Structural Causal Bottleneck Models

Simon Bing[*,1,2] Jonas Wahl[*,3] Jakob Runge[2,4]

[*]Equal contribution.
[1]Technische Universität Berlin, Germany
[2]Department of Computer Science, University of Potsdam, Germany
[3]German Research Centre for AI (DFKI), Saarbrücken, Germany
[4]ScaDS.AI Dresden/Leipzig, TU Dresden, Germany

Abstract

We introduce structural causal bottleneck models (SCBMs), a novel class of structural causal models. At the core of SCBMs lies the assumption that causal effects between high-dimensional variables only depend on low-dimensional summary statistics, or *bottlenecks*, of the causes. SCBMs provide a flexible framework for task-specific dimension reduction while being estimable via standard, simple learning algorithms in practice. In addition to an analysis of identifiability in SCBMs, we provide experimental results evidencing that we can estimate bottlenecks in practice. We also demonstrate the benefit of bottlenecks for effect estimation in low-sample transfer learning settings.

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]. These phenomena are modeled as high-dimensional random vectors that are consequently simplified, abstracted or transformed.

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 $\mathbf{X}_i := m_i(\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_k}, \boldsymbol{\eta}_i)$, one for each quantity of interest, that describe how each variable is brought about by its *causal parents* $\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_k}$ and an exogeneous noise term $\boldsymbol{\eta}_i$ through the mechanism function m_i. Although in most applications of SCMs the variables $\mathbf{X}_i$ and noise-terms $\boldsymbol{\eta}_i$ are assumed to be one-dimensional, the SCM setup does not necessitate this assumption. However, modeling interactions of high-dimensional vectors $\mathbf{X}_i$ may quickly become infeasible in practice without additional assumptions. Even if the mechanism functions m_i are assumed linear additive, i.e. $\mathbf{X}_i := \sum_{j \in \mathrm{pa}(i)} \mathbf{A}^i_j \mathbf{X}_j + \boldsymbol{\eta}_i$, with matrices $\mathbf{A}^i_j$, 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 dimensionality is particularly daunting. The estimation may not only require as input the treatment and outcome variables of but also confounding covariates that need be conditioned on to remove spurious correlations and that increase the dimension of the input space further.

In this work, we investigate SCMs of high-dimensional random vectors in which the causal variables only depend on their parents through low-dimensional sufficient statistics or *bottlenecks*. That is to say that we assume that for any $\mathbf{X}_i$ and any of its parents $\mathbf{X}_j$ there exists a deterministic bottleneck function b^i_j that maps $\mathbf{X}_j$ to a lower-dimensional variable $\mathbf{Z}^i_j = b^i_j(\mathbf{X}_j)$ such that

$$\mathbf{X}_i := f_i(\mathbf{Z}^i_{j_1}, \ldots, \mathbf{Z}^i_{j_k}, \boldsymbol{\eta}_i) \tag{1}$$

depends on its parents only through their bottlenecks. For linear mechanisms, this requirement translates to the assumption that the matrices $\mathbf{A}^i_j$ are of low rank compared to the dimensions of $\mathbf{X}_i$ and $\mathbf{X}_j$.

The bottleneck assumption seems reasonable when modeling causal interactions between high-dimensional phenomena, in which a causal child does not depend on all information encoded in its parents but on emergent properties— captured for instance by a weighted average or specific system states. To model rainfall patterns over West Africa it may be sufficient to include information on whether El Niño Southern Oscillation (ENSO) is in an El Niño or a La Niña phase rather than modeling the full temperature distribution over the Pacific Ocean. However, reducing dimensions before estimating causal effects can discard or misidentify important information [Wahl et al., 2024, Ninad et al., 2025]. Further, different children may rely on

Accepted for the 1st Workshop on Causal Abstractions and Representations (CAR) at UAI 2025 (CAR 2025).

different aspects of a parent variable. East Asian and South American rainfall patterns may respond to different ENSO region anomalies, calling for *target-dependent* dimension reduction. While *sufficient dimension reduction* has a long history [Izenman, 1975], most approaches focus on linear models with single treatment-outcome pairs [Globerson and Tishby, 2003, Li, 2007, 2018]. In addition, with the notable exception of [Aoi and Pillow, 2018], the outcome vector is typically one-dimensional and only the input may be of high-dimensionality.

