# Learning Causal Response Representations through Direct Effect Analysis

**Homer Durand**[1] **Gherardo Varando**[2] **Gustau Camps-Valls**[1]

[1]Image Processing Lab, University of Valencia, Valencia, Spain
[2]Department of Statistics and Operational Research, University of Valencia, Valencia, Spain

## Abstract

We propose a novel approach for learning causal response representations. Our method aims to extract directions in which a multidimensional outcome is most directly caused by a treatment variable. By bridging conditional independence testing with causal representation learning, we formulate an optimisation problem that maximises the evidence against conditional independence between the treatment and outcome, given a conditioning set. This formulation employs flexible regression models tailored to specific applications, creating a versatile framework. The problem is addressed through a generalised eigenvalue decomposition. We show that, under mild assumptions, the distribution of the largest eigenvalue can be bounded by a known $F$-distribution, enabling testable conditional independence. We also provide theoretical guarantees for the optimality of the learned representation in terms of signal-to-noise ratio and Fisher information maximisation. Finally, we demonstrate the empirical effectiveness of our approach in simulation and real-world experiments. Our results underscore the utility of this framework in uncovering direct causal effects within complex, multivariate settings.

## 1 INTRODUCTION

Representation learning has been a foundational tool in modern machine learning, enabling models to automatically extract features from high-dimensional data (Bengio et al., 2013; LeCun et al., 2015). However, traditional approaches often fail to capture the causal mechanisms that underlie data generation, leading to poor generalisation under data distribution shifts. To address these shortcomings, *causal representation learning* (CRL) has emerged as a crucial approach to integrate causality into representation learning (Schölkopf et al., 2021). By learning representations that reflect the causal structure of the data, models can become more robust to distribution shifts and provide better causal insights for downstream tasks. This enables the modelling of intervention effects and the construction of counterfactuals, allowing for the analysis of questions that classical statistical models may struggle with, such as estimating the effects of policies. CRL has often been used to describe the recovery of ground-truth latent causal variables, in this work we adopt a broader perspective, focusing on the extraction of causally meaningful features from multi-dimensional, unstructured data—even in the absence of a single identifiable latent causal model—to enable inference for downstream tasks.

A key focus of causal inference literature is understanding how variables influence one another along different pathways (Pearl, 2014). Of particular interest to this work is the direct effect of a cause on an outcome variable while controlling for confounders and mediators (Pearl, 2022). Studying direct effects rather than total effects is essential for several reasons. For instance, it allows isolating specific mechanisms in science, such as assessing the effect of greenhouse gas emissions on local temperature while controlling for natural climate variations (that emissions may also influence). Finally, it helps disentangle immediate effects from delayed downstream effects, which may have a longer-term impact on the outcome.

When the outcome is multi-dimensional, identifying a subspace where the causes maximally influence it can benefit various tasks. In these cases, it is often impractical to observe how each dimension's distribution is shifted by the intervention. Therefore, it is of interest to examine this shift in a lower-dimensional space (e.g., 1-D or 2-D). This approach can also help discover simple, low-dimensional representations that capture relevant information about the intervention's effect. Additionally, it can help disentangle the direction in which the outcome is affected by the intervention from the direction where the distribution remains

unchanged. We will demonstrate that this has important implications in different application domains with a focus on climate change attribution.

While considerable work has focused on learning representations for confounder adjustment in causal effect estimation (Louizos et al., 2017), modelling representations of causes (Arjovsky et al., 2020; Peters et al., 2016), or uncovering latent causal graphs (Locatello et al., 2019), the representation of effects remains largely underexplored. Here, we aim to bridge this gap by learning a mapping of the response variable through the maximisation of a conditional independence statistic. Under certain structural assumptions, the method identifies the direction in which the effect of interventions is most observable. By using conditional expectation estimators, it adapts to different data types through various regression models.

## 2 PRELIMINARY

Let $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^d$, and $Z \in \mathbb{R}^r$ be three random vectors with density function $p(x, y, z)$ and assume that their joint distribution is absolutely continuous with respect to the Lebesgue measure. We also assume that $X$ and $Z$ are known causes of $Y$, but the relation between $X$ and $Z$ is left unspecified, allowing it to be a confounder, a mediator, or both. We aim to identify the component of $Y$ that is most directly caused by $X$, by finding $\mathbf{w}$ that maximises the causal relationship between $X$ and $\mathbf{w}^\top Y$. In the following, we clarify key terms related to the concepts of direct effect and conditional independence. We begin by considering James Woodward's *manipulationist* definition (Woodward, 2005) of a direct cause:

> *A necessary and sufficient condition for $X$ to be a direct cause of $Y$ with respect to some variable set $[Z]$ is that there be a possible intervention on $X$ that will change $Y$ (or the probability distribution of $Y$) when all other variables in $[Z]$ besides $X$ and $Y$ are held fixed at some value by interventions.*

The distribution of $Y$ under intervention is called the direct effect (DE) of $X$ on $Y$. For simplicity, we assume that the effects of $X$ and $Z$ on $Y$ are additive, as formalised in the model assumption in Sec. 4. Thus, DE can be written as:

$$DE(x) = p(Y|do(X = x), do(Z = z)). \tag{1}$$

The variable $Y$ under the intervention $do(X = x)$ is denoted as $Y^x$. We note that under the assumption of additivity, the direct effect is equivalent to the natural direct effect (see Pearl, 2009, section 4.5). In some contexts, it is described in terms of conditional expectation—referred to as the expected direct effect (EDE) (see Pearl, 2009, section 4.5.4). However, we avoid this reduction, as the (conditional) expectation masks valuable information needed to identify the direction in which $Y$ is most caused by $X$, namely $Y$'s noise structure. Additionally, the term *Gradient DE* (GDE) will be used to denote the vector of the partial derivative of Eq. (1) with respect to $x$, capturing how small variations in the intervention affect $Y$. In some cases, the Gradient Direct Effect (GDE) lies in a subspace $\mathbb{R}^q \subset \mathbb{R}^d$, meaning that the distribution $P(Y|do(X = x))$ is affected by the intervention only in this subspace, while the remaining dimensions of the space are unaffected. We refer to this as the direct effect subspace (DES). Our work focuses on recovering this reduced space, with its basis ordered by the variance in $Y$ explained by $X$, while controlling for $Z$, analogous to how Principal Component Analysis (PCA) identifies directions of maximum variance in a random vector.

We now summarise conditional independence testing, which plays an important role in our work. We say that $X$ is conditionally independent of $Y$ given $Z$, denoted $X \perp\!\!\!\perp Y \mid Z$, if for all $x \in \mathbb{R}^p$, $y \in \mathbb{R}^d$, and $z \in \mathbb{R}^r$, $p(y \mid x, z) = p(y \mid z)$, or equivalently, $p(x, y \mid z) = p(x \mid z)p(y \mid z)$. This means that, given $Z$, $X$ adds no additional information about $Y$. Let $P$ denote the joint distribution of $(X, Y, Z)$ such that $P \in \mathcal{P}$ if $X \perp\!\!\!\perp Y \mid Z$ holds (the null hypothesis), and $P \in \mathcal{Q}$ if $X \not\perp\!\!\!\perp Y \mid Z$ (the alternative hypothesis). A Conditional Independence Test (CIT) is formulated as $H_0 : P \in \mathcal{P}$ vs. $H_1 : P \in \mathcal{Q}$. Given i.i.d. observations $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{Y} \in \mathbb{R}^{n\times d}$, and $\mathbf{Z} \in \mathbb{R}^{n\times r}$, we use a statistic $\mathbf{T}_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, and reject $H_0$ when $\mathbf{T}_n$ deviates sufficiently from its expected distribution under $P \in \mathcal{P}$.

### 2.1 INTRODUCTORY EXAMPLE

Through a simple example, we demonstrate that EDE is generally suboptimal for distinguishing $Y$ distributions under different interventions and that strategically maximising a CIT statistic may be more effective. Let us consider the simple linear model $Y = \mathbf{b}X + \mathbf{c}Z + N$ with $Y \in \mathbb{R}^d$, $X \in \mathbb{R}$, $Z \in \mathbb{R}^p$, and $N \in \mathbb{R}^d$. Let also $\mathbf{\Sigma}$ denote the covariance matrix of $Y^x$. The relationship between $X$ and $Z$ is not relevant in this context, as we focus on the intervention distribution $Y^x$, and such an intervention breaks the statistical association between $X$ and $Z$.

In a linear model, the Gradient EDE is given by the weight vector $\mathbf{b}$ for interventions on $X$ (Peters et al., 2017, ex. 6.42), often called the direct effect. This means that the distribution of $Y$ is only shifting along the $\vec{\mathbf{b}}$ axis when intervening on $X$ (see Fig. 1). The most common approach to find $\mathbf{b}$ is by analysing the weights of $X$ in the conditional expectation $\mathbb{E}[Y|X, Z]$. Alternatively, $\mathbf{b}$ can be obtained by maximising the partial correlation between $X$ and $\mathbf{w}^\top Y$ given $Z$. When $N$ is isotropic, the vector $\mathbf{w}$ that maximises this partial correlation is indeed $\mathbf{b}$. Since partial correlation is used in CITs, through Fisher's $Z$ transformation (Fisher,

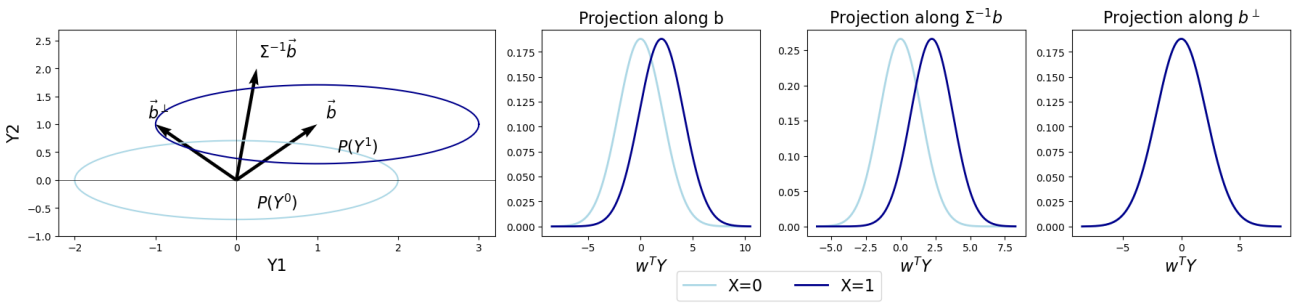



Figure 1: Illustration of the linear model from Sec. 2.1 with $\mathbf{b} = (1, 1)^\top$ and $\mathbf{\Sigma} = (4, 0; 0, 1/2)$, showing the one-sigma ellipsoid for $Y^0$ and $Y^1$. For one-dimensional $X$, $Y^x$ shifts along $\mathbf{b}$, but projection along $\mathbf{b}$ is suboptimal. In contrast, projection along $\mathbf{\Sigma}^{-1}\mathbf{b}$ is optimal, with $(\mathbf{\Sigma}^{-1}\mathbf{b}, \mathbf{b}^\perp)$ forming a natural basis for the intervention space, where the first axis captures the intervention effect and the second contains no information.

