Unsupervised Causal Abstraction

Yuchen Zhu[*]
Centre for Artificial Intelligence
University College London
yuchen.zhu.18@ucl.ac.uk

Sergio Hernan Garrido Mejia[*]
Max Planck Institute for Intelligent Systems
Amazon
Tübingen, Germany
shgm@tuebingen.mpg.de

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Michel Besserve
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Abstract

Causal abstraction aims at mapping a complex causal model into a simpler ("reduced") one. Causal consistency constraints have been established to link the initial "low-level" model to its "high-level" counterpart, and identifiability results for such mapping can be established when we have access to some information about high-level variables. In contrast, we study the problem of learning a causal abstraction in an *unsupervised* manner, that is, when we do not have any information on the high-level causal model. In such setting, there typically exists multiple causally consistent abstractions, and we need to put additional constraints to unambiguously select a high-level model. To achieve this, we supplement a Kullback-Leibler-divergence-based consistency loss with a projection loss, which aims at finding the causal abstraction that best captures the variations of the low-level variables, thereby eliminating trivial solutions. The projection loss bears similarity to the Principal Component Analysis (PCA) algorithm; in this work it is shown to have a causal interpretation. We experimentally show how the abstraction preferred by the reconstruction loss varies with respect to the causal coefficients.

1 Introduction

In this paper we are interested in the problem of unsupervised causal reduction; that is, we are interested in finding aggregation functions from a low-level structural causal model (SCM) to a high-level SCM of lower dimensions, where the high-level SCM is unknown to us. This is important because a low-level SCM might be too intricate to analyse for policy makers (after all, policy is one of the end goals of causality), and a high-level SCM could offer such a simplification in a causal "consistent" way. Furthermore, if the aggregation function with which the low-level model is itself interpretable, then we can easily trace back the way in which the variables in the low-level SCM compose the high-level model.

We propose an *unsupervised* causal abstraction algorithm. The proposed algorithm can be applied to linear low-level SCM without restricting the SCM to entail a Directed Acyclic Graph (DAG), so that cyclic low-level SCM are allowed. The aggregation function is constrained to be linear and *constructive* to guarantee an interpretable reduction for policymakers. Beyond these inductive biases, we introduce a "Projection" or "Reconstruction" loss that prevents the algorithm from choosing trivial aggregation functions.

[*] Equal contribution, exact order decided by coin flip.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

The method can be applied to a class of Input-Output (IO or Leontief) models in economics, which use a matrix to represent the relations between industries in an economy. The matrix represents the produce (output) of each one of the industries in the model as a linear function of the produce of other industries (input). The industries in such a model typically represent fine-grained low-level sectors and thus policy makers find it useful to aggregate into larger sectors. However, they are usually aggregated by hand, using the intuition of an analyst, or by minimising some measure of bias Kymn and Norsworthy (1976). We believe that using a causal abstraction method could be an alternative to the current aggregation methods.

2 Related Work

Causal Abstraction. Since Rubenstein et al. (2017); Beckers and Halpern (2019), the causality community has been interested in the notion of *causal abstraction* or *representation*, aiming at finding a mapping that aggregates variables of a "microscopic" low-level causal model into "macroscopic" variables that form a high-level/abstracted causal model, such that interventions at both levels are consistent with each other. This notably led to exploring the notion of exact transformations (Rubenstein et al., 2017) and what happens when a non-exact transformation is used for aggregation.

The problem of learning such a causal abstraction with partial information has further been investigated. Zennaro et al. (2023) investigates, in a discrete setting, the case where both low-and high-level models are known, but the mapping is only partially known. Kekić et al. (2024) investigate the case where only one "target" node is known in the high level model, and interventional distributions can be sampled from the low-level model. In the opposite direction, Massidda et al. (2024) characterize the class of low-level model compatible with a predefined high-level model. To the best of our knowledge, the case of learning an abstraction of a low-level SCM with observed variables, without knowledge on the high-level variables, has not yet been addressed successfully. We call this setting unsupervised abstraction learning.

Zhu et al. (2024) study confoundedness in macro models and define low-level realisations of high-level interventions. Both concepts are essential in understanding what can we do with the high-level models, provided we have already defined what they call aggregation maps.

Causal Representation Learning. We can make a parallel between this aim and work in unsupervised causal representation learning (Schölkopf et al., 2021), Locatello et al. (2019). In this setting, observed variables are assumed to be generated from the mixing of latent variables forming an unknown latent SCM, using an unknown (typically deterministic) mapping. It is know that such causal representation cannot be recovered without any further inductive biases. This sparked a series of works von Kügelgen et al. (2024); Wendong et al. (2024) focusing on the conditions under which we can identify latent variables and a "mixing function" from observed variables alone without knowledge of causal structure. They conclude that under certain functional forms of the aggregation function and/or enough data from different environments where some causal mechanism is altered (i.e., interventions) (Squires et al., 2023; Buchholz et al., 2024), or different views (i.e., counterfactuals), identification holds. Crucially, works in causal representation learning assumes that there *is* a ground truth high-level model to be discovered, whereas we do not make this assumption.