We formally introduce structural causal bottleneck models (SCBMs), a class of graphical causal models that address the aforementioned shortcomings and provide a flexible framework for targeted dimension reduction for causal effects. We discuss special cases of SCBMs that are not covered in the standard causal inference literature, provide an identifiability result showing the degree to which we can learn the bottleneck variables from data, and establish a connection between SCBMs and the Information Bottleneck method of [Tishby et al., 2000]. Finally, we provide experimental evidence to support our theoretical identifiability results, and highlight the benefits of SCBMs in a transfer learning setting where joint observations of all variables are rare.

2 PRELIMINARIES

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 exogeneous (i.e. parentless) nodes, while $\mathcal{V}_{end} = \mathcal{V} \backslash \mathcal{V}_{ex}$ is the set of exogeneous nodes. We reserve the letter d to refer to dimensions of vector spaces, e.g. $\mathbb{R}^d$. We will use the convention that $d = 0$ refers to discrete spaces. Unless specified otherwise, we use the word space in the sense of measurable space, i.e. a set endowed with a σ-algebra. All maps between such spaces are assumed to be measurable.

We recall that a *structural causal model (SCM)* is a tuple $\mathfrak{A} = \langle \mathcal{G}, \mathcal{X}, \mathcal{H}, \boldsymbol{\eta}, \mathcal{M}, \mathbf{X} \rangle$ 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 terms $\boldsymbol{\eta} = (\boldsymbol{\eta}_i)_{i \in \mathcal{V}}$, where $\boldsymbol{\eta}_i$ takes values in $\mathcal{H}_i$.
- a family $\mathcal{M} = (m_j)_{j \in \mathcal{V}_{end}}$ of mechanism functions $m_j : \left(\prod_{i \in \mathrm{pa}(j)} \mathcal{X}_i\right) \times \mathcal{H}_j \to \mathcal{X}_j$;
- a family of random vectors $\mathbf{X} = (\mathbf{X}_i)_{i \in \mathcal{V}}$ that solves the structural assignments

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

3 DEFINITION OF STRUCTURAL CAUSAL BOTTLENECK MODELS

We first define structural causal bottleneck models in full generality. Afterwards, we will add additional assumptions that narrow down the model class.

Definition 1. 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 $\boldsymbol{\eta} = (\boldsymbol{\eta}_i)_{i \in \mathcal{V}}$. We assume that each $\boldsymbol{\eta}_i$ 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 $\mathfrak{C} = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}, \mathcal{H}, \boldsymbol{\eta}, \mathcal{B}, \mathcal{F}, \mathbf{X} \rangle$ straightforwardly induces an SCM $\mathfrak{A}(\mathfrak{C}) = \langle \mathcal{G}, \mathcal{X}, \boldsymbol{\eta}, \mathcal{M}, \mathbf{X} \rangle$ in the classical sense 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*.

Definition 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, structural equivalent SCBMs share the same interventional distributions (interventional equivalence) and observational distribution (observational equivalence).

We now introduce additional assumptions that reduce the degrees of freedom of SBMs to render them more practical.

In Definition 1, the parents of any endogeneous node j are allowed to mix arbitrarily in the bottleneck space $\mathcal{Z}_j$. A reasonable assumption to impose on such models is that the bottleneck can be subdivided into separate bottlenecks for each parent.

Assumption 1. (a) Each bottleneck is *factored* in the sense that $\mathcal{Z}_j$ and b_j can be decomposed as $\mathcal{Z}_j = \prod_{i \in \mathrm{pa}(j)} \mathcal{Z}_{(i,j)}$ and

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

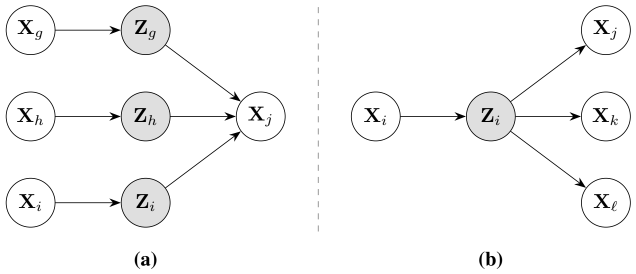

Figure 1: Examples of **(a)** factored bottleneck and effect functions and **(b)** intrinsic bottlenecks.

with $\dim \mathcal{Z}_{(i,j)} \leq d_i$. In other words, we assume that there is a separate bottleneck space for every parent of the endogeneous 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 $f_{i,j} : \mathcal{Z}_{(i,j)} \to \mathcal{X}_j$ such that

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

We will call SCBMs for which both (a) and (b) hold *factored SCBMs*.

