RATE | 1e−7 |
| WARMUP STEPS | 1000 |
| DECAY STEPS | 10000 |

Each bottleneck is estimated with an encoder–decoder network: an MLP encoder (hidden layer sizes $[64, 32]$, swish activations) mapping $\mathbf{X}_1 \in \mathbb{R}^2$ to a one-dimensional latent encoding, followed by a single linear decoder. Training uses AdamW (Loshchilov and Hutter, 2019) with learning rate $10^{-3}$; $b_{12}$ is trained for 2000 epochs and $b_{13}$ for 5000 epochs to account for the harder discrete target.

#### E.3.4 Adjustment experiments

The mediator and backdoor adjustment experiments (Figure 5) use a linear three-node SCBM with $d_{\mathbf{X}} = 100$ and $d_{\mathbf{Z}} = 5$. All weight matrices are sampled from Uniform$(0.5, 1.5)$ and are of full rank $d_{\mathbf{Z}}$. The root node $\mathbf{X}_0$ (the treatment in the mediator setting, the confounder in the backdoor setting) has noise covariance $I_{100}$; all other nodes have noise covariance $0.01\, I_{100}$.

**Mediator adjustment.** The graph is $\mathbf{X}_0 \to \mathbf{X}_1 \to \mathbf{X}_2$, $\mathbf{X}_0 \to \mathbf{X}_2$. The target quantity is the direct effect $M_{02}$. We compare conditioning on the full mediator $\mathbf{X}_1 \in \mathbb{R}^{100}$ against the oracle outcome-side bottleneck $\mathbf{Z}_{12} = b_{12}(\mathbf{X}_1) \in \mathbb{R}^5$.

**Backdoor adjustment.** The graph is $\mathbf{X}_0 \to \mathbf{X}_1 \to \mathbf{X}_2$, $\mathbf{X}_0 \to \mathbf{X}_2$, with $\mathbf{X}_0$ acting as a confounder. The target quantity is the direct effect $M_{12}$. We compare conditioning on the full confounder $\mathbf{X}_0 \in \mathbb{R}^{100}$ against the oracle outcome-side bottleneck $\mathbf{Z}_{02} = b_{02}(\mathbf{X}_0) \in \mathbb{R}^5$.

In both settings, oracle bottlenecks (i.e., the ground-truth linear projections) are used as conditioning variables, and the estimator is ordinary least squares. Results are averaged over 5 random seeds.

#### E.3.5 Approximately Low-Rank Mechanisms

We use the backdoor data-generating process of Appendix E.3.4, with identical weight sampling, so that for $\rho = 0$ the mechanisms coincide with those of the backdoor experiment for the same seed. The perturbation $E$ has i.i.d. standard Gaussian entries and is rescaled such that $\|E\|_F^2 = \rho \, \|W_{CY}^b W_{CY}^f\|_F^2$. For each seed, the mechanisms, the perturbation direction and all noise draws are held fixed across $\rho$, so that curves over $\rho$ are paired. We sweep $\rho \in \{0\} \cup \{0.02, 0.05, 0.2\} \cup \{37 \text{ log-spaced values in } [10^{-9}, 1]\}$, $d_{\mathbf{X}} \in \{20, 50, 100\}$, $d_{\mathbf{Z}} \in \{2, 5\}$ and $n_{\text{train}} \in \{10^3, 5 \cdot 10^3, 3 \cdot 10^4\}$, with 10 seeds per configuration. The confounder has noise covariance $I$, and treatment and outcome have noise covariance $0.01\, I$. For the noise-level sweep, we fix $d_{\mathbf{X}} = 100$, $d_{\mathbf{Z}} = 5$, $n_{\text{train}} = 5000$ and $\sigma_X^2 = 0.01$, and vary $\sigma_Y^2 \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. Both estimators are OLS with intercept of $\mathbf{Y}$ on $(\mathbf{X}, \mathbf{Z}_{CY})$ and on $(\mathbf{X}, \mathbf{C})$, respectively, using the oracle $W_{CY}^b$ of the low-rank part. The crossover $\rho^*$ is computed per seed by log-linear interpolation of the log-MSE ratio between adjacent grid points; we report its median and interquartile range over seeds.