1915), which applies the *arctanh* function to the partial correlation, the Gradient EDE can be recovered by finding the direction that maximises a CIT statistic for $\mathbf{w}^\top Y$.

However, when the noise $N$ is non-isotropic, $\mathbf{b}$ may not be optimal for isolating the causal effect of $X$ on $Y$. In this case, the direction $\mathbf{b}$ may align too closely with the noise structure of $Y$, making the intervention's effect less discernible. While regression-based approaches fail to account for the noise structure, CIT statistics balance signal detection (the effect of the intervention) and noise reduction to obtain optimal power. For non-isotropic noise, it can be shown that the most discriminative direction for the intervention is $\mathbf{\Sigma}^{-1}\mathbf{b}$. While EDE generally fails, maximising the partial correlation recovers this optimal direction. This illustrates how identifying the direction in which $Y$ maximises a conditional independence statistic can effectively uncover the subspace of $Y$ most caused by $X$.

This example is illustrated in Fig. 1, where we observe that projection along $\mathbf{\Sigma}^{-1}\mathbf{b}$ improves the separability of distributions under different interventions (here $X = 0$ and $X = 1$). A natural basis for representing interventions on $Y$ is then $(\mathbf{\Sigma}^{-1}\mathbf{b}, \mathbf{b}^\perp)$, where the first vector captures all information about the intervention, and the second contains no information. These axes need not be orthogonal. Under favorable conditions, such as a rapid decay in the eigenvalues of the covariance matrix $\mathbf{\Sigma}$, the noise in the distribution of $Y$ along the optimal direction $\mathbf{\Sigma}^{-1}\mathbf{b}$ diminishes as the dimensionality increases, concentrating the distribution's mass in a single point and achieving optimal separability of the intervention distributions.

## 2.2 RELATED WORK

Although the idea of learning representations of effects of causes is, to our knowledge, novel, there are important connections between our work and other fields. It intersects two key areas of statistical learning: conditional independence testing and causal representation learning. Below, we summarise the most relevant results in these and other relevant fields.

**Conditional Independence Testing:** A variety of methods address this problem, broadly classified into nonparametric and parametric approaches. Nonparametric methods, like kernel-based tests (Zhang et al., 2011), nearest-neighbour methods (Runge, 2018), and mutual information-based tests (Fukumizu et al., 2008), offer flexibility but are computationally expensive. Regression-based approaches (Shah and Peters, 2018b) test residual dependencies or whether $X$ improves prediction of $Y$ given $Z$ (Chow, 1960). Parametric methods, such as partial canonical correlation analysis (CCA) (Rao, 1969), assume linearity and Gaussianity, providing computational efficiency at the cost of strong assumptions. While these methods balance complexity, power, and robustness, they do not explicitly recover an optimal subspace for testing, though they may indirectly solve an optimisation problem that achieves this, as we will demonstrate.

**Causal Representation Learning (CRL):** CRL (Schölkopf et al., 2021) aims to learn representations that capture causal mechanisms, enhancing generalisation, interpretability, and robustness. Leveraging invariance across environments (Arjovsky et al., 2020), recent methods focus on learning representations for confounders or predictors to estimate causal effects (Yao et al., 2018; Yang et al., 2021; Locatello et al., 2019), with some extending to temporal data (Lachapelle et al., 2022; Lippe et al., 2022). Recent work from Saengkyongam et al. (2023) shows how CRL is a promising avenue for out-of-distribution generalisation. While prior work targets confounder or predictor representations, our method focuses on causal effect representation of the outcome, filling a gap in previous approaches.

**Connections to Signal Detection:** Our framework relates to signal detection (Macmillan, 2002; Kay, 1998, 1993), aiming to identify a deterministic signal $X$ in noisy observations $Y = X + N$. In climate science, this is addressed by the "optimal fingerprint" (Hasselmann, 1993), which

maximises the signal-to-noise ratio of a linear projection of observations. This enables a direct test for the detection of climate change while recovering a useful climate pattern.

**Sufficient Dimensionality Reduction (SDR):** There are also similarities with the SDR framework (Globerson and Tishby, 2003; Fukumizu et al., 2009), which aims to find a sufficient statistic $\mathbf{w}^\top X$ such that $p(Y|X) = p(Y|\mathbf{w}^\top X)$. The reduced space therefore contains all the relevant information in $X$ to predict $Y$. Our work focuses on finding a sufficient statistic specifically for the DE, to know, a subspace that retains all relevant information about the DE.

# 3 LEARNING FRAMEWORK

Our goal is to identify the components of $Y$ that are most caused by $X$, conditional on $Z$, assuming all confounders $C \subseteq Z$ are observed and the causal relationship $X \to Y$ is known. Specifically, we aim to find a subspace of $Y$ that encapsulates all information about interventions on $X$. To achieve this, we represent the subspace as a linear transformation, $\tilde{Y} = \mathbf{W}^\top Y \in \mathbb{R}^q$, where $\mathbf{W} \in \mathbb{R}^{d\times q}$. For simplicity, we focus on the case where $q = 1$, and identify a vector $\mathbf{w} \in \mathbb{R}^d$ such that $\mathbf{w}^\top Y \in \mathbb{R}$ captures the maximum amount of information that a one-dimensional representation of $Y$ can convey about the intervention on $X$. The case for $q > 1$ is discussed in Section A.3.

## 3.1 MAXIMISATION OF A CIT STATISTIC

We propose a class of learning algorithms that maximise a CIT statistic to find $\mathbf{w}$, following the optimisation problem:

$$\mathbf{w}^\star = \arg\max_{\mathbf{w}} T(X, \mathbf{w}^\top Y, Z). \tag{2}$$

Here, $X$, $\mathbf{w}^\top Y$, and $Z$ are treated as random variables, as we consider a *population* version of the test statistic. This formulation provides theoretical guarantees for recovering the latent structure (see Sec. 4.1) and the optimality of the learned representation in terms of Fisher information. We denote $T$ the population loss and $\mathbf{T}_n$ its empirical counterpart.

Building on this idea, we propose a flexible framework based on nested predictive models of $Y$. This approach assesses conditional independence by analysing the residuals from two regression models. The restricted model regresses $Y$ on $Z$ alone, while the full model includes both $X$ and $Z$. Conditional independence is evaluated by comparing the residuals of these models, without assuming a specific functional form between $X$ and $Y$. This flexibility makes the framework broadly applicable across various settings, accommodating complex, nonlinear relationships between variables. Let us define $R^2_{\text{full}}(\mathbf{w}) = \mathbb{E}\left[(\mathbf{w}^\top Y - \mathbb{E}[\mathbf{w}^\top Y|X, Z])^2\right]$ and $R^2_{\text{res}}(\mathbf{w}) = \mathbb{E}\left[(\mathbf{w}^\top Y - \mathbb{E}[\mathbf{w}^\top Y|Z])^2\right]$ as the population mean squared error when predicting $\mathbf{w}^\top Y$ from the full model (including both $X$ and $Z$) and the restricted model (including only $Z$), respectively. A straightforward way of enforcing conditional dependence–maximising the *distance* between $p(y|x, z)$ and $p(y|z)$–is to maximise the distance between the residuals of the full regression model and the restricted one. This leads to the simple loss function:

$$T_S(X, Y, Z; \mathbf{w}) = R^2_{\text{res}}(\mathbf{w}) - R^2_{\text{full}}(\mathbf{w}). \tag{3}$$

Under the null hypothesis, both regression models have equal predictive power, but the full model, with more degrees of freedom, yields smaller residuals. This can also be viewed through an information theory perspective, detailed further in Sec. A.1. However, this loss function is unbounded with respect to $\mathbf{w}$; thus, it is necessary to impose additive constraints on $\mathbf{w}$ to avoid trivial solutions. The most straightforward way to constrain the loss is to limit $\mathbf{w}$ to be a unit norm vector, i.e., $\|\mathbf{w}\| = 1$. We show in Lemma B.1 in supplementary material that this approach recovers the EDE and is thus suboptimal for non-isotropic noises.