We illustrate Assumption 1 (a) and (b) in Figure 1(a).

Intrinsic Structural Bottleneck Models Factored bottleneck models are still very flexible and allow for any variable $\mathbf{X}_i$ to affect its children $\mathbf{X}_k,\ k \in \mathrm{ch}(i)$ through different bottleneck space and different bottleneck functions. In other words, the bottleneck variables $\mathbf{Z}_{(i,k)} = b_{(i,k)}(\mathbf{X}_i)$ and $\mathbf{Z}_{(i,k')} = b_{(i,k')}(\mathbf{X}_i)$ do not need to be related and can be of different dimensions for different children $k \neq k'$. On the other hand, we might often believe that there exists an underlying low-dimensional *emergent quantity* that describes the high-dimensional $\mathbf{X}_i$ and its effect on *all* of its targets. This can be captured by the following definition.

Definition 3 (Intrinsic bottlenecks)**.** A node $i \in \mathcal{V}$ in a factored SCBM that has children admits an intrinsic bottleneck if $\mathbf{Z}_{i,j} = \mathbf{Z}^i$ and $b_{(i,j)} = b_i$ do not depend on the child j. The effect function $f_{i,j}$ is still allowed to depend on j.

We illustrate intrinsic bottlenecks in Figure 1(b).

Equivalence Relations Induced by Bottlenecks. If (i,j) is an edge in the graph $\mathcal{G}$ underlying a factored SCBM $\mathfrak{C}$, then the function $b_{(i,j)} : \mathcal{X}_i \to \mathcal{Z}_{(i,j)}$ can be considered a quotient map that induces an equivalence relation on $\mathcal{X}_i$ by declaring two states $\mathbf{x}, \mathbf{x}' \in \mathcal{X}_i$ *bottleneck-equivalent relative to child* j if $b_{(i,j)}(\mathbf{x}) = b_{(i,j)}(\mathbf{x}')$. Thus, two states of the random vector $\mathbf{X}_i$ are bottleneck-equivalent relative to j if both lead to the same state of the bottleneck for child j. As a consequence, two bottleneck-equivalent states (w.r.t. j) are *causally equivalent* w.r.t. the random vector $\mathbf{X}_j$ in the sense of Chalupka et al. [2017], i.e. they satisfy

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

If bottlenecks are assumed intrinsic, then bottleneck-equivalence is no longer relative to a specific child, and bottleneck-equivalent states $\mathbf{x}, \mathbf{x}' \in \mathcal{X}_i$ are causally equivalent for any descendant of i, i.e.

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

for every descendant k of i.

Factorization of the Observational Distribution. In any SCM, the distribution over the observed node vectors $P = P_{\mathbf{X}}$ factorizes according to the graph $\mathcal{G}$ as

$$P(\mathbf{x}) = \prod_{i \in \mathcal{V}} P(\mathbf{x}_i | \mathbf{x}_{\mathrm{pa}(i)}).$$

In SCBMs, since $\mathbf{X}_i$ only depends on its parents through the bottleneck variable $\mathbf{Z}_{\mathrm{pa}(i)}$, this can be rewritten as

$$P(\mathbf{x}) = \prod_{i \in \mathcal{V}} P(\mathbf{x}_i | \mathbf{z}_{\mathrm{pa}(i)}).$$

In addition, if the SCBM is factored, we can further decompose this expression as

$$P(\mathbf{x}) = \prod_{i \in \mathcal{V}} P(\mathbf{x}_i | (\mathbf{z}_{(k,i)})_{k \in \mathrm{pa}(i)}).$$

The SCM over the Bottleneck Variables. In a factored SCBM, the bottleneck variables $\mathbf{Z}_{(i,j)}$ can be expressed as

$$\mathbf{Z}_{(i,j)} = b_{(i,j)} \left(\sum_{k \in \mathrm{pa}(i)} F_{(i,k)}(\mathbf{Z}_{(k,i)}) + \boldsymbol{\eta}_i \right).$$