### E.4 Compute resources

For the identifiability experiments involving estimating nonlinear bottleneck functions we utilize an Nvidia A100 GPU; our experiments required approximately 200 GPU hours. All other experiments were performed on a local machine with an Apple M1 chip, requiring in the order of five to ten minutes, each.

## F On randomly sampling nonlinear functions

While visualizing ground-truth bottleneck spaces in nonlinear SCBMs, we stumbled across an unexpected phenomenon: randomly sampled ground-truth bottleneck spaces $\mathcal{Z}$ seemed to collapse to a one-dimensional subspace. Recalling that bottleneck functions map from observed to bottleneck spaces, i.e., $b_{ij} : \mathcal{X}_i \to \mathcal{Z}_{ij}$, we can see how this phenomenon has its explanation in the choice of ground-truth bottleneck functions. Following common practice, we sample these nonlinear bottleneck functions by randomly initializing MLPs.

A nonlinear function $m_{ij} : \mathbb{R}^{d_i} \to \mathbb{R}^{d_j}$ is implemented by an MLP with input dimension $d_i$ and output dimension $d_j$. Such an MLP is nothing more than the product of randomly sampled matrices—all layers with randomly sampled weights—composed with element-wise nonlinearities. Let us disregard the nonlinearities for now and take a closer look at products of random matrices, which we write as

$$M = \prod_{i=1}^{h} W_i,$$

where, for simplicity, we assume all factors $W_i$ to have compatible dimensions and full column rank. The number of factors $h$ can be understood as the number of hidden layers in the corresponding MLP.

Let $r = \min_{i \in \{1..h\}} \text{rank}(W_i)$ denote the smallest rank of any factor in this product. We know that for the product it holds that $r = \text{rank}(M)$. However, empirically, we observe that the numerical rank of $M$ quickly approaches 1. Quickly, in this context, means that this behaviour is observed for as few as three factors. This phenomenon also seems robust to the base distribution from which entries are chosen, as well as the size of the factor matrices. Understanding and characterizing this behaviour is active work in progress in an orthogonal research project.

While we cannot yet fully explain the origins of this collapse, we do propose an effective remedy in practice. When sampling the random matrices corresponding to individual matrices, we append an orthogonalization step. This is achieved by taking the $QR$-decomposition of the factor $W_i$ and taking $Q$ as the layer matrix. We thereby guarantee that we have the same span as the originally sampled $W_i$, but with an orthogonal basis. Empirically, this results in much more numerically stable full-rank products.

The technical difference between our proposed, new approach and the original one (without orthogonalization) is that we have proper control of the dimension of the space to which our MLP maps. Neither of the two is per se right or wrong, but their difference is worth being aware of. Rather, we believe our findings raise the question of how to "properly" sample nonlinear functions in the first place. In our experiments, we see that the orthogonalization approach leads to significantly more complex functions (requiring more samples and resulting in slightly lower identifiability metrics), but how representative such synthetic functions are of practical problems remains an open question.

## G Declaration of LLM usage

This work used LLMs to assist with some of the proofs (primarily in Section 4), as well as for some of the code for the experiments. All proofs were verified for correctness by a human.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We state our main contributions at the end of Section 1 and support these claims in Section 3 and Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations of our work in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All theoretical results and proofs are presented in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all necessary details to reproduce our experiments in Appendix E. All code to implement our experiments, along with instructions to reproduce the results we show is available at https://github.com/simonbing/StructuralCausalBottleneckModels.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code together with instructions on how to reproduce our results is available at https://github.com/simonbing/StructuralCausalBottleneckModels.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We provide all experimental details in Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Figure 2 and all ables show mean ± std; Figures 9 and 10 show median and IQR over 10 seeds; all other Figures show 95% confidence interval bars, calculated over five random seeds via bootstrapping.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details on compute resources in Appendix E.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in this work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work introduces and discusses a novel modeling framework, which amounts to basic research. While we hope this work can contribute to the development of practical methods with a positive societal impact, the results we discuss in this work are not expected to have a direct societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release any code or assets that have a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the authors of all used software packages.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code and its documentation is available at `https://github.com/simonbing/StructuralCausalBottleneckModels`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We declare LLM usage in Appendix G.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.