Another approach, is to constrain the full residuals to be fixed, leading to the following loss function:

$$T_F(X, Y, Z; \mathbf{w}) = \frac{R^2_{\text{res}}(\mathbf{w}) - R^2_{\text{full}}(\mathbf{w})}{R^2_{\text{full}}(\mathbf{w})}. \tag{4}$$

In the context of a linear Gaussian SCM, this statistic can be interpreted as an F-test between nested models (aka Chow test (Chow, 1960)), which is commonly used for variable selection (Hocking, 1976) or causal discovery (Nogueira et al., 2022). When the conditioning set $Z$ consists of the past values of $Y$, the empirical version of $T_F$ corresponds to the statistic of the well-known Granger causality test (Granger, 1969). In this context, the maximisation of $T_F$ with respect to $\mathbf{w}$ leads to a causal representation method known as Granger PCA (Varando et al., 2022). This further emphasises how maximising a conditional independence testing statistic can be leveraged to uncover the direction in which $Y$ is most strongly caused by $X$. Another possible constraint is grounded in detection theory (Macmillan, 2002; Kay, 1993). Considering that $Y$ can be decomposed into a signal term $S$ (variance related to $X$) and a noise term $N$ (variance related to $Z$ and $Y$'s intrinsic noise), we constrain the variance of $\mathbf{w}^\top N$. Assuming that the signal and noise are additive in $Y$, this constraint relates to constraining $R^2_{\text{noise}} = \mathbb{E}[(\mathbf{w}^\top Y - \mathbb{E}[\mathbf{w}^\top Y \mid X, Z = 0])^2]$. We thus propose the loss function:

$$T_D = \frac{R^2_{\text{res}}(\mathbf{w}) - R^2_{\text{full}}(\mathbf{w})}{R^2_{\text{noise}}(\mathbf{w})}. \tag{5}$$

It will be shown in Sec. 4.1 that this formulation is optimal under certain structural assumptions.

Canonical Correlation Analysis (CCA) (Hotelling, 1992) and its partial variant Rao (1969) also seek a subspace that

captures reduced information between $X$ and $Y$ (conditioning on $Z$ in partial CCA), enabling (conditional) independence testing. In Sec. A.2, we demonstrate that partial CCA aligns with our framework by interpreting it as the maximisation of a conditional independence statistic.

### 3.2 EMPIRICAL ESTIMATORS

We now present the practical optimisation procedure to estimate $\mathbf{w}^\star$. Given observation (or design) matrices $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{Y} \in \mathbb{R}^{n\times d}$, and $\mathbf{Z} \in \mathbb{R}^{n\times r}$, we now present empirical estimators for $\mathbf{w}_S$, $\mathbf{w}_F$ and $\mathbf{w}_D$.

Similarly, we assume that we have two estimators $\hat{g}_{\text{full}}(X, Z)$ and $\hat{g}_{\text{res}}(Z)$ for the conditional expectations $\mathbb{E}[Y \mid X, Z]$ and $\mathbb{E}[Y \mid Z]$, respectively. The learning algorithms employed to estimate these conditional expectations are not restricted, allowing users to tailor them based on their assumptions about the relationships within the data and their prior knowledge. We denote by $\hat{\mathbf{\Sigma}}_{\text{full}}$, $\hat{\mathbf{\Sigma}}_{\text{res}}$, and $\hat{\mathbf{\Sigma}}_{\text{noise}}$ the sample covariance matrices of the residuals from the full and restricted models, as well as the noise covariance. The three population losses can be maximised by solving the general eigenvalue problem $\hat{\mathbf{M}}\mathbf{w} = \lambda\hat{\mathbf{N}}\mathbf{w}$, where $\hat{\mathbf{M}} = \hat{\mathbf{\Sigma}}_{\text{res}} - \hat{\mathbf{\Sigma}}_{\text{full}}$ and $\mathbf{N}$ corresponds to the constraints on $\mathbf{w}$: $\hat{\mathbf{N}} = \mathbf{I}$ for $T_S$, $\hat{\mathbf{N}} = \hat{\mathbf{\Sigma}}_{\text{full}}$ for $T_F$, and $\hat{\mathbf{N}} = \hat{\mathbf{\Sigma}}_{\text{noise}}$ for $T_D$. Given random realisations of $(X, Y, Z)$, the population matrices $\mathbf{M}$ and $\mathbf{N}$ are random, typically following a Wishart distribution. Under this condition, the first eigenvalue of the GEV problem, denoted by $\Lambda_1$, is also random. Upon observing data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, the empirical covariances $\hat{\mathbf{M}}$ and $\hat{\mathbf{N}}$ are fixed, and we obtain a realisation $\lambda_1$ of $\Lambda_1$ with corresponding eigenvector $\mathbf{w}_1$. We denote the eigen-pairs $(\lambda_S, \mathbf{w}_S)$, $(\lambda_F, \mathbf{w}_F)$, and $(\lambda_D, \mathbf{w}_D)$ as those corresponding to the first eigenvalues for the losses $T_S$, $T_F$, and $T_D$, respectively.

An algorithm for the estimation of the eigen-pairs is available in Algo. 1. The convergence properties of these estimators are presented in Th. B.10 in the supplementary materials. Additional details on the estimation of the conditional expectations, as well as the estimation of other components and the stability of the solution, can be found in Section A.3.

**Algorithm 1:** Direct Effect Analysis Algorithm

**Input:** Data matrices $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, components $K$, $solver \in \{T_S, T_F, T_D\}$, learning algorithms $g_{res}, g_{full}$.
**Output:** Matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$

Initialize $\mathbf{W} \leftarrow [\ ]$, $\mathbf{Y}^{(1)} \leftarrow \mathbf{Y}$
**for** $k = 1$ **to** $K$ **do**
  Train $\hat{g}_{res}$, $\hat{g}_{full}$ using $(\mathbf{Y}^{(k)}, \mathbf{X}, \mathbf{Z})$
  $\mathbf{N}_{\text{res}} \leftarrow \mathbf{Y}^{(k)} - \hat{g}_{res}(\mathbf{Z})$
  $\mathbf{N}_{\text{full}} \leftarrow \mathbf{Y}^{(k)} - \hat{g}_{full}(\mathbf{X}, \mathbf{Z})$
  **if** *solver* == $T_D$ **then**
    $\mathbf{N}_{\text{noise}} \leftarrow \mathbf{Y}^{(k)} - \hat{g}_{full}(\mathbf{X}, \mathbf{0})$
  Compute $\mathbf{\Sigma}_{\text{res}}$, $\mathbf{\Sigma}_{\text{full}}$, and (if $T_D$) $\mathbf{\Sigma}_{\text{noise}}$ from residuals.
  $\hat{\mathbf{M}} \leftarrow \mathbf{\Sigma}_{\text{res}} - \mathbf{\Sigma}_{\text{full}}$
  $\hat{\mathbf{N}} \leftarrow \mathbf{I}$ (if $T_S$), $\hat{\mathbf{N}} \leftarrow \mathbf{\Sigma}_{\text{full}}$ (if $T_F$), $\hat{\mathbf{N}} \leftarrow \mathbf{\Sigma}_{\text{noise}}$ (if $T_D$).
  Solve GEV: $\hat{\mathbf{M}}\mathbf{w} = \lambda\hat{\mathbf{N}}\mathbf{w}$, normalize and append $\mathbf{w}_k$ to $\mathbf{W}$.
  $\mathbf{Y}^{(k+1)} \leftarrow \mathbf{Y}^{(k)} - \sum_{i=1}^{k} \mathbf{Y}^{(k)}\mathbf{w}_i\mathbf{w}_i^\top$
**return** $\mathbf{W}$

## 4 THEORETICAL GUARANTEES

In this section, we discuss the theoretical properties of the maximisation of the statistics introduced earlier. We consider the distribution of $(X, Y, Z) \sim P$ entailed within the following Structural Causal Model (SCM):

$$Y := \mathbf{b}\phi(X) + \psi(Z) + N_y, \tag{6}$$

where $\phi(x) : \mathbb{R}^p \to \mathbb{R}$, $\psi(z) : \mathbb{R}^r \to \mathbb{R}^d$, $\mathbf{b} \in \mathbb{R}^d$ and with $N_y \sim \mathcal{N}(0, \mathbf{\Sigma})$. Again, the relationship between $X$ and $Z$ is left undefined as applying $do(X)$ breaks any statistical dependencies that existed in the observational setting. We denote by $\mathbf{\Sigma}_{\psi(z)}$ the covariance of $\psi(Z)$. For the remainder of this section, we assume that the intervention $\phi(x)$ is bounded. Concretely, we assume that $X$ goes threw an information bottleneck of dimension one. The vector $\mathbf{b}$ thus gives the direction of the causal effect as intervention on $X$ will shift along axis $\mathbf{b}$. Note that if $\phi(x)$ is linear, it corresponds to the Gradient EDE. All the proofs are given in Sec. B in supplementary materials.

The notion of the "direction most caused by" an intervention is ambiguous when considering the full distribution rather than just its mean, as in EDE. We seek a direction along which interventions on $X$ induce the most significant changes in the distribution of $Y$, capturing the overall distribution shift rather than relying solely on mean effects. To quantify this, we analyze the signal-to-noise ratio (SNR) and Fisher information, which serve as natural measures of intervention-induced variation. In the Gaussian case, where distributions are fully characterized by their first two moments, these measures provide a principled way to identify the most affected direction. Given our generative model, we show that our algorithm optimally identifies directions that maximize SNR, effectively separating interventional distributions. Furthermore, under the assumption that $\phi(x)$ is linear and that noise is gaussian, we prove that these directions also maximize the Fisher information of $\mathbf{w}^\top Y$ with respect to the intervention, ensuring the most effective disentanglement of $Y$ under infinitesimal intervention shifts.

### 4.1 CAUSAL EFFECT REPRESENTATION

To better understand the properties of the different learning algorithms, it is useful to decompose the intervention distribution $Y^x$ into a signal term and a noise term $Y^x = S(x) + N$ where $S(x) = \mathbf{b}\phi(x)$ represents the EDE, a non-random component of $Y^x$, and the noise term is given by $N = \psi(Z) + N_y$, which remains random. We define the SNR of the transformed variable $\mathbf{w}^\top Y^x = \mathbf{w}^\top S(x) + \mathbf{w}^\top N$ as

$$\gamma^2(\mathbf{w}) = \frac{(\mathbf{w}^\top S(x))^2}{\mathbf{w}^\top \mathbf{\Sigma}_N \mathbf{w}}, \tag{7}$$

where $\mathbf{\Sigma}_N$ is the covariance matrix of the noise term. Notably, when the conditioning set $Z$ is accounted for, the noise covariance $\mathbf{\Sigma}_N$ simplifies to $\mathbf{\Sigma}$. In this case, the optimality results that will be established for $\mathbf{w}_D$ also apply to $\mathbf{w}_F$. We now present some optimality results related to the SNR. This metric is tied to an optimal representation because, as the SNR increases, the distribution becomes more concentrated around the signal $S(x)$ Thus, the direction that maximises the SNR is the one for which small perturbations of the intervention are most observable. We thus say that a weight vector $\mathbf{w}$ is optimal if it maximises $\gamma^2(\mathbf{w})$. For general noise structures, $\mathbf{w}_D$ is shown to be optimal.