This expression no longer contains any $\mathbf{X}$-vectors but describes $\mathbf{Z}_{(i,j)}$ fully in terms of other bottleneck nodes and noise terms. In contrast to the model over the $\mathbf{X}$-vectors the noise terms in the equations for the $\mathbf{Z}_{(i,j)}$ are no longer guaranteed to be independent. In fact $\mathbf{Z}_{(i,j)}$ and $\mathbf{Z}_{(i,k)},\ k \neq j$ share the same noise term $\boldsymbol{\eta}_i$.[1]

[1] This also does not imply that the shared noise terms necessarily induce dependence: the noise term may be two-dimensional and $b_{(i,j)}$ may only depend on the first component while $b_{(i,k)}$ depends on the second.

4 SPECIAL CASES AND EXAMPLES

The class of SCBMs is large, hence we will present several example models here that are typically not covered in the standard causal inference literature where the focus is on one-dimensional node spaces.

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 a weighted spatial means, e.g.

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

where α, β, γ are normalized to one, e.g. $\int_{\mathcal{D}_1} \alpha(y)\, dy = 1$. In this case, the spaces are $\mathcal{X}_i$ are function spaces over the respective domains, e.g. $\mathcal{X}_i = L^2(\mathcal{D}_i)$, the bottleneck spaces are one dimensional $\mathcal{Z}_i \cong \mathbb{R}$, the bottleneck functions are given by the weighted integrals. The effect functions are simple embeddings by multiplication with a constant, e.g. $F_{(1,2)} : \mathcal{Z}_{(1,2)} \to L^2(\mathcal{D}_2) : z \mapsto Az \cdot \mathbb{1}_{\mathcal{D}_2}$, 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 $\mathbf{X}_2 \in L$ 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 and Peters, 2024] 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.

5 IDENTIFIABILITY

The first straightforward observation is that it is always possible to create a new, structurally equivalent SCBM from an existing one by inserting invertible mappings on the bottleneck spaces.

Lemma 1. *Let $\mathfrak{C} = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\eta}, \mathcal{B}, \mathcal{F}, \mathbf{X} \rangle$ be an SBM. 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. The 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 2. *Let $\mathfrak{C} = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\eta}, \mathcal{B}, \mathcal{F}, \mathbf{X} \rangle$ and $\mathfrak{C}' = \langle \mathcal{G}, \mathcal{X}, \mathcal{Z}', \boldsymbol{\eta}, \mathcal{B}', \mathcal{F}', \mathbf{X} \rangle$ be two SBMs 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. 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 $\psi_j = f'^{-1}_j \circ F_j$ is well-defined and has the desired properties.

□

Injectivity of the effect functions is a natural minimality assumption that together with the surjectivity imposes our bottlenecks to be as small as possible without information loss. In the linear case, this guarantees that the rank of the linear map $f_j \circ b_j$ is $\operatorname{rank}(f_j \circ b_j) = \dim \mathcal{Z}_j$ and not lower. In other words, we can always reparametrize the bottleneck space in a different basis if we adapt the bottleneck and effect functions accordingly.

6 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 is implemented through 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 intrinsic SCBM over variables $\mathbf{X}_i$ with bottleneck variables $\mathbf{Z}_i = b_i(\mathbf{X}_i)$ for every non-sink node i. i.e.

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

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 also 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 $\mathrm{pa}(i) \subset \bigsqcup_{q<s} \mathcal{V}_q$. We then fix a causal order $(i_1,\ldots,i_n)$ respecting the causal grading and solve 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. For a full paper version of this workshop submission, we plan to investigate algorithms that solve the optimization problem, for instance those outlined in [Hassanpour et al., 2017].

7 ESTIMATING SCBMS IN PRACTICE

General Estimation Procedure. As per Assumption 1, 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 a model's graph. Since each edge between variables $\mathbf{X}_i$ and $\mathbf{X}_j$ can be decomposed into a bottleneck function b_j, that maps to the corresponding bottleneck space $\mathbf{Z}_{(i,j)}$ and an effect function f_j that maps from the bottleneck space to $\mathcal{X}_j$, the targets of the estimation procedure are the bottleneck function b_j and the effect function f_j. We recover the joint map $m_j := f_j \circ b_j : \mathcal{X}_i \times \mathcal{H}_j \to \mathcal{X}_j$ by fitting an estimator from $\mathbf{X}_i$ to $\mathbf{X}_j$. The conditioning sets required for this estimation procedure are analogous to standard causal effect. As our identifiability results presented in Section 5 tell us, from the estimated composed map $\hat{m}_j$ we are able to recover both b_j and f_j up to an invertible map ψ_j, i.e., $\hat{b}_j := \psi_j \circ b_j$ and $\hat{f}_j := f_j \circ \psi_j^{-1}$. How $\hat{b}_j$ and $\hat{f}_j$ are recovered from $\hat{m}_j$ is described in detail below.