3 Background

Definition 3.1 (Structural Causal Model (SCM) Peters et al. (2017))**.** A D-dimensional structural causal model is a triplet $M = (G, \mathcal{S}, P_{\boldsymbol{N}})$ consisting of: (i) a joint distribution $P_{\boldsymbol{N}}$ over exogenous variables $\mathcal{N}_{\mathcal{J}_{j\leq n}}$, (ii) a directed graph G with D vertices, (iii) a set $\mathcal{S} = \{X_d := \mathrm{f}_d(\boldsymbol{Pa}_d, N_d),\ d = 1, \cdots, D\}$ of structural equations, where $\boldsymbol{Pa}_d$ are the parents of variable X_d in G, such that the system $\{x_d := \mathrm{f}_d(\boldsymbol{pa}_d, n_d)\}$ has a unique solution in $\mathbf{x}$, $P_{\boldsymbol{N}}$-almost everywhere.

In this work, we focus on the class of **linear structural causal models**: a linear structural causal model is an SCM (Def. 3.1) where there is some $\boldsymbol{A} \in \mathbb{R}^{D\times D}$ such that $\mathcal{S}$ contains equations in the linear system $\boldsymbol{X} := \boldsymbol{A}\boldsymbol{X} + \boldsymbol{N}$. The existence of solution is equivalent to requiring that $(\boldsymbol{I}_D - \boldsymbol{A})$ is invertible. In this case, for every $\boldsymbol{n}$, the solution of M is $\boldsymbol{x} = (\boldsymbol{I}_{(D)} - \boldsymbol{A})^{-1}\boldsymbol{n}$, and the joint distribution over $\boldsymbol{X}$, $P_{\boldsymbol{X}}$ is well-defined as a pushforward through $(\boldsymbol{I}_{(D)} - \boldsymbol{A})^{-1}$ of $P_{\boldsymbol{N}}$.

Moreover, we focus on a class of interventions named *shift interventions*.

Definition 3.2 (Shift interventions and entailed distribution)**.** Given a D-dimensional linear structural causal model M, a shift intervention is represented by a vector $\boldsymbol{i} \in \mathbb{R}^D$, and it transforms M into $M^{(\boldsymbol{i})} := (G, \mathcal{S}^{(\boldsymbol{i})}, P_{\boldsymbol{N}})$, where $\mathcal{S}^{(\boldsymbol{i})}$ now contains the equations given by the linear system $\boldsymbol{X} := \boldsymbol{A}\boldsymbol{X} + \boldsymbol{N} + \boldsymbol{i}$. We denote the joint distribution over $\boldsymbol{X}$ under a shift intervention $\boldsymbol{i}$ by $P_{\boldsymbol{X}}^{(\boldsymbol{i})}$.

4 Unsupervised Causal Abstraction

We learn a mapping from a low-level SCM model to a high-level SCM with aggregated variables. From now on, we use M to denote our D-dimensional low-level linear SCM and $\bar{M}$ the $\bar{D}$-dimensional high-level linear SCM which we learn a mapping to. Let $\boldsymbol{X}$ denote the endogenous variables of M and $\boldsymbol{Z}$ of $\bar{M}$.

Our first learning objective is to encourage the exact transformation between M and $\bar{M}$ based on shift interventions (Rubenstein et al., 2017). This requires that there is a mapping between *states* $\boldsymbol{\tau} : \mathbb{R}^D \to \mathbb{R}^{\bar{D}}$ and a mapping between *shift interventions* $\boldsymbol{\omega} : \mathbb{R}^D \to \mathbb{R}^{\bar{D}}$, such that for any shift intervention $\boldsymbol{i} \in \mathbb{R}^D$,

$$\widehat{P_{\boldsymbol{Z},\boldsymbol{\tau}}}^{(\boldsymbol{i})} = P_{\boldsymbol{Z}}^{(\boldsymbol{\omega}(\mathbf{i}))} \tag{1}$$

where $\widehat{P_{\boldsymbol{Z},\boldsymbol{\tau}}}^{(\boldsymbol{i})} = \boldsymbol{\tau}_{\#}\left(P_{\boldsymbol{X}}^{(\boldsymbol{i})}\right)$, the pushforward via $\boldsymbol{\tau}$ of the $\boldsymbol{i}$-intervened distribution over $\boldsymbol{X}$.

$\boldsymbol{\tau}$ and $\boldsymbol{\omega}$ are standard constructions in the causal abstraction literature (Rubenstein et al., 2017; Beckers and Halpern, 2019; Kekić et al., 2024), $\boldsymbol{\tau}$ is sometimes also known as the aggregation map (Zhu et al., 2024). Additionally, we place the restriction that $\boldsymbol{\tau}$ and $\boldsymbol{\omega}$ form a *constructive* abstraction (Beckers and Halpern, 2019). That is, there is a so-called alignment function π from $\{1, \cdots, \bar{D}\}$ to the non-overlapping subsets of $\{1, \cdots, D\}$, such that $\boldsymbol{\tau}$ and $\boldsymbol{\omega}$ decompose as

$$\boldsymbol{\tau} = (\tau_1, \cdots, \tau_{\bar{D}}) \text{ with } \tau_{\bar{\mathrm{d}}} : \boldsymbol{x} \mapsto \bar{\tau}_{\bar{\mathrm{d}}}\left(\boldsymbol{x}_{\pi(\bar{d})}\right) \tag{2}$$

$$\boldsymbol{\omega} = (\omega_1, \cdots, \omega_{\bar{D}}) \text{ with } \omega_{\bar{\mathrm{d}}} : \boldsymbol{i} \mapsto \bar{\omega}_{\bar{\mathrm{d}}}\left(\boldsymbol{i}_{\pi(\bar{d})}\right) \tag{3}$$

Our set-up closely follows Kekić et al. (2024), but they require a known target variable Y, whereas we do not require this.