**Proposition 4.1** (General optimality)**.** *Assuming $P$ is entailed in the SCM in* (6)*, we have that $\mathbf{w}_D$ is optimal.*

Under stronger assumptions – isotropy of the noises – both $\mathbf{w}_S$ and $\mathbf{w}_F$ are shown to be optimal.

**Proposition 4.2** (Optimality under isotropic noise)**.** *Assuming that $P$ is entailed in the SCM in* (6) *and that $\mathbf{\Sigma}_N$ is isotropic, we have that both $\mathbf{w}_S$ and $\mathbf{w}_D$ are optimal. Moreover, if $\mathbf{\Sigma}$ is also isotropic, then $\mathbf{w}_F$ is also optimal.*

This proposition implies that when the effects of $X$ and $Z$ are assumed to be separable, $\mathbf{w}_D$ is optimal in the sense that it maximises the SNR.

We now present different guarantees for the learned representation, demonstrating that in the large-dimensional regime, and under specific conditions on the characteristics of $\mathbf{b}$, $\mathbf{\Sigma}$, and $\mathbf{\Sigma}_{\psi(z)}$, the signal-to-noise ratio improves as the dimensionality of $Y$ increases, such that the signal of $\mathbf{w}^\top Y$ completely dominates its noise.

**Proposition 4.3** (Noise term behavior)**.** *Let $\|\mathbf{b}\|^2 = o(\nu_1(d))$, $\mathbf{b}^\top(\mathbf{\Sigma} + \mathbf{\Sigma}_{\psi(z)})\mathbf{b} = o(\nu_2(d))$, $\mathbf{b}^\top\mathbf{\Sigma}^{-1}\mathbf{b} = o(\nu_3(d))$, $\mathbf{b}^\top(\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\psi(\mathbf{Z})}\mathbf{\Sigma}^{-1})\mathbf{b} = o(\nu_4(d))$, and $\mathbf{b}^\top(\mathbf{\Sigma} + \mathbf{\Sigma}_{\psi(z)})^{-1}\mathbf{b} = o(\nu_5(d))$. Here $\nu_i$ denotes the rates of growth with regard to $d$.*

*Assume the distribution $P$ follows the structural causal model in Eq. 6, and the following conditions hold:* ***1.*** $\lim_{d\to\infty} \frac{\nu_1(d)}{\nu_2(d)} \to \infty$, ***2.*** $\lim_{d\to\infty} \frac{\nu_3^2(d)}{\nu_4(d)} \to \infty$ *and* ***3.*** $\lim_{d\to\infty} \nu_5(d) \to \infty$.

*The following convergence properties hold: $\gamma^2(\mathbf{w}_S) \to \infty$ if condition **1** holds, $\gamma^2(\mathbf{w}_F) \to \infty$ if condition **2** holds, $\gamma^2(\mathbf{w}_D) \to \infty$ if condition **1**, **2** or **3** holds.*

In general, the above conditions reflect the fact that $\mathbf{b}$ is unaligned with $\mathbf{\Sigma}$, meaning that large values of $\mathbf{b}$ correspond to small values of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{\psi(Z)}$. This relationship can also be interpreted in terms of the growth of the largest eigenvalue of $\mathbf{\Sigma}$ or of $\|\mathbf{b}\|^2$, independently. All of these conditions are related to the observation that as the dimensionality increases, $Y^x$'s distribution contains 'more signal' relative to its noise level. This phenomenon occurs, for example, when the sources of noise are limited and the resolution of the observations is increased. We provide further details and insights on these assumptions in Sec. A.4. As discussed above, a strong SNR indicates that the recovered signal is closer to the information bottleneck $\phi(x)$. More importantly, it also implies better separability of the distributions of $Y^x$ along the projected axis. This can be formalised by considering the Fisher information of $\mathbf{w}^\top Y^x$ with respect to $x$, given by:

$$I_{\mathbf{w}}(x) = \mathbb{E}\left[U(x)U(x)^\top\right],$$

with $U(x) = \nabla_x \log P(\mathbf{w}^\top Y \mid do(X = x))$ denoting the score function. We now show that for linear models, the Fisher information and the SNR of $\mathbf{w}^\top Y^x$ are equivalent up to a positive scaling factor.

**Proposition 4.4** (Equivalence between Fisher information and SNR)**.** *Consider a SCM as described in* (6)*, and let the intervention function be $\phi(x) = \mathbf{v}^\top x$, where $\mathbf{v} \in \mathbb{R}^d$. Then, the SNR is proportional to the Fisher Information of the intervention, i.e. $I_{\mathbf{w}}(x) = \alpha\gamma^2(\mathbf{w})$ with $\alpha \in \mathbb{R}^+$.*

Applying this result to $T_D$, we obtain an optimality guarantee in terms of Fisher information.

**Corollary 4.5.** *Under the assumptions of Prop. 4.4, the optimal solution $\mathbf{w}_D$ maximises the Fisher information $I_{\mathbf{w}}(x)$.*

A similar result to Prop. 4.3 can also be derived for Fisher information under the assumption of a linear effect of $X$ on $Y$. Thus, the optimality conditions for recovering the bottleneck structure $\phi(x)$ translate into conditions for the discriminative power of the learned representation. In this setting, maximising the SNR is equivalent to maximising the Fisher information, which quantifies the sensitivity of the projected distribution to changes in the intervention parameter. This can be better understood by examining the relationship between Fisher information and the Kullback-Leibler divergence (see A.1). Specifically, it measures the distance between parametric distributions, where in this case, the parameter corresponds to the intervention value. For linear

models, $T_D$ is optimal as it maximises the distributional divergence induced by infinitesimal perturbations of the intervention. This enhances the discriminative power of the learned representation across different interventions. Moreover, higher Fisher information indicates that the learned representation retains more information about the intervention.

### 4.2 TESTING THE PRESENCE OF A DIRECT EFFECT

We now explore a direct implication of our problem formulation. Since we are maximising a test statistic for conditional independence testing, we can derive the distribution of the loss function under the null hypothesis. Consequently, one can reject the hypothesis of conditional independence at level $\alpha$ if the value of the loss function, specifically the largest eigenvalue $\lambda_1$, exceeds a critical threshold.

**Proposition 4.6** (Distribution of $\lambda_F$ under conditional independence)**.** *Let the distribution $P$ be induced by the SCM in* (6) *with linear assignments and Gaussian noise, and assume $p = q = 1$. Under the null hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$, the largest root $\Lambda_F$ is $F$-distributed such that $(dfn/dfd)\Lambda_F \sim F(dfd, dfn)$ where $dfn = d$ and $dfd = n - p - r - 1$.*

Finding the distribution of $\Lambda_D$ is more challenging. Instead, we establish an upper bound on $\Lambda_D$'s distribution, allowing the distribution of $\Lambda_F$ to serve as a proxy for computing upper bounds on the p-values of $\Lambda_D$.

**Proposition 4.7** (Upper bound on $\Lambda_D$ under conditional independence)**.** *Under similar assumptions as in Prop 4.6 we have under the null hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$ that $P(\Lambda_D \geq \lambda_D | H_0) \leq P(\Lambda_F \geq \lambda_D | H_0)$.*

Testing is straightforward by rejecting the null hypothesis if $(dfd/dfn)\lambda_1$ deviates sufficiently from $F(dfn, dfd)$. This property is useful for testing whether the learned representation (Sec. 4.1) captures a meaningful effect of $X$ on $Y$.

## 5 EXPERIMENTS

In this section, we present the results of our extensive simulation experiments designed to support our theoretical findings. Additionally, we provide a straightforward use case from climate science detection and attribution to illustrate the practical relevance of our approach. The code for all experiments is available at this github repository.

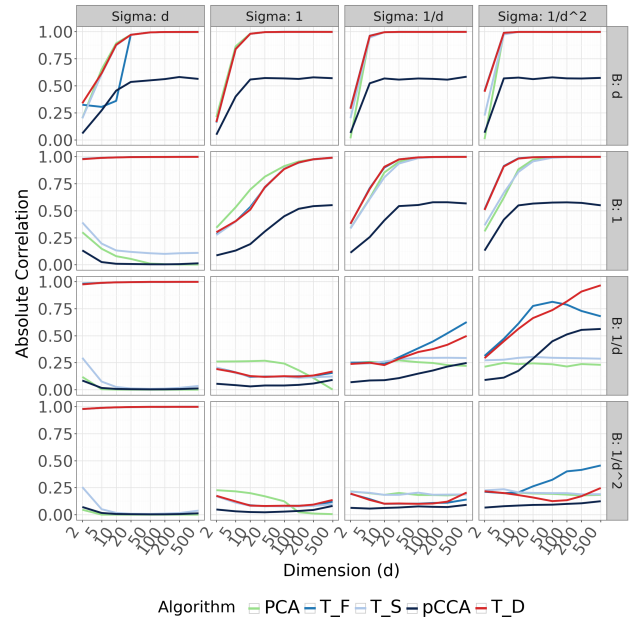



Figure 2: Correlation between $\mathbf{w}^\top Y$ and $\phi(X)$ as $d$ increases. $T_D$ consistently outperforms all methods, recovering $\phi(X)$ as $d$ grows, provided that $\mathbf{b}$ faster than $\mathbf{\Sigma}$. When $T_F$, $T_S$ or PCA are not visible, they are overlapped by $T_D$. See Fig. 4 for the (5, 95) percentiles.