Conditioning Sets. The procedure for estimating the bottleneck space corresponding to the edge $\mathbf{X}_i \to \mathbf{X}_j$ is analogous to estimating the direct effect of $\mathbf{X}_i$ on $\mathbf{X}_j$. As such, we can use any conditioning set for this estimation that is valid for effect estimation. Given the source node $\mathbf{X}_i$ and target node $\mathbf{X}_j$, a valid conditioning set is $\{\mathbf{X}_{\mathrm{pa}(i)}, \mathbf{X}_{\mathrm{pa}(j)}\}$. Since bottleneck variables are deterministic transforms of their parent nodes, we can equivalently use the set of (lower dimensional) bottleneck variables $\{(\mathbf{Z}_{(k,i)})_{k\in\mathrm{pa}(i)}, (\mathbf{Z}_{(\ell,j)})_{\ell\in\mathrm{pa}(j)}\}$ for conditioning.

Linear vs. Nonlinear Estimators. Let $\mathbf{X}_i \in \mathbb{R}^{d_i}, \mathbf{X}_j \in \mathbb{R}^{d_j}$ and $\mathbf{Z}_{(i,j)} \in \mathbb{R}^{d_{(i,j)}}$. Fitting an estimator between $\mathbf{X}_i$ and $\mathbf{X}_j$ results in a map $\hat{m}_j : \mathbb{R}^{d_i} \mapsto \mathbb{R}^{d_j}$ from which we must extract the estimates of the bottleneck function $\hat{b}_j : \mathbb{R}^{d_i} \mapsto \mathbb{R}^{d_{(i,j)}}$ and the effect function $\hat{f}_j : \mathbb{R}^{d_{(i,j)}} \mapsto \mathbb{R}^{d_j}$. In the linear case, fitting an estimator returns a weight matrix $\hat{\mathbf{M}} \in \mathbb{R}^{d_i \times d_j}$, which by the assumption of our data generating process has rank $d_{(i,j)} << d_i, d_j$. Extracting $\hat{b}_j$ and $\hat{f}_j$ amounts to finding a matrix factorization $\hat{\mathbf{M}} = \hat{\mathbf{B}}_j\hat{\mathbf{F}}_j$, where $\hat{\mathbf{B}}_j \in \mathbb{R}^{d_i \times d_{(i,j)}}$ and $\hat{\mathbf{F}}_j \in \mathbb{R}^{d_{(i,j)} \times d_j}$. We recover $\hat{\mathbf{B}}_j$ by selecting $d_{(i,j)}$ linearly independent columns from $\hat{\mathbf{M}}$