Consistency loss. Similar to Kekić et al. (2024), we encourage exact transformation using a KL-divergence based loss:

$$\mathrm{L}_{\mathrm{kl}}(\boldsymbol{\tau}, \boldsymbol{\omega}, \bar{M}) = \mathbb{E}_{\boldsymbol{i} \sim P(\boldsymbol{i})}\left[\mathrm{KL}\left(\widehat{P_{\boldsymbol{Z},\boldsymbol{\tau}}}^{(\boldsymbol{i})} \| P_{\boldsymbol{Z}}^{(\boldsymbol{\omega}(\boldsymbol{i}))}\right)\right] \tag{4}$$

noting that $P_{\boldsymbol{Z}}^{(\boldsymbol{\omega}(\boldsymbol{i}))}$ is a function of $\bar{M}$.

Linearity of $\boldsymbol{\tau}$ and $\boldsymbol{\omega}$. We assume both $\boldsymbol{\tau}$ and $\boldsymbol{\omega}$ are linear maps for interpretability. Thus, onwards we abuse notation and use $\boldsymbol{\tau}$ and $\boldsymbol{\omega}$ as matrices in $\mathbb{R}^{\bar{D} \times D}$, so the mappings can be viewed as a matrix multiplied by a vector:

$$\boldsymbol{\tau}(\boldsymbol{x}) = \boldsymbol{\tau}\boldsymbol{x} \tag{5}$$

$$\boldsymbol{\omega}(\boldsymbol{i}) = \boldsymbol{\omega}\boldsymbol{i} \tag{6}$$

Moreover, the constructive abstraction constraint is equivalent to requiring that every column of $\boldsymbol{\tau}$ and $\boldsymbol{\omega}$ has at most one non-zero element.

Projection loss. A trivial solution which would achieve $\mathrm{L}_{\mathrm{kl}} = 0$ is to set $\boldsymbol{\tau} = \boldsymbol{\omega} = \mathbf{0}$. To discourage this meaningless solution, we introduce a reconstruction loss to encourage the high-level SCM to retain the maximum amount of information from the low-level data. Specifically, we parameterise a projection operator based on the abstraction matrix $\boldsymbol{\tau}$, and require that over all populations of interventional distributions of $\boldsymbol{X}$, $\boldsymbol{\tau}$ minimises the *projection* loss:

$$\mathrm{L}_{\mathrm{proj}}(\boldsymbol{\tau}) = \mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}},\, \boldsymbol{X}\sim P_{\boldsymbol{X}}^{(\boldsymbol{i})}}\left[\|\boldsymbol{X} - \mathbf{P}_{\boldsymbol{\tau}}(\boldsymbol{X})\|_2^2\right] \tag{7}$$

Various choices of $\mathbf{P}_{\boldsymbol{\tau}}$ are possible. One option is to construct an orthogonal projection matrix with τ, and another is to construct a projection using conditional expectation. In this work, we focus on the former, which we detail in the following.

Orthogonal projection recovers PCA. We will denote the orthogonal projection as $\mathbf{P}_{\boldsymbol{\tau}}^{(\mathrm{orth})}$, construct it as

$$\mathbf{P}_{\boldsymbol{\tau}}^{(\mathrm{orth})} = \boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-1}\boldsymbol{\tau} \tag{8}$$

It can be easily verified that $\mathbf{P}_{\boldsymbol{\tau}}^{(\mathrm{orth})}$ is a projection and is symmetric, and therefore is an orthogonal projection.

The property that $\mathbf{P}_{\boldsymbol{\tau}}^{(\mathrm{orth})}$ is an orthogonal projection helps us connect it to the well-known algorithm *principal component analysis (PCA).*

Proposition 4.1. *Assuming $\boldsymbol{\tau}$ is full-rank, the following equality is true:*

$$\arg\min_{\boldsymbol{\tau}\in\mathbb{R}^{\bar{D}\times D}} \mathbb{E}_{P_{\boldsymbol{X}}^{(0)}}\left[\|\boldsymbol{X} - \mathbf{P}_{\boldsymbol{\tau}}^{(\mathrm{orth})}\boldsymbol{X}\|_2^2\right] = \arg\max_{\boldsymbol{\tau}\in\mathbb{R}^{\bar{D}\times D}} \mathbb{E}_{P_{\boldsymbol{X}}^{(0)}}\left[\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X}\|_2^2\right] \tag{9}$$

Remark 4.2. If we assume $P_{\boldsymbol{X}}^{(0)}$ has mean 0, then the right hand side of (4.1) achieve the same maximum as that achieved by PCA over $P_{\boldsymbol{X}}^{(0)}$, so we recover the connection between our reconstruction loss and PCA:

$$\begin{aligned}\max_{\boldsymbol{\tau}\in\mathbb{R}^{\bar{D}\times D}} \mathbb{E}_{P_{\boldsymbol{X}}^{(0)}}\left[\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X}\|_2^2\right] &= \max_{\substack{\boldsymbol{\tau}\in\mathbb{R}^{\bar{D}\times D},\\ \|\boldsymbol{\tau}_{\bar{d}:}\|_2=1 \text{ for } \bar{d}=1,\cdots,\bar{D},\\ \boldsymbol{\tau}_{\bar{d}:}\text{orthogonal}}} \mathbb{E}_{P_{\boldsymbol{X}}^{(0)}}\left[\|\boldsymbol{\tau}\boldsymbol{X}\|_2^2\right] & (10)\\ &= \max_{\substack{\boldsymbol{\tau}\in\mathbb{R}^{\bar{D}\times D},\\ \|\boldsymbol{\tau}_{\bar{d}:}\|_2=1 \text{ for } \bar{d}=1,\cdots,\bar{D},\\ \boldsymbol{\tau}_{\bar{d}:}\text{orthogonal}}} \sum_{\bar{d}=1}^{\bar{D}} \boldsymbol{\tau}_{\bar{d},:}\operatorname{cov}_{P_{\boldsymbol{X}}^{(0)}}[\boldsymbol{X}]\boldsymbol{\tau}_{\bar{d},:}^\top & (11)\end{aligned}$$