### 5.1 SIMULATION EXPERIMENTS

We simulate data from a linear SCM with Gaussian noise, where $Z$ acts as a confounder for both $X$ and $Y$ (6).

$$\begin{aligned} N_x, N_z &\sim \mathcal{N}(0, \mathbf{I}), \\ N_y &\sim \mathcal{N}(0, \mathbf{\Sigma}), \\ Z &:= N_z, \\ X &:= f_a(\mathbf{C}^\top Z) + N_x, \\ Y &:= u\mathbf{b}^\top f_a(\Gamma^\top X) + v f_a(\mathbf{D}^\top Z) + w N_y. \end{aligned} \tag{8}$$

The noise terms $N_x$ and $N_z$ are independent. For the nonlinear case, we define $f_a(z) := \exp(-z^2/2)\sin(az)$, where $a$ controls nonlinearity. In the linear case, $f_a$ is set to the identity. The coefficients $\Gamma, \mathbf{b}, \mathbf{C}, \mathbf{D}$ are uniformly sampled from $[0, 1]$. We run 10 repetitions for each sample size $n$ and dimension $d$, reporting median values and quartiles. The scalars $u$, $v$ and $w$ allow us to balance the weights of the effect of $X$, $Z$ and of the noise $N_y$.

**Causal Effect Representation** We assess the performance of our algorithm in recovering the direct effect of $X$ on $Y$, modeled as $f_a(\Gamma^\top X)$. The recovery is tested as $d$ increases and with varying noise structures. We set $p = r = 10$, and use $n = 2000$ samples for robust evaluation. Performance is evaluated by the absolute correlation between $\mathbf{w}^\top Y$ and $f_a(\Gamma^\top X)$, comparing nested models ($T_S$, $T_F$, $T_D$) against PCA and partial CCA (pCCA)

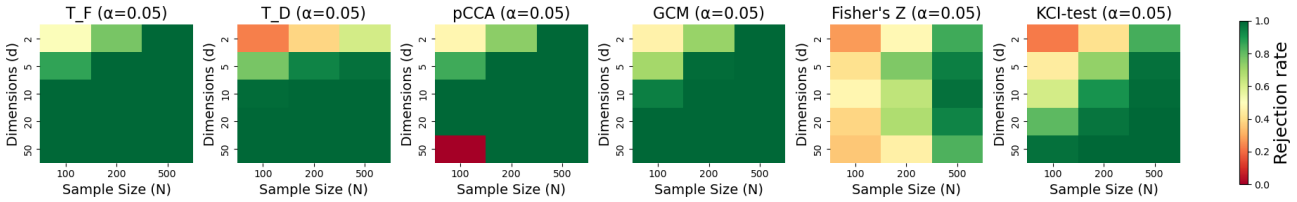



Figure 3: Power of the test for $\alpha = 0.05$. A detailed experiments with different values $\alpha$ is available in Fig. 10

as baselines. To understand the contexts where learning algorithms may fail to fully recover $\phi(X)$, we consider various configurations of $\mathbf{\Sigma}$ and $\mathbf{b}$. We set $\mathbf{\Sigma}$ to be diagonal and explore four sets of entries for $Diag(\mathbf{\Sigma})$ and $\mathbf{b}$: $(1, \ldots, i, \ldots, d)$, $(1, \ldots, 1)$, $(1, \ldots, 1/i, \ldots, 1/d)$, and $(1, \ldots, 1/i^2, \ldots, 1/d^2)$. Our main observation is that when $\mathbf{b}$ grows slowly relative to $\mathbf{\Sigma}$, none of the methods fully recover the signal. Specifically, pCCA tends to converge to a correlation of approximately 0.75, as it only recovers the part of $\phi(X)$ independent of $Z$—the signal correlated with $\psi(Z)$ is regressed out from both residuals before regression. This behavior is clearer in Appendix figure 8, where $X$ and $Z$ are simulated as independent variables, and pCCA can recover $\phi(X)$. Additionally, we observe that $T_F$ and $T_D$ outperform $T_S$ when $\mathbf{b}$ grows too slowly relative to the noise. Both $T_F$ and $T_D$ effectively control the variance contributions from $\mathbf{\Sigma}$, resulting in better performance in these challenging contexts (figure 2 in appendix). We analyse recovery across various noise configurations for $\mathbf{\Sigma}$: Diagonal, Full-rank, and Low-rank (rank = 10). We also test three weighting schemes: *equal*, *strong_N_Y*, and *strong_Z*, setting $(u, v, w)$ in (8) by $(1/3, 1/3, 1/3)$, $(0.1, 0.1, 0.8)$, and $(0.1, 0.8, 0.1)$, respectively. As shown in Appendix figure 5, $T_D$ consistently outperforms other methods, with correlation approaching 1 as $d$ increases. Similar trends are observed in nonlinear and high-dimensional cases (Appendix figures 7, 6).

**Level and Power of the Test** We assume that the data are generated from a linear SCM with Gaussian noise, where $f_a(Z) = Z$ and set $p = r = q = 1$. Our analysis compares tests based on the optimisation of $T_F$ and $T_D$ against four common conditional independence (CI) tests: partial CCA (Rao, 1969), the Generalised Covariance Measure (GCM) (Shah and Peters, 2018b), Fisher's Z test (Kalisch and Bühlman, 2007), and the Kernel Conditional Independence (KCI) test (Zhang et al., 2012). The primary focus is on test performance with respect to sample size and $Y$'s dimensionality. All tests maintain valid control of false positives when $d < n$ (see figure 11), ensuring effective Type I error control. However, for test power (see figure 3), Fisher's Z and KCI show lower performance, especially for small samples and large $d$, due to their broader hypothesis set $\mathcal{P}$, which includes potentially nonlinear relationships. Tests based on $T_F$, $T_D$ and pCCA leverage $Y$'s dimensionality, show better performance with higher dimensions for fixed sample sizes. This contrasts with Fisher's $Z$, which performance does not increases with $d$.

## 5.2 REAL-WORLD EXPERIMENTS

We present two real-world climate detection and attribution experiments: the first leverages the algorithm's ability to learn disentangled representations, and the second applies $T_D$ to test causal effects.

**Separating internal climate variability from the externally forced response.** We evaluate the ability of our method to disentangle internal climate variability from the externally forced response using temperature fields from CESM2 historical climate simulations (Danabasoglu et al., 2020). Use of the optimal projection $\mathbf{w}_D$ is compared against two commonly used baselines in climate science: *Detrending* and *Dynamical Adjustment* (Sippel et al., 2019). To achieve this, we model internal variability using Sea Level Pressure (SLP) as a proxy and estimate the externally forced response using a smoothed version of the Global Mean Temperature (GMT). $T_D$ learns a projection that isolates the internal component of temperature fluctuations while preserving their dynamical structure. Once trained, the model allows us to separate the forced and internal components of temperature fields. Figure 12 presents the mean squared error (MSE) for trend estimation across different algorithms. $T_D$ performs comparably to Detrending for reconstructing forced trends but performs better in recovering internal variability trends, providing better worst-case control. The spatial distribution of estimated internal trends (Figures 13 and 14) further highlights that both methods capture large-scale patterns but tend to underestimate trends in polar regions. Additionally, Figure 15 illustrates that $T_D$ effectively reconstructs the forced response across different locations, although both $T_D$ and Detrending struggle in highly variable regions. Overall, our approach provides a principled framework for disentangling forced and internal climate variability.

**Climate change attribution.** In this experiment, we examine the direct effects of external forcing and investigate whether external forcing factors—such as aerosols, $CO_2$, $CH_4$, and land use have a direct effect on the annual mean temperature field ($Y_{\text{factual}}$). Using 50 historical climate sim-

Table 1: Performance comparison of different approaches for detecting various effects. Bold values indicate the lowest Type II Error and Type I Error at level $5\%$.

| Effect | Approach | Type II Err. | Type I Err. |
|---|---|---|---|
| $CO_2$ | DEA | **0.00** | **0.00** |
| | GMT Reg | 0.06 | 0.30 |
| | EOF | 0.06 | 0.30 |
| $CH_4$ | DEA | **0.52** | **0.00** |
| | GMT Reg | 0.70 | 0.30 |
| | EOF | 0.74 | 0.26 |
| Aerosol | DEA | **0.00** | **0.04** |
| | GMT Reg | 0.76 | 0.24 |
| | EOF | 0.76 | 0.24 |
| Land Use | DEA | **0.00** | **0.14** |
| | GMT Reg | 0.36 | 0.64 |
| | EOF | 0.74 | 0.26 |

ulations from CESM2, we compute counterfactual temperature fields ($Y_{\text{counterfactual}}$), following the methodology described in Eq. 36 in the supplementary materials. We apply the algorithm $T_D$ to test for the significance of each forcing ($X$) while controlling for the effects of the others ($Z$). Our results are compared to two common approaches in climate attribution (Lean and Rind, 2008): regression-based tests where forcings are assessed for their significance in predicting climate patterns, specifically Global Mean Temperature (spatial average) and the first Empirical Orthogonal Function (EOF) of the climate field. The findings demonstrate that our method effectively controls type I error (when applied to $Y_{\text{factual}}$) and type II error (when applied to $Y_{\text{counterfactual}}$) and outperforms the other approaches. These results highlight the potential of our method in attributing causal effects of external forcing, with implications for its use in analysing observational data, such as the ERA5 or HADCRUT datasets.

# 6 CONCLUSION

This paper proposes a novel framework for recovering the direct effect of low-rank interventions in multivariate response variables. Our approach combines conditional independence testing and causal representation learning, enabling robust estimation of direct causal effects in multivariate settings. We showed that the choice of test statistic $T$ significantly influences algorithm performance, with different choices yielding varying effectiveness. Notably, the learning algorithm that controls noise variance exhibits stronger theoretical guarantees and improved performance in simulations, even in nonlinear settings. Our results highlight that performance depends on noise matrix assumptions, particularly as dimensionality $d$ increases, leading to better discriminative power of intervention distributions. Furthermore, the loss function serves as a statistic in CI tests, allowing us to assess whether $X$ significantly affects $Y$ while enhancing interpretability. Our approach ensures robustness in multivariate settings and enables extensions to other CI test statistics and regression models, fostering broader applicability. Future work will derive the distribution of the optimal learning loss under null and alternative hypotheses to enhance test power. We will also explore nonlinear representations via projection into a reproducing kernel Hilbert space and assess cases where the effects of $X$ and $Z$ on $Y$ are not linearly separable. Additionally, we aim to further investigate this problem from an information-geometric perspective.