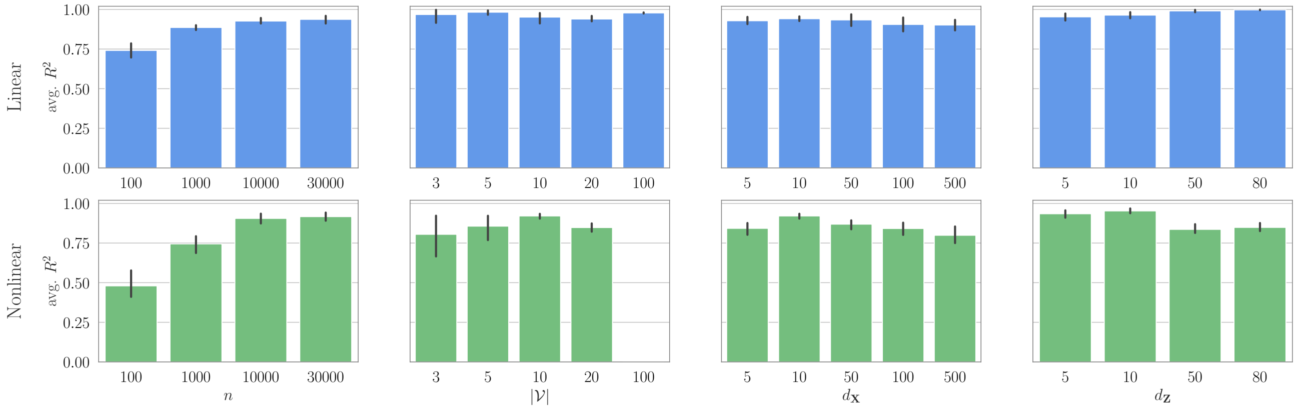

Figure 2: Results of the identifiability experiments across various settings. We report the mean R^2 along with its standard deviation shown as error bars. **Top:** The results for multiple varying parameter settings for linear SCBMs. **Bottom:** The results for the same parameter settings for nonlinear SCBMs. R^2 scores across models and settings of close to one indicate that we successfully learn the bottleneck variables to to a bijection. See section 8.1 for a detailed discussion of the results.

and $\hat{\mathbf{F}}_j$ by computing $\hat{\mathbf{F}}_j = \hat{\mathbf{B}}_j^+ \hat{\mathbf{M}}$, where $(\cdot)^+$ denotes the Moore-Penrose inverse [Moore, 1920, Bjerhammar, 1951, Penrose, 1955]. In the case of a nonlinear estimator we go about this factorization by means of the chosen network architecture. We employ an encoder-decoder structure where we train an encoder network $\hat{b}_\theta$, parametrized by weights θ, to map from the source node $\mathbf{X}_i$ to the bottleneck $\mathbf{Z}_{(i,j)}$ and a decoder network $\hat{f}_\phi$, parametrized by weights ϕ, to map from the bottleneck to the target node $\mathbf{X}_j$.

8 EXPERIMENTS

We present experiments for the estimation of SCBMs in practice, as well as their merit for transfer learning problems.

8.1 IDENTIFIABILITY

Setup. 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_j, as well as an effect function f_j. For linear models, we sample b_j by sampling a random matrix $\mathbf{B} \in [0,1]^{d_{\mathbf{X}} \times d_{\mathbf{Z}}}$ and f_j by sampling a random matrix $\mathbf{F} \in [0,1]^{d_{\mathbf{Z}} \times d_{\mathbf{X}}}$, both with rank $= d_{\mathbf{Z}}$. For nonlinear models, b_j and f_j are implemented by randomly initialized, 4-layer multilayer perceptrons (MLPs). Unless specified otherwise, by default we use the parameters $n = 30000$, $|\mathcal{V}| = 10$, $d_{\mathbf{X}} = 5$ and $d_{\mathbf{Z}} = 2$. We conduct experiments where we vary one of these parameters while keeping all others fixed.

Since our notion of identifiability amounts to estimating the ground truth bottlenecks up to a bijection (cf. Section 5), we fit an estimator between each ground truth $\mathbf{Z}_{(i,j)}$ and its estimate $\hat{\mathbf{Z}}_{(i,j)}$ in *both directions* and use the average of the R^2 of both fits as our metric for successful recovery of the ground truth bottleneck variable. The final score we report is the mean of this metric across all nodes. Fitting an estimator in both directions is required as we wish to test equivalence up to a bijection; a surjective map between $\mathbf{Z}_{(i,j)}$ and $\hat{\mathbf{Z}}_{(i,j)}$ would achieve a perfect score in one direction, but not in the other. For linear models, we fit a ordinary least squares estimator in each direction and for the nonlinear case we fit an MLP with six hidden layers and linear last layer.

Results. In linear SCBMs, we recover bottleneck variables with high accuracy across all settings. Performance improves quickly with sample size, saturating around $n = 10000$, and remains strong even as the number of nodes increases, suggesting minimal error propagation. As node dimension $d_{\mathbf{X}}$ increases, R^2 scores slightly drop due to greater compression demands. When varying bottleneck dimension $d_{\mathbf{Z}}$ (with $d_{\mathbf{X}} = 100$), performance stays high and converges to perfect as $d_{\mathbf{Z}}$ approaches $d_{\mathbf{X}}$, which is expected due to reduced compression.