We next show that, for a general aggregation matrix $\boldsymbol{\tau}\in\mathbb{R}^{\bar{D}\times D}$, L_{proj} with the orthogonal projection $\mathbf{P}_{\boldsymbol{\tau}}^{(\mathrm{orth})}$ is the same as the sum of the projections into $\bar{D}$ orthogonal directions of the sum of second order statistics of the unintervened $\boldsymbol{X}$ and the shift interventions $\boldsymbol{i}$.

Theorem 4.3. *For any $\boldsymbol{\tau}\in\mathbb{R}^{\bar{D}\times D}$ of full rank, define the orthogonal projection loss as*

$$\mathrm{L}_{\text{proj}}^{(\text{orth})}(\boldsymbol{\tau}) := \mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}},\, \boldsymbol{X}\sim P_{\boldsymbol{X}}^{(\boldsymbol{i})}}\left[\|\boldsymbol{X} - \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}(\boldsymbol{X})\|_2^2\right]. \tag{12}$$

Moreover, let $\boldsymbol{O} := \left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-1}\boldsymbol{\tau}$. Then

$$\mathrm{L}_{\text{proj}}^{(\text{orth})}(\boldsymbol{\tau}) = -\sum_{\bar{d}=1}^{\bar{D}}\left(\boldsymbol{O}_{\bar{d}:}\underbrace{\left(\operatorname{cov}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}[\boldsymbol{X}] + \left(\boldsymbol{I}_{(D)} - \boldsymbol{A}\right)^{-1}\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}\left[\boldsymbol{i}\boldsymbol{i}^\top\right]\left(\boldsymbol{I}_{(D)} - \boldsymbol{A}\right)^{-1,\top}\right)}_{\boldsymbol{\Psi}}\boldsymbol{O}_{\bar{d}:}^\top\right) + \text{Constant} \tag{13}$$

Remark 4.4. Ψ in the summand in (13) is the sum of the unintervened covariance of $\boldsymbol{X}$ and the second order statistic of impact of the shift intervention $\boldsymbol{i}$ on $\boldsymbol{X}$; therefore it is certainly positive (semi-)definite. Moreover, $\boldsymbol{O}$ has orthonormal rows since $\boldsymbol{O}\boldsymbol{O}^\top = \left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}} = \boldsymbol{I}_{\bar{D}}$. The summand in (13) thus computes the projection of the *sum* second order statistic *for unintervened variables and intervention effects* Ψ in the $\bar{d}$-th orthogonal direction.

Remark 4.5. When $\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}[\boldsymbol{i}] = 0$, $\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}\left[\boldsymbol{i}\boldsymbol{i}^\top\right] = \operatorname{cov}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}[\boldsymbol{i}]$. In this case, Ψ recovers the covariance of the marginal over $\boldsymbol{X}$ in the linear model

$$\boldsymbol{i}\sim P_{\boldsymbol{i}},\ \boldsymbol{N}\sim P_{\boldsymbol{N}},\ \boldsymbol{X} = \left(I_{(\bar{D})} - \boldsymbol{A}\right)^{-1}(\boldsymbol{N}+\boldsymbol{i}),\ \boldsymbol{i}\perp\boldsymbol{N}.$$

Sum over clusters of PCA projected variances. So far we have not exploited the construction property of the abstraction matrix $\boldsymbol{\tau}$. A desirable property follows from constructive abstraction - that the rows of $\boldsymbol{\tau}$ are orthogonal. Furthermore, projection along the direction of a row of $\boldsymbol{\tau}$ only attends to a subset of the variables in $\boldsymbol{X}$. This leads us to the following decomposition of $\mathrm{L}_{\text{proj}}^{(\text{orth})}$ for constructive $\boldsymbol{\tau}$.

Corollary 4.6. *Let $\boldsymbol{\tau} \in \mathbb{R}^{\bar{D} \times \bar{d}}$ be a full-rank constructive abstraction matrix, then the following is true.*

$$\begin{aligned}
&\mathrm{L}_{\text{proj}}^{(\text{orth})}(\boldsymbol{\tau}) \\
&= \textstyle\sum_{\bar{d}=1}^{\bar{D}} \frac{-1}{\left\|\tau_{\bar{d},\pi(\bar{d})}\right\|_2^2} \left(\boldsymbol{\tau}_{\bar{d},\pi(\bar{d})} \underbrace{\left(\mathrm{cov}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\boldsymbol{X}_{\pi(\bar{d})}\right] + \left(\boldsymbol{I}_{(D)} - \boldsymbol{A}\right)^{-1}_{\pi(\bar{d}),:} \mathbb{E}_{\boldsymbol{i} \sim P_{\boldsymbol{i}}}\left[\boldsymbol{i}\boldsymbol{i}^\top\right] \left(\boldsymbol{I}_{(D)} - \boldsymbol{A}\right)^{-1,\top}_{:,\pi(\bar{d})} \right)}_{\boldsymbol{\Psi}_{\text{cons}}} \boldsymbol{\tau}_{\bar{d},\pi(\bar{d})}^\top \right) \\
&\qquad + \text{ Constant}
\end{aligned} \tag{14}$$