### Author Contributions

Gherardo Varando and Homer Durand conceived the original idea of Direct Effect Analysis, generalising an earlier concept introduced by Gustau Camps-Valls and Gherardo Varando. Durand and Varando developed the methodology and algorithmic framework, with Durand deriving the theoretical guarantees. Durand implemented the code and conducted the experiments. Gustau Camps-Valls supervised the project and provided critical guidance. All authors collaboratively wrote and edited the manuscript.

### Acknowledgements

Authors acknowledge funding from the Horizon projects AI4PEX (grant agreement 101137682) and ThinkingEarth (grant agreement 101130544), the European Research Council (ERC) support under the ERC Synergy Grant USMILE (grant agreement 855187).

### References

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, 1958.

T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 2003.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Martin Bilodeau and David Brenner. Theory of multivariate statistics. In *Theory of multivariate statistics*, 1999. URL `https://api.semanticscholar.org/CorpusID:117047324`.

Gregory C Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, pages 591–605, 1960.

Gokhan Danabasoglu, J-F Lamarque, J Bacmeister, DA Bailey, AK DuVivier, Jim Edwards, LK Emmons, John Fasullo, R Garcia, Andrew Gettelman, et al. The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, 12(2): e2019MS001916, 2020.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 489–496, 2008.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Kernel dimension reduction in regression. 2009.

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

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Klaus Hasselmann. Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate*, 6 (10):1957–1971, 1993.

Ronald R Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, pages 1–49, 1976.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pages 162–190. Springer, 1992.

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. 2013.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8 (3), 2007.

Steven M Kay. Fundamentals of statistical signal processing: Estimation theory, 1993.

Steven M. Kay. *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*. Prentice Hall, 1998.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

Judith L Lean and David H Rind. How natural and anthropogenic influences alter global and regional surface temperatures: 1889 to 2006. *Geophysical Research Letters*, 35(18), 2008.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

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. *arXiv preprint arXiv:2206.06169*, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *International Conference on Machine Learning*, pages 4114–4124, 2019.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Rich Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

Neil A Macmillan. Signal detection theory. *Stevens' handbook of experimental psychology: Methodology in experimental psychology*, 4:43–90, 2002.

R. V. Mises and H. Pollaczek-Geiringer. Praktische verfahren der gleichungsauflösung . *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 9(2):152–164, 1929. doi: https://doi.org/10.1002/zamm.19290090206. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/zamm.19290090206`.

Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. Methods and tools for causal discovery and causal inference. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 12(2):e1449, 2022.

J Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014.

Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. 2022.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78 (5):947–1012, 10 2016. ISSN 1369-7412.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

B. Raja Rao. Partial canonical correlations. *Trabajos de estadistica y de investigacion operativa*, 20(2):211–219, Jun 1969. ISSN 0041-0241. doi: 10.1007/BF03028532. URL `https://doi.org/10.1007/BF03028532`.

Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 5381–5390, 2018.

Sorawit Saengkyongam, Elan Rosenfeld, Pradeep Ravikumar, Niklas Pfister, and Jonas Peters. Identifying representations for intervention extrapolation. *arXiv preprint arXiv:2310.04295*, 2023.

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.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalized covariance measure. *The Annals of Statistics*, 46(3):1516–1553, 2018a.

Rajen Dinesh Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 2018b. URL `https://api.semanticscholar.org/CorpusID:88516900`.

Sebastian Sippel, Nicolai Meinshausen, Anna Merrifield, Flavio Lehner, Angeline G. Pendergrass, Erich Fischer, and Reto Knutti. Uncovering the forced climate response from a single ensemble member using statistical learning. *Journal of Climate*, 32(17):5677–5699, 2019.

MTCAJ Thomas and A Thomas Joy. *Elements of information theory*. Wiley-Interscience, 2006.

Gherardo Varando, Miguel-Ángel Fernández-Torres, Jordi Muñoz-Marí, and Gustau Camps-Valls. Learning causal representations with granger pca. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

Xinghao Yang, Weifeng Liu, Wei Liu, and Dacheng Tao. A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2349–2368, 2021.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31, 2018.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 804–813, 2011.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 804–813, 2012.

# Supplementary Material

**Homer Durand**[1] **Gherardo Varando**[2] **Gustau Camps-Valls**[1]

[1]Image Processing Lab, University of Valencia, Valencia, Spain
[2]Department of Statistics and Operational Research, University of Valencia, Valencia, Spain

## CONTENTS

**1 Introduction** **1**

**2 Preliminary** **2**

2.1 Introductory example . . . . . 2

2.2 Related work . . . . . 3

**3 Learning Framework** **4**

3.1 Maximisation of a CIT statistic . . . . . 4

3.2 Empirical estimators . . . . . 5

**4 Theoretical guarantees** **5**

4.1 Causal effect representation . . . . . 6

4.2 Testing the presence of a direct effect . . . . . 7

**5 Experiments** **7**

5.1 Simulation experiments . . . . . 7

5.2 Real-world experiments . . . . . 8

**6 Conclusion** **9**

**A Further methodological details** **14**

A.1 An information theory perspective on nested models test maximisation . . . . . 14

A.2 Partial correlation analysis as the maximisation of a conditional independence test statistic . . . . . 14

A.3 Empirical estimators . . . . . 15

A.4 Noise term behavior . . . . . 15

**B Proofs** **16**

B.1 Auxiliary lemma . . . . . 16
B.2 Signal-to-Noise optimality . . . . . 17
B.3 Noise term behavior . . . . . 17
B.4 Equivalence of Signal-to-Noise ratio and Fisher information . . . . . 18
B.5 Distribution of leading eigenvalues under conditional independence hypothesis . . . . . 19
B.6 Convergence rates . . . . . 20

**C Experiments** **22**
C.1 Simulation experiments . . . . . 22
C.2 Real-World experiments . . . . . 27

# A FURTHER METHODOLOGICAL DETAILS

We provide some further details and intuitions about our approach, starting by giving some insights it has with information theory.

## A.1 AN INFORMATION THEORY PERSPECTIVE ON NESTED MODELS TEST MAXIMISATION

In an information-theoretic framework (Thomas and Joy, 2006), the nested models residuals (3) can be interpreted as measures of uncertainty—quantified by entropy—regarding $Y$ given $X$ and $Z$. The proposed statistic then aims to maximise the conditional mutual information

$$I(X; \mathbf{w}^\top Y \mid Z) = H(\mathbf{w}^\top Y \mid Z) - H(\mathbf{w}^\top Y \mid Z, X),$$

where $H(\mathbf{w}^\top Y \mid Z)$ and $H(\mathbf{w}^\top Y \mid Z, X)$ denote the corresponding conditional entropies. This aligns with the well-established connection between conditional mutual information, causality, and conditional independence (Janzing et al., 2013).

Moreover, Proposition 4.4 establishes that, in the linear case with information bottleneck, maximising the signal-to-noise ratio (SNR) is equivalent to maximizing Fisher information. This implies that the proposed algorithm optimally distinguishes between the interventional distributions $p(Y \mid do(X = x))$ and $p(Y \mid do(X = x + \delta x))$, improving their separability under small interventions. This follows from the well established connection between Fisher information and the Kullback–Leibler divergence.

**Proposition A.1.** *Let $P(Y \mid x)$ be a probability distribution over $Y$ parameterised by $x \in \mathbb{R}^d$. Consider a small perturbation $\delta x$ such that $P(Y \mid x + \delta x)$ remains close to $P(Y \mid x)$. Then, the Kullback–Leibler divergence between these two distributions admits the following second-order expansion:*

$$D_{\mathrm{KL}}(P(Y \mid x) \,\|\, P(Y \mid x + \delta x)) = \frac{1}{2} \delta x^\top I(x) \delta x + O(\|\delta x\|^3),$$

*where $I(x)$ is the* Fisher information matrix*, given by*

$$I_{\mathbf{w}}(x) = \mathbb{E}\left[U(x) U(x)^\top\right],$$

*with $U(x) = \nabla_x \log P(\mathbf{w}^\top Y \mid X = x)$ denoting the score function.*

The proof is provided in Appendix B.4.

This result formalises the intuition that our algorithm identifies a subspace that maximally separates distributions under infinitesimal intervention perturbations, enhancing their distinguishability.

## A.2 PARTIAL CORRELATION ANALYSIS AS THE MAXIMISATION OF A CONDITIONAL INDEPENDENCE TEST STATISTIC

We briefly outline how the partial Canonical Correlation Analysis (CCA) test, originally introduced by Rao (1969), can be interpreted within our framework. Specifically, we show that it can be viewed as the maximisation of a partial correlation test between $\mathbf{w}^\top Y$ and $X$ when adjusted for $Z$.

Let the population residuals after regressing out $Z$ be defined as:

$$R_x(\mathbf{v}) = \mathbf{v}^\top X - \mathbb{E}[\mathbf{v}^\top X \mid Z],$$
$$R_y(\mathbf{w}) = \mathbf{w}^\top Y - \mathbb{E}[\mathbf{w}^\top Y \mid Z].$$

Assuming a linear relationship between $X$ and $Y$, the conditional independence statistic can be expressed as:

$$T_{\mathrm{C}}(X, Y, Z; \mathbf{w}, \mathbf{v}) = \operatorname{artanh}(\operatorname{corr}(R_x(\mathbf{v}), R_y(\mathbf{w}))). \tag{9}$$

Under the null hypothesis of conditional independence, and assuming that $R_x$ and $R_y$ are linearly related and follow a Gaussian distribution, it can be shown that $T_C$ is asymptotically normally distributed. Since the artanh function is monotonic, maximising the CIT statistic is equivalent to maximizing the partial correlation test statistic.