Compared to the linear setting, the scores for estimating bottlenecks in nonlinear SCBMs are slightly lower, although not beyond what is to expected from a harder estimation problem. When increasing the number of nodes $|\mathcal{V}|$, performance tends to decrease, indicating larger error propagation effects than for linear models. However, performance stabilizes and error propagation does not lead to catastrophic

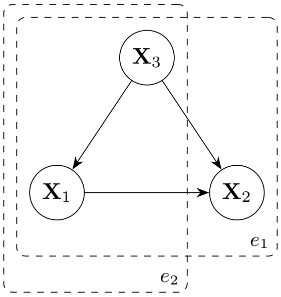

Figure 3: Graph of the SCBM used for the transfer learning experiments. We assume that samples from the environment e_1, where all variables are jointly observed, are relatively scarce compared to the number of samples of environment e_2, where we only jointly observe $\mathbf{X}_1$ and $\mathbf{X}_3$.

failure for larger models.[2] Larger internal node dimension $d_{\mathbf{X}}$ tends to diminish performance, but overall our method still successfully manages to estimate the bottlenecks. For the experiment where we vary the bottleneck dimension $d_{\mathbf{Z}}$, we again set the internal node dimension to $d_{\mathbf{X}} = 100$. There is a slight dip in performance as the bottleneck dimension approaches the internal node dimension. This could however also be an artifact of the metric in the nonlinear case, for which fitting a neural network is required, which for fixed network size we can expect to perform worse for large source and target spaces.

8.2 TRANSFER LEARNING

Setup. Consider a three variable SCBM with the associated graph $\mathcal{G}$ depicted in Figure 3. Our query of interest is the effect of $\mathbf{X}_1$ on $\mathbf{X}_2$, which is confounded by $\mathbf{X}_3$. Further, we assume that we have access to few samples of the joint distribution of $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$, while having orders of magnitude more samples of the joint distribution of $\{\mathbf{X}_1, \mathbf{X}_3\}$. This problem is sometimes referred to as the causal marginal problem [Gresele et al., 2022], where one has samples from different environments, where not all variables are included in all environments. This setting can be motivated from ecology, where $\mathbf{X}_1$ describes rainfall in some area, $\mathbf{X}_2$ the growth of vegetation in that area and $\mathbf{X}_3$ describes the cloud coverage. Measurements of $\mathbf{X}_1$ and $\mathbf{X}_3$ may come from a local measurement station that can collect samples at a high frequency, while measurements of all three variables jointly, i.e. including the vegetation growth, are collected from a satellite that crosses the area of interest at a much lower frequency. Estimating the effect of the edge $\mathbf{X}_1 \to \mathbf{X}_2$ requires conditioning on the confounder $\mathbf{X}_3$. Given the high dimensionality of the nodes, paired with the small sample size available of the joint distribution, this estimation problem is likely ill-conditioned.

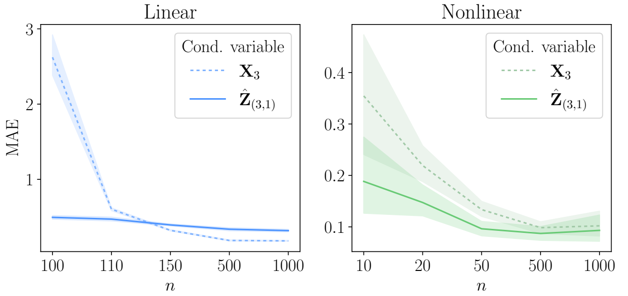

Figure 4: Mean absolute error (MAE) and 95% confidence interval of estimating the effect $\mathbf{X}_1 \to \mathbf{X}_2$ using different conditioning variables. For both linear and nonlinear SCBMs, using the bottleneck variable is beneficial for small samples sizes.

We study if we can leverage the abundant data from the environment that contains joint observations of $\{\mathbf{X}_1, \mathbf{X}_3\}$ to improve the effect estimation in the low sample regime. Specifically, we will use the samples of the joint distribution of $\{\mathbf{X}_1, \mathbf{X}_3\}$ to estimate the bottleneck variable $\hat{\mathbf{Z}}_{(3,1)}$, which can be used equivalently used for conditioning as $\mathbf{X}_3$, since it is a deterministic transformation of $\mathbf{X}_3$. Since we assume that bottleneck variables are lower dimensional than observed variables, we expect that using bottlenecks as conditioning variables is particularly beneficial in settings where the number of jointly observed samples of $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$ is low, but observations of $\{\mathbf{X}_1, \mathbf{X}_3\}$ are abundant. In the joint sample, the low dimensional bottlenecks lead to a larger effective sample size compared to using the observed variables for conditioning directly. For the linear case we consider a model with $d_{\mathbf{X}} = 50$ and $d_{\mathbf{Z}} = 2$, for the nonlinear case we set $d_{\mathbf{X}} = 500$. We use $n = 20000$ samples to first estimate the bottleneck and then study the performance of the resulting effect estimation for varying number of samples of the full joint distribution.