Proof. Immediate by substituting in a constructive $\boldsymbol{\tau}$ and observing that $\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)$ becomes diagonal with $\left\|\boldsymbol{\tau}_{\bar{d},:}\right\|_2^2$ on the $\bar{d}$-th diagonal element. □

We focus on strongly enforcing consistency conditions, that is, we require that for the optimum $\boldsymbol{\tau}^*, \boldsymbol{\omega}^*, \bar{M}^*$, $\mathrm{L}_{\text{kl}}\left(\boldsymbol{\tau}^*, \boldsymbol{\omega}^*, \bar{M}^*\right) = 0$. We thus optimise the following constrained loss using the definition of $\mathrm{L}_{\text{proj}}^{(\text{orth})}(\boldsymbol{\tau})$ in Theorem 4.3 and L_{kl}:

$$\boldsymbol{\tau}^*, \boldsymbol{\omega}^*, \bar{M}^* = \arg \min_{\substack{\boldsymbol{\tau}, \boldsymbol{\omega} \in \mathbb{R}^{\bar{D} \times D},\ \bar{M} \\ \boldsymbol{\tau}\text{constructive}}} \mathrm{L}_{\text{proj}}^{(\text{orth})}(\boldsymbol{\tau}) \qquad \text{subject to} \qquad \mathrm{L}_{\text{kl}}\left(\boldsymbol{\tau}, \boldsymbol{\omega}, \bar{M}\right) = 0 \tag{15}$$

5 Experiments

Data generation process and optimisation. We consider the linear chain as in Kekić et al. (2024),

$$X_1 \longrightarrow X_2 \longrightarrow X_3 \longrightarrow X_4$$

with causal mechanism

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{N} \qquad \boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ p_{21} & 0 & 0 & 0 \\ 0 & p_{32} & 0 & 0 \\ 0 & 0 & p_{43} & 0 \end{pmatrix} \qquad \boldsymbol{N} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_{(4)}\right) \tag{16}$$

We will explore variations of the parameters p_{ij} in two experiments we will describe below. In both experiments, we sample 10k observations from the data generation process. We consider all bi-partitions of this chain such that each subset is formed of contiguous variables in the topological ordering. That is, we consider the partitions $\{\{X_1\}, \{X_2, X_3, X_4\}\}$, $\{\{X_1, X_2\}, \{X_3, X_4\}\}$, $\{\{X_1, X_2, X_3\}, \{X_4\}\}$. Kekić et al. (2024) provides a relation in $\boldsymbol{\tau}_{1,:}$ and $\boldsymbol{\tau}_{2,:}$ such that $\mathrm{L}_{\text{kl}} = 0$,

$$\bar{\tau}_{1,:} \propto \left(0, \cdots, 0, \bar{\tau}_{2,1}\right). \tag{17}$$

Combined with our objective $\mathrm{L}_{\text{proj}}^{(\text{orth})}$ which normalises each $\boldsymbol{\tau}_{\bar{d},:}$ the result of Kekić et al. (2024) is equivalent to setting:

$$\bar{\tau}_{1,:} = \left(0, \cdots, 0, 1\right). \tag{18}$$

We then choose $\boldsymbol{\tau}_{2,:}$ based on minimising $\mathrm{L}_{\text{proj}}^{(\text{orth})}$, and finally choose the partition with the highest summed projected covariance for our reduction.

First experiment (one parameter equals 0). In this experiment we consider 3 different scenarios. In each scenario we will set one of the parameters p_{ij} to 0 and the other two to 1. We then compute the projected covariance for each one of the partitions as explained above. We repeat this process 10 times for each one of the scenarios. The results are shown in Figure 1

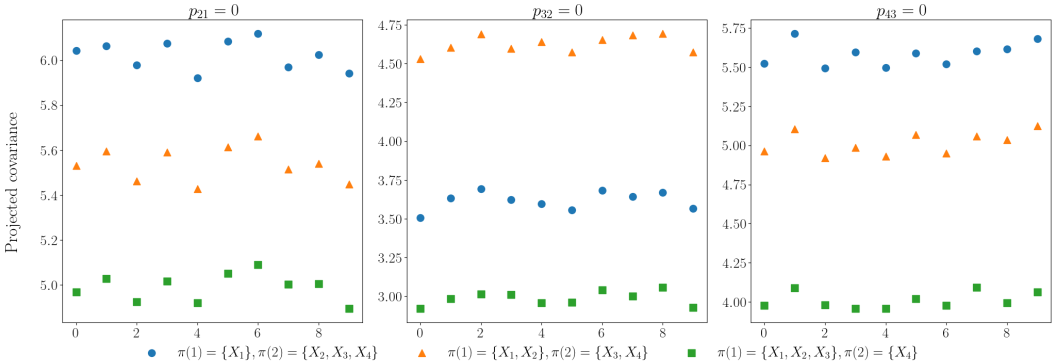

Figure 1: Projected covariance for the three variations of the first experiment. In each figure, we set one of the parameters of the linear chain to be 0 and the other two to be 1. The x-axis represents the experiment number.

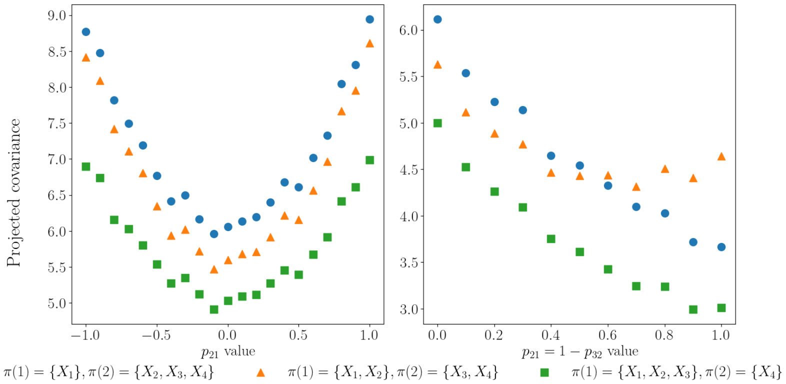

Figure 2: Projected covariance for the second experiment. The x-axis represents the value of p_{21} (left) and $p_{21} = 1 - p_{32}$ (right)

Second experiment (parameter values vary). In the second experiment we see how the projected covariance varies when some of the parameter values vary. First, we fix p_{32} and p_{43} to be 1 and vary p_{21} from -1 to 1 in increments of 0.1 totalling 21 different SCM. Second, we fix p_{43} to be 1 and vary p_{21} from 0 to 1 in increments of 0.1 , and let $p_{32} = 1 - p_{21}$, totalling 11 different SCM. As in the first experiment, we compute the projected covariance for each of the partitions and see how the chosen partition changes with changes in the value of the parameter. The results of this experiment are shown in Figure 2. In Appendix B we include a figure where p_{32} and p_{43} are varied independently as in p_{21} here.

Results: first experiment. When we set $p_{21} = 0$ and $p_{32} = 0$, we see that the chosen partition (that is, the one with the highest projected covariance) is that where p_{ij} is set to zero. This follows our intuition that causally related variables are grouped together by the variance maximisation algorithm. When p_{43} is set to 0, the result is more interesting: the variance maximisation algorithms chooses the partition between X_1 and X_2. We interpret this as follows: consider the causal model obtained from marginalising out X_4; this is a chain $X_1 \to X_2 \to X_3$, with constant causal coefficients equal to 1. Running our algorithm on this model will result in a partition between X_1 and X_2, because variance accumulates down the chain. Now add in another low-level variable independent from the existing variables, with the inductive bias that it is a child of X_3, but with coefficient 0. Suppose the variance of the added variable is 0, this is identical to the three-variable model - in the sense that the set of admissible states in this model can be identified with that of the three-variable model - so the partition should not change. As we increase the variance of the added variable, we would reach a critical point beyond which the variable varies enough that we should have a separate high-level variable to represent it, so the partition should be between between the added variable and X_3. This

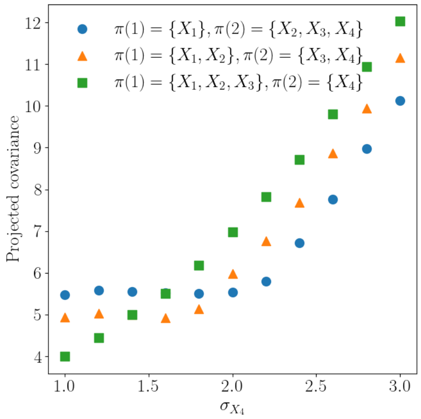

Figure 3: Projected covariance for the alternative second experiment. The x-axis represents the value of the standard deviation of X_4.

is also confirmed by our algorithm: as $\mathrm{Var}(X_4) \to \infty$, the partition will at some point jump to be between X_3 and X_4. The reason why the partition is currently between X_1 and X_2 is because the variance of X_4 has not reached the critical value. We showcase this phenomenon on Figure 3.

Results: second experiment. We observe that for all values of p_{21} the first partition has the highest projected covariance. This is expected, as given that the other two parameters are 1, the variance of the other variables is amplified by their noise variables, so that the total variance of X_2, X_3 and X_3 are larger than that of X_1. This behaviour, however, is reversed whenever we also vary p_{32}. Again, this is expected because as p_{32} gets smaller, the causal influence of X_2 over X_3 diminishes whereas there is a causal connection between X_1 and X_2 so that it is sensible to aggregate those variables into a single high-level variable.

6 Discussion

We present an algorithm do causal reduction of low-level linear SCM. We force the aggregation map to be linear and constructive. In addition to these inductive biases we include a "projection" loss that prevents our algorithm to find trivial solutions. A particular case of the projection loss has similarities with the classic PCA algorithm. In future work we plan to apply the algorithm to real-world data, in particular, we are interested in economic Input-Output models.

References

Beckers, S. and Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 2678–2685.

Buchholz, S., Rajendran, G., Rosenfeld, E., Aragam, B., Schölkopf, B., and Ravikumar, P. (2024). Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36.

Kekić, A., Schölkopf, B., and Besserve, M. (2024). Targeted reduction of causal models. In *The 40th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Kymn, K. O. and Norsworthy, J. (1976). A review of industry aggregation in input-output models. *The American Economist*, 20(1):5–10.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR.

Massidda, R., Magliacane, S., and Bacciu, D. (2024). Learning causal abstractions of linear structural causal models. In *The 40th Conference on Uncertainty in Artificial Intelligence*.

Peters, J., Janzing, D., and Schlkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Rubenstein, P., Weichwald, S., Bongers, S., Mooij, J., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. (2017). Causal consistency of structural equation models. In *Proceedings of the 17th 33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Squires, C., Seigal, A., Bhate, S. S., and Uhler, C. (2023). Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, pages 32540–32560. PMLR.

von Kügelgen, J., Besserve, M., Wendong, L., Gresele, L., Kekić, A., Bareinboim, E., Blei, D., and Schölkopf, B. (2024). Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36.

Wendong, L., Kekić, A., von Kügelgen, J., Buchholz, S., Besserve, M., Gresele, L., and Schölkopf, B. (2024). Causal component analysis. *Advances in Neural Information Processing Systems*, 36.

Zennaro, F. M., Drávucz, M., Apachitei, G., Widanage, W. D., and Damoulas, T. (2023). Jointly learning consistent causal abstractions over multiple interventional distributions. In *Conference on Causal Learning and Reasoning*, pages 88–121. PMLR.

Zhu, Y., Budhathoki, K., Kübler, J. M., and Janzing, D. (2024). Meaningful causal aggregation and paradoxical confounding. In *Causal Learning and Reasoning*, pages 1192–1217. PMLR.

A Proofs

Proof of Proposition 4.1. Since $\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}$ is an orthogonal projection, $\text{Im}\left(\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\right) \perp \text{Ker}\left(\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\right)$. But $\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\left(\boldsymbol{X} - \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\right) = \left(\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})} - \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})^2}\right)\boldsymbol{X} = \mathbf{0}$, so $\left(\boldsymbol{X} - \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\right) \perp \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}$. Therefore, by Pythagoras' theorem,

$$\|\boldsymbol{X} - \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2 = \|\boldsymbol{X}\|_2^2 - \|\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2 \tag{19}$$

But $\boldsymbol{X}$ is constant in $\boldsymbol{\tau}$, so

$$\arg\min_{\boldsymbol{\tau}\in\mathbb{R}^{\bar{D}\times D}} \mathbb{E}_{P_{\boldsymbol{X}}^{(0)}}\left[\|\boldsymbol{X} - \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2\right] \tag{20}$$

$$= \arg\min_{\boldsymbol{\tau}\in\mathbb{R}^{\bar{D}\times D}} \mathbb{E}_{P_{\boldsymbol{X}}^{(0)}}\left[\|\boldsymbol{X}\|_2^2 - \|\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2\right] \tag{21}$$

$$= \arg\max_{\boldsymbol{\tau}\in\mathbb{R}^{\bar{D}\times D}} \mathbb{E}_{P_{\boldsymbol{X}}^{(0)}}\left[\|\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2\right] \tag{22}$$

But assuming $\boldsymbol{\tau}$ is full rank,

$$\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})} = \boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-1}\boldsymbol{\tau} \tag{23}$$

$$= \boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau} \tag{24}$$

But note that the rows of $\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}$ are orthogonal, since $\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}} = \boldsymbol{I}_{\bar{D}}$. Therefore

$$\|\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2 \tag{25}$$

$$= \|\boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X}\|_2^2 \tag{26}$$

$$= \|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X}\|_2^2 \tag{27}$$

Therefore, the hypothesis follows. □

Proof of Proposition 4.3. Since $\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}$ is an orthogonal projection, $\text{Im}\left(\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\right) \perp \text{Ker}\left(\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\right)$. But $\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\left(\boldsymbol{X} - \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\right) = \left(\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})} - \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})^2}\right)\boldsymbol{X} = \mathbf{0}$, so $\left(\boldsymbol{X} - \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\right) \perp \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}$. Therefore, by Pythagoras' theorem,

$$\|\boldsymbol{X} - \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2 = \|\boldsymbol{X}\|_2^2 - \|\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2 \tag{28}$$

But $\boldsymbol{X}$ is constant in $\boldsymbol{\tau}$, so

$$\begin{aligned}
&\mathrm{L}_{\text{proj}}^{(\text{orth})}(\boldsymbol{\tau}) && (29)\\
=&\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}, P_{\boldsymbol{X}}^{(\boldsymbol{i})}}\left[\|\boldsymbol{X} - \mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2\right] && (30)\\
=&\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}, P_{\boldsymbol{X}}^{(\boldsymbol{i})}}\left[\|\boldsymbol{X}\|_2^2 - \|\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2\right] && (31)\\
=& \text{ Constant wrt } \boldsymbol{\tau} - \mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}, P_{\boldsymbol{X}}^{(\boldsymbol{i})}}\left[\|\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2\right] && (32)
\end{aligned}$$

But assuming $\boldsymbol{\tau}$ is full rank,

$$\begin{aligned}
\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})} &= \boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-1}\boldsymbol{\tau} && (33)\\
&= \boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau} && (34)
\end{aligned}$$

But note that the rows of $\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}$ are orthogonal, since $\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}} = \boldsymbol{I}_{\bar{D}}$. Therefore

$$\begin{aligned}
&\|\mathbf{P}_{\boldsymbol{\tau}}^{(\text{orth})}\boldsymbol{X}\|_2^2 && (35)\\
=&\|\boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X}\|_2^2 && (36)\\
=&\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X}\|_2^2 && (37)
\end{aligned}$$

Taking expectation over $P_{\boldsymbol{i}}$. Consider:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}\left[\mathbb{E}_{P_{\boldsymbol{X}}^{(\boldsymbol{i})}}\left[\left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X}\right\|_2^2\right]\right] && (38)\\
=&\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}\left[\mathbb{E}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\left(\boldsymbol{X} + \left(\boldsymbol{I}_{(D)} - \boldsymbol{A}\right)^{-1}\boldsymbol{i}\right)\right\|_2^2\right]\right] && (39)
\end{aligned}$$

Let $\boldsymbol{S} := \left(\boldsymbol{I}_{(D)} - \boldsymbol{A}\right)^{-1}$ denote the solution map for the micro model. The above can be continued as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}\left[\mathbb{E}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\left(\boldsymbol{X} + \left(\boldsymbol{I}_{(D)} - \boldsymbol{A}\right)^{-1}\boldsymbol{i}\right)\right\|_2^2\right]\right] && (40)\\
=&\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}\left[\mathbb{E}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X} + \left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{S}\boldsymbol{i}\right\|_2^2\right]\right] && (41)\\
=&\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}\left[\mathbb{E}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X}\right\|_2^2 + \left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{S}\boldsymbol{i}\right\|_2^2 + 2\boldsymbol{X}^\top\boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{S}\boldsymbol{i}\right]\right] && (42)\\
=&\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}\left[\mathbb{E}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X}\right\|_2^2 + \left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{S}\boldsymbol{i}\right\|_2^2 + 2\boldsymbol{X}^\top\boldsymbol{\tau}^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-1}\boldsymbol{\tau}\boldsymbol{S}\boldsymbol{i}\right]\right] && (43)\\
=&\mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}\left[\mathbb{E}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X}\right\|_2^2 + \left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{S}\boldsymbol{i}\right\|_2^2\right]\right] && (44)
\end{aligned}$$

since X has mean $\mathbf{0}$, and continuing the above:

$$=\mathbb{E}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{X}\right\|_2^2\right] + \mathbb{E}_{\boldsymbol{i}\sim P_{\boldsymbol{i}}}\left[\left\|\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\boldsymbol{\tau}\boldsymbol{S}\boldsymbol{i}\right\|_2^2\right] \tag{45}$$

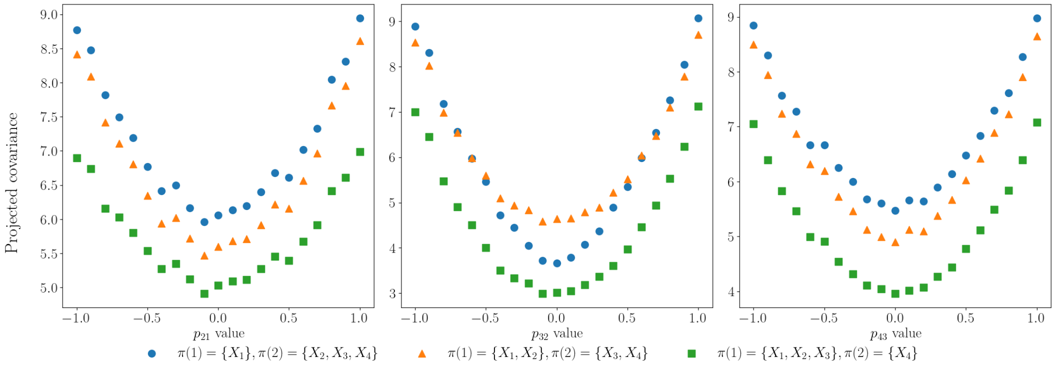

Figure 4: Projected covariance for the second experiment applied to all parameters The x-axis represents the value of p_{21} (left), p_{32} (center) and p_{43} (right).

But now using the trace properties, we can manipulate the above to be in terms of second order statistics:

$$= \mathbb{E}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\operatorname{Tr}\left(\boldsymbol{X}^\top \boldsymbol{\tau}^\top \left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-1} \boldsymbol{\tau}\boldsymbol{X}\right)\right] + \mathbb{E}_{\boldsymbol{i}\sim P_i}\left[\operatorname{Tr}\left(\boldsymbol{i}^\top \boldsymbol{S}^\top \boldsymbol{\tau}^\top \left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-1} \boldsymbol{\tau}\boldsymbol{S}\boldsymbol{i}\right)\right] \tag{46}$$

$$= \mathbb{E}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\operatorname{Tr}\left(\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}} \boldsymbol{\tau}\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\tau}^\top \left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\right)\right] + \mathbb{E}_{\boldsymbol{i}\sim P_i}\left[\operatorname{Tr}\left(\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}} \boldsymbol{\tau}\boldsymbol{S}\boldsymbol{i}\boldsymbol{i}^\top \boldsymbol{S}^\top \boldsymbol{\tau}^\top \left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\right)\right] \tag{47}$$

Now, by linearity of expectation:

$$= \operatorname{Tr}\left(\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}} \boldsymbol{\tau}\, \mathbb{E}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\boldsymbol{X}\boldsymbol{X}^\top\right] \boldsymbol{\tau}^\top \left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\right) + \operatorname{Tr}\left(\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}} \boldsymbol{\tau}\boldsymbol{S}\, \mathbb{E}_{\boldsymbol{i}\sim P_i}\left[\boldsymbol{i}\boldsymbol{i}^\top\right] \boldsymbol{S}^\top \boldsymbol{\tau}^\top \left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\right) \tag{48}$$

By linearity of trace:

$$= \operatorname{Tr}\left(\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}} \boldsymbol{\tau}\left(\mathbb{E}_{P_{\boldsymbol{X}}^{(\mathbf{0})}}\left[\boldsymbol{X}\boldsymbol{X}^\top\right] + \boldsymbol{S}\, \mathbb{E}_{\boldsymbol{i}\sim P_i}\left[\boldsymbol{i}\boldsymbol{i}^\top\right] \boldsymbol{S}^\top\right) \boldsymbol{\tau}^\top \left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)^{-\frac{1}{2}}\right) \tag{49}$$

Substituting back into (32), we obtain the hypothesis. □

B Extra results