It also share similarities to the statistic proposed in (Shah and Peters, 2018a), with the key difference being the normalisation used. While Shah and Peters (2018a) proposed a statistic based on the covariance of the residuals normalised by the variance of their product, we normalise by the product of the variances of the residuals, resulting in a correlation coefficient. Although the approach in (Shah and Peters, 2018a) is known to have power against alternatives under very weak assumptions—specifically, that the convergence rate of the estimators of the conditional expectations results in an error product rate of $o(n^{-1})$—it is less straightforward to derive explicit formulations for $\mathbf{v}$ and $\mathbf{w}$ from their method.

**Empirical estimator for partial CCA** We are given two unbiased estimators $\hat{f}_x(Z)$ and $\hat{f}_y(Z)$ of respectively $\mathbb{E}[X|Z]$ and $\mathbb{E}[Y|Z]$. We denote by $\hat{\mathbf{R}}_x$ and $\hat{\mathbf{R}}_y$ the (empirical) residuals obtained from the predictions of $\mathbf{X}$ and $\mathbf{Y}$, respectively $\hat{\mathbf{R}}_x = \mathbf{X} - \hat{f}_x(\mathbf{Z})$ and $\hat{\mathbf{R}}_y = \mathbf{Y} - \hat{f}_y(\mathbf{Z})$. Similarly, the maximisation of the loss can be obtained via a generalised eigenvalue decomposition

$$\hat{\mathbf{\Sigma}}_{\mathbf{R}_x\mathbf{R}_y}^{\top} \hat{\mathbf{\Sigma}}_{\mathbf{R}_y}^{-1} \hat{\mathbf{\Sigma}}_{\mathbf{R}_x\mathbf{R}_y} \mathbf{w} = \lambda \hat{\mathbf{\Sigma}}_{\mathbf{R}_y} \mathbf{w}, \tag{10}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{R}_x\mathbf{R}_y}$ is the sample covariance of $\hat{\mathbf{R}}_x$ and $\hat{\mathbf{R}}_y$, and $\hat{\mathbf{\Sigma}}_{\mathbf{R}_y}$ is the sample covariance of $\hat{\mathbf{R}}_y$.

### A.3 EMPIRICAL ESTIMATORS

We provide additional details regarding our estimators, focusing on three key aspects: the estimation of conditional expectations, the extraction of multiple components, and the stability of the solutions obtained through the Generalised Eigenvalue (GEV) problem.

**Estimation of the Conditional Expectation** We have, so far, assumed the availability of unbiased estimators for the conditional expectations $\mathbb{E}[Y|X, Z]$ and $\mathbb{E}[Y|Z]$. In practice, these estimators should be selected based on domain-specific knowledge. In our case, we use the OLS estimator for the linear case and random forests for the nonlinear case.

**Further Components** Until now, we have considered the case where $q = 1$, assuming that the dimensionality of the direct effect of $X$ on $Y$ is rank one. In a manner analogous to the power iteration method (Mises and Pollaczek-Geiringer, 1929), we can extract additional components by employing a deflation technique.

**Stability of the Solution** The stability of the solutions is influenced by the covariance matrices $\hat{\mathbf{\Sigma}}_{\text{full}}$ and $\hat{\mathbf{\Sigma}}_{R_y}$, which may be ill-conditioned due to the characteristics of the noise term $N_Y$. This can complicate the GEV resolution. To mitigate this issue, we use a regularisation strategy that modifies the covariance matrices as $\hat{\mathbf{\Sigma}}_{\text{full}} + \delta\mathbf{I}$, and $\hat{\mathbf{\Sigma}}_{R_y} + \delta\mathbf{I}$, where $\delta$ is a small constant (typically $10^{-8}$) that stabilises the smallest eigenvalues.

Optimising the regularisation parameter more effectively might be crucial in the context of high-dimensional response variables. A promising approach could be the Ledoit-Wolf regularisation strategy, as proposed by Ledoit and Wolf (2004).

**Complexity** The computational complexity of the Direct Effect Analysis algorithm scales with the number of components ($K$), the number of samples ($n$), and the dimensionality of the variables ($d$). Let the training complexities of the conditional expectation models $g_{\text{res}}$ and $g_{\text{full}}$ be $\mathcal{O}(w_{\text{res}}(n, d, r))$ and $\mathcal{O}(w_{\text{full}}(n, d, r))$, respectively. Then, the overall complexity is

$$\mathcal{O}\left(K\left(w_{\text{res}}(n, d, r) + w_{\text{full}}(n, d, r) + nd^2 + Kd^3\right)\right).$$

Here, the first two terms account for training the conditional expectation estimators, the third term corresponds to computing the residual covariance matrix, and the final term arises from solving the generalized eigenvalue (GEV) problem.

### A.4 NOISE TERM BEHAVIOR

The conditions outlined in Proposition 4.3 may initially seem complex, so we provide a more intuitive explanation here. In many practical scenarios, improving the signal-to-noise ratio becomes crucial as the dimensionality of $Y$ increases, which can occur when enhancing image resolution or adding sensors. Higher-dimensional data provides a richer representation of the system, enabling better separation of signal and noise and improving inference accuracy.

**Example.** In climate science, we analyze global temperature patterns using climate observations. Let $Y$ represent the observed temperature field and $N_Y$ represent the observational errors arising from sensor limitations or model imperfections. The function $\phi(X)$ may capture internal climate variability (e.g., El Niño), while $\psi(Z)$ represents external forcing (e.g., greenhouse gas emissions). Increasing data granularity—through higher-resolution climate models, more sensors, or longer historical records—enhances the detectability of systematic climate responses while averaging out transient noise. As a result, the signal-to-noise ratio improves, making it easier to discern causal relationships and understand climate drivers.

The key insight is that algorithm performance depends on the structure of $\mathbf{b}$ and its interaction with noise terms. Performance improves when signal variance increases ($\|\mathbf{b}\|^2$ grows with $d$) in directions where noise covariances $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{\psi(Z)}$ are small. We can for example think about the simple case where the eigenvalue of the covariance matrix $\mathbf{\Sigma}_N$ decay quadractically as $d$ increases and where $\mathbf{b} = (1, \ldots, 1)$. The estimator $T_D$ is optimal in that it maximises the signal-to-noise ratio under mild conditions (see Proposition 4.1). Convergence issues arise only in rare cases where $\mathbf{b}$, $\mathbf{\Sigma}$, and $\mathbf{\Sigma}_{\psi(Z)}$ decay at similar rates, as illustrated in Figure 2.

While these guarantees hold in the idealised population setting with infinite data, real-world applications often involve limited samples. In such cases, the theoretical insights may not directly translate into robust performance, necessitating regularisation techniques to prevent overfitting and improve estimation reliability in small datasets.

# B PROOFS

We now provide proof of our main theoretical results. As this will be useful for most of the theoretical development, we first get a result for the first eigenvector of each optimisation problem.

## B.1 AUXILIARY LEMMA

**Lemma B.1.** *Let $\mathbf{w}_S$, $\mathbf{w}_F$, and $\mathbf{w}_D$ denote the first eigenvectors associated with the optimisation problems in Eq.* (3)*, Eq.* (4)*, and Eq.* (5)*, respectively. The following properties hold:*

1. *The eigenvector $\mathbf{w}_S$ is proportional $\mathbf{b}$, i.e., $\mathbf{w}_S \propto \mathbf{b}$, when maximizing Eq.* (3)*. In the linear case it corresponds to the Gradient EDE.*
2. *The eigenvector $\mathbf{w}_F$ is proportional to the direction of the inverse covariance-weighted true causal effect, i.e., $\mathbf{w}_F \propto \mathbf{\Sigma}^{-1}\mathbf{b}$, when maximizing Eq.* (4)*.*
3. *The eigenvector $\mathbf{w}_D$ is proportional to the inverse of the sum of the covariances of the noise and confounding variables, i.e., $\mathbf{w}_D \propto (\mathbf{\Sigma} + \mathbf{\Sigma}_{\psi(Z)})^{-1}\mathbf{b}$, when maximizing Eq.* (5)*.*

*Proof.* Recall the definitions:

$$
\begin{aligned}
R^2_{\text{full}}(\mathbf{w}) &= \mathbb{E}[(\mathbf{w}^\top Y - \mathbb{E}[\mathbf{w}^\top Y | X, Z])^2], \\
R^2_{\text{res}}(\mathbf{w}) &= \mathbb{E}[(\mathbf{w}^\top Y - \mathbb{E}[\mathbf{w}^\top Y | Z])^2], \\
R^2_{\text{noise}}(\mathbf{w}) &= \mathbb{E}[(\mathbf{w}^\top Y - \mathbb{E}[\mathbf{w}^\top Y \mid X, Z = 0])^2].
\end{aligned}
$$

From the model $Y^x = \mathbf{b}\phi(x) + \psi(Z) + N_y$, we derive:

$$
\begin{aligned}
R^2_{\text{full}}(\mathbf{w}) &= \mathbf{w}^\top \mathbf{\Sigma}\mathbf{w}, \\
R^2_{\text{res}}(\mathbf{w}) &= \mathbf{w}^\top \mathbf{\Sigma}\mathbf{w} + \phi(x)^2 \mathbf{w}^\top \mathbf{b}\mathbf{b}^\top \mathbf{w}, \\
R^2_{\text{noise}}(\mathbf{w}) &= \mathbf{w}^\top \mathbf{\Sigma}\mathbf{w} + \mathbf{w}^\top \mathbf{\Sigma}_{\psi(Z)}\mathbf{w}.
\end{aligned}
$$

The difference between the residual and full terms is:

$$
R^2_{\text{res}}(\mathbf{w}) - R^2_{\text{full}}(\mathbf{w}) = \phi(x)^2 \mathbf{w}^\top \mathbf{b}\mathbf{b}^\top \mathbf{w}.
$$

Each optimisation problem for $T_S$, $T_F$, and $T_D$ corresponds to a generalised eigenvalue problem of the form $\mathbf{N}^{-1}\mathbf{M}$, where $\mathbf{M} = \phi(x)^2 \mathbf{b}\mathbf{b}^\top$ is rank-1. Therefore, the first eigenvector $\mathbf{w}_1$ is proportional to $\mathbf{N}^{-1}\mathbf{b}$. The optimal solutions are:

1. $\mathbf{w}_S \propto \mathbf{b}$ for Eq. (3),
2. $\mathbf{w}_F \propto \mathbf{\Sigma}^{-1}\mathbf{b}$ for Eq. (4),
3. $\mathbf{w}_D \propto (\mathbf{\Sigma} + \mathbf{\Sigma}_{\psi(Z)})^{-1}\mathbf{b}$ for Eq. (5).

□

### B.2 SIGNAL-TO-NOISE OPTIMALITY

**Proposition B.2** (General optimality)**.** *Assuming $P$ is entailed in the SCM in Eq.* (6)*, we have that $\mathbf{w}_D$ is optimal.*

*Proof.* Recall that the optimal detector loss is defined as:

$$T_D = \frac{R^2_{\text{res}}(\mathbf{w}) - R^2_{\text{full}}(\mathbf{w})}{R^2_{\text{noise}}(\mathbf{w})} = \frac{\phi(x)^2\mathbf{w}^\top\mathbf{b}\mathbf{b}^\top\mathbf{w}}{\mathbf{w}^\top\mathbf{\Sigma}\mathbf{w} + \mathbf{w}^\top\mathbf{\Sigma}_{\psi(Z)}\mathbf{w}}.$$

Using the results from Theorem B.1, the signal-to-noise ratio is given by:

$$\gamma^2(\mathbf{w}) = \frac{(\mathbf{w}^\top S(x))^2}{\mathbf{w}^\top\mathbf{\Sigma}_N\mathbf{w}} = \phi(x)^2\frac{\mathbf{w}^\top\mathbf{b}\mathbf{b}^\top\mathbf{w}}{\mathbf{w}^\top(\mathbf{\Sigma} + \mathbf{\Sigma}_{\psi(Z)})\mathbf{w}},$$

where we note that $\psi(Z)$ and $N_y$ are assumed to be independent.

This completes the proof. □

**Proposition B.3** (Optimality under isotropic noise)**.** *Assuming that $P$ is entailed in the SCM in Eq.* (6) *and that $\mathbf{\Sigma}_N$ is isotropic, we have that both $\mathbf{w}_S$ and $\mathbf{w}_D$ are optimal. Moreover, if $\mathbf{\Sigma}_y$ is also isotropic, then $\mathbf{w}_F$ is also optimal.*

*Proof.* The proof follows straightforwardly from the assumption that $\mathbf{\Sigma}_N = \mathbf{\Sigma}_{\phi(z)} + \mathbf{\Sigma}$ is isotropic. Under this assumption, the constraint $\mathbf{w}^\top(\mathbf{\Sigma}_{\phi(z)} + \mathbf{\Sigma})\mathbf{w}$ is equivalent to the constraint $\|\mathbf{w}\| = 1$. Therefore, the loss function $T_D$ simplifies to the form of $T_S$. As a consequence, the optimality of $\mathbf{w}_D$ stated in Prop. 4.1 implies the optimality of $\mathbf{w}_S$.

Similarly, when both $\mathbf{\Sigma}_N$ and $\mathbf{\Sigma}_y$ are isotropic, we observe that $T_F$ becomes equivalent to $T_S$. Since it has been established that if $\mathbf{\Sigma}_N$ is isotropic, $\mathbf{w}_S$ is optimal, we conclude that $\mathbf{w}_F$ is also optimal. □

### B.3 NOISE TERM BEHAVIOR

**Proposition B.4** (Noise Term Behavior)**.** *Let $\|\mathbf{b}\|^2 = o(\nu_1(d))$, $\mathbf{b}^\top(\mathbf{\Sigma} + \mathbf{\Sigma}_{\psi(z)})\mathbf{b} = o(\nu_2(d))$, $\mathbf{b}^\top\mathbf{\Sigma}^{-1}\mathbf{b} = o(\nu_3(d))$, $\mathbf{b}^\top(\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\psi(\mathbf{Z})}\mathbf{\Sigma}^{-1})\mathbf{b} = o(\nu_4(d))$, and $\mathbf{b}^\top(\mathbf{\Sigma} + \mathbf{\Sigma}_{\psi(z)})^{-1}\mathbf{b} = o(\nu_5(d))$.*

*Assume that $\phi(X) \in \mathbb{R}$, the distribution $P$ follows the structural causal model in 6, and the following conditions hold:*

1. $\lim_{d\to\infty} \frac{\nu_1(d)}{\nu_2(d)} \to \infty$,
2. $\lim_{d\to\infty} \frac{\nu_3^2(d)}{\nu_4(d)} \to \infty$,
3. $\lim_{d\to\infty} \nu_5(d) \to \infty$.

*Under these conditions, the following convergence properties hold:*

1. $\gamma^2(\mathbf{w}_S) \to \infty$ *if condition 1) holds,*
2. $\gamma^2(\mathbf{w}_F) \to \infty$ *if condition 2) holds,*
3. $\gamma^2(\mathbf{w}_D) \to \infty$ *if condition 1), 2), or 3) holds.*

*Proof.* Recalling that we have

$$R^2_{\text{full}}(\mathbf{w}) = \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} \tag{11}$$

$$R^2_{\text{res}}(\mathbf{w}) - R^2_{\text{full}}(\mathbf{w}) = \phi(x)^2 \mathbf{w}^\top \mathbf{b}\mathbf{b}^\top \mathbf{w} \tag{12}$$

$$R^2_{\text{noise}} = \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} + \mathbf{w}^\top \mathbf{\Sigma}_{\psi(Z)} \mathbf{w} \tag{13}$$

Substituting these into the signal-to-noise ratio in Eq. (7), we obtain:

$$\gamma^2(\mathbf{w}_S) = \frac{\|\mathbf{b}\|^2 \phi(x)^2}{\mathbf{b}^\top \mathbf{\Sigma} \mathbf{b} + \mathbf{b}^\top \mathbf{\Sigma}_{\psi(Z)} \mathbf{b}}, \tag{14}$$

$$\gamma^2(\mathbf{w}_F) = \frac{(\mathbf{b}^\top \mathbf{\Sigma}^{-1} \mathbf{b})^2 \phi(x)^2}{\mathbf{b}^\top \mathbf{\Sigma}^{-1} \mathbf{b} + \mathbf{b}^\top \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_{\psi(Z)} \mathbf{\Sigma}^{-1} \mathbf{b}}, \tag{15}$$

$$\gamma^2(\mathbf{w}_D) = \mathbf{b}^\top (\mathbf{\Sigma} + \mathbf{\Sigma}_{\psi(Z)})^{-1} \mathbf{b} \phi(x)^2. \tag{16}$$

The convergence properties of the signal-to-noise ratio follow directly from these formulations, assuming $\phi(x)$ is bounded. Since $\mathbf{w}_D$ maximises the SNR, if either $\mathbf{w}_S$ or $\mathbf{w}_F$ has an SNR that grows to infinity, then the SNR of $\mathbf{w}_D$ will also tend to infinity. □

### B.4 EQUIVALENCE OF SIGNAL-TO-NOISE RATIO AND FISHER INFORMATION

**Proposition B.5** (Equivalence between Fisher Information and SNR)**.** *Consider a SCM as described in* (6)*, and let the intervention function be $\phi(x) = \mathbf{v}^\top x$, where $\mathbf{v} \in \mathbb{R}^d$. Then, the SNR is proportional to the Fisher Information of the intervention, i.e. $I_{\mathbf{w}}(x) = \alpha \gamma^2(\mathbf{w})$ with $\alpha \in \mathbb{R}^+$.*

*Proof.* Let $\mathbf{w}^\top Y^x \sim \mathcal{N}(\mathbf{w}^\top \mu(x), \mathbf{w}^\top \mathbf{\Sigma}_N \mathbf{w})$ denote the distribution of $\mathbf{w}^\top Y^x$.

The log-likelihood for the intervention is given by:

$$\log p(Y \mid do(X = x)) = C - \frac{1}{2} \mathbf{w}^\top (Y - \mu(x))^\top (\mathbf{w}^\top (\mathbf{\Sigma}_{\psi(z)} + \mathbf{\Sigma}) \mathbf{w})^{-1} (Y - \mu(x)) \mathbf{w},$$

where $C$ is a constant relative to $x$.

The *informant* $U(x)$ is the derivative of the log-likelihood with respect to $x$:

$$\frac{\partial}{\partial x} \log p(Y \mid do(X = x)) = \frac{1}{2} \mu'(x)^\top (\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w})^{-1} (Y - \mu(x)) \mathbf{w}.$$

The Fisher information $I_{\mathbf{w}}(x)$ is the variance of the informant. Since the informant at the maximum likelihood has mean zero (see Lehmann and Casella, 2006, section 6), we write:

$$\begin{aligned} I_{\mathbf{w}}(x) &= \mathbb{E}[U(x) U(x)^\top] \\ &= \mathbb{E}\left[\mu'(x)^\top (\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w})^{-1} (Y - \mu(x)) \mathbf{w}\mathbf{w}^\top (Y - \mu(x))^\top (\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w})^{-1} \mu'(x)\right]. \end{aligned}$$

Using the fact that $\mathbb{E}[(Y - \mu(x))(Y - \mu(x))^\top] = \mathbf{\Sigma}$, we obtain:

$$I_{\mathbf{w}}(x) = \mathbf{w}^\top \mu'(x) (\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w})^{-1} \mu'(x) \mathbf{w}.$$

Since $\mu(x) = \mathbf{b}\mathbf{v}^\top x$, we have $\mu'(x) = \mathbf{b}\mathbf{v}$. Additionally, $\mathbf{\Sigma} = \mathbf{\Sigma}_{\psi(z)} + \mathbf{\Sigma}$. Substituting these expressions into the Fisher information formula, we get:

$$\begin{aligned} I_{\mathbf{w}}(x) &= \mathbf{w}^\top \mathbf{b}\mathbf{v}^\top (\mathbf{w}^\top (\mathbf{\Sigma}_{\psi(z)} + \mathbf{\Sigma}) \mathbf{w})^{-1} \mathbf{v}\mathbf{b}^\top \mathbf{w} \\ &= \frac{\mathbf{w}^\top \mathbf{