Results. For linear models—as shown in Figure 4—in the very low sample regime, using the bottleneck as conditioning variables provides a substantial benefit in terms of incurred error w.r.t. directly conditioning on the observed variable. This benefit subsides as the sample size increases, but nevertheless highlights the settings where bottleneck variables may be useful in practice. For nonlinear models, the same holds, however only becoming pronounced for larger internal node dimension $d_{\mathbf{X}}$.

[2] Experiments for $|\mathcal{V}| = 100$ exceeded our available compute, as graphs of this size typically have $\sim$ 3500 edges, for all of which a neural network must be trained to estimate the bottleneck variables. In practice, one would rarely estimate *all* bottlenecks of such a large graph, but only those required for a specific query.

9 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., 2023, Bing et al., 2024, Lachapelle et al., 2024, 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], while our notion of abstraction is a within-model operation that reduces a random vector to the essentials needed for downstream effect estimations.

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 pair of a cause and an outcome variable 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., 2024] 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. [2024] 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 generatiing 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. The exact relationship between both frameworks remains unexplored.

Dimension Reduction. Principle 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.

10 OUTLOOK

The work presented here is still in progress, and we plan to extend it in a number of ways. A natural first step is to investigate learning methods for discrete bottleneck spaces in greater detail (perhaps via clustering algorithms). We also intend to extend our experimental results to (more) realistic scenarios in the future. As our estimation method is comparatively simple w.r.t. most CRL approaches (that fail on real data [Gamella* et al., 2025]), we hope that it is more robust to misspecifications that inevitably arise in real-world settings. In our current experiments, we assume to know the dimension of the estimated bottleneck spaces, which is arguably unrealistic. We believe that this can be weakened straightforwardly, as the actual bottleneck dimension is just a lower bound on the dimension chosen during estimation. In practice, one could either make a conservative (over) estimate of the necessary dimension, or cross-validate on a small held out dataset, slowly decreasing the bottleneck dimension until performance begins to deteriorate. We will examine this in a future version.

Author Contributions

J.W. conceived the original idea and formalized the framework. S.B. developed the estimation method, created the code and ran the experiments. J.W. and S.B. jointly developed the theory and wrote the paper. J.R. provided supervision.

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 Education and Research (BMBF) as part of the project MAC-MERLin (Grant Agreement No. 01IW24007). 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.

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.

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.

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.

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. *arXiv preprint arXiv:2502.00407*, 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.

Juan L. Gamella*, Simon Bing*, and Jakob Runge. Sanity checking causal representation learning on a simple real-world system. *arXiv preprint arXiv:2502.20099*, 2025. *equal contribution.

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

Luigi Gresele, Julius Von Kügelgen, Jonas Kübler, Elke Kirschbaum, Bernhard Schölkopf, and Dominik Janzing. Causal inference through the structural causal marginal problem. In *Proceedings of the 39th International Conference on Machine Learning*, pages 7793–7824, 2022.

Shayan Hassanpour, Dirk Wübben, and Armin Dekorsy. Overview and investigation of algorithms for the information bottleneck method. In *SCC 2017; 11th International ITG Conference on Systems, Communications and Coding*, pages 1–6. VDE, 2017.

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.

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. *arXiv preprint arXiv:2401.04890*, 2024.

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.

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.

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.

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 and Jonas Peters. Decor: Deconfounding time series with robust regression. *arXiv preprint arXiv:2406.07005*, 2024.

Francisco NFQ Simoes, Mehdi Dastani, and Thijs van Ommen. Optimal causal representations and the causal information bottleneck. *arXiv preprint arXiv:2410.00535*, 2024.

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.

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 with Interventions. *arXiv preprint arXiv:2301.08230*, 2023.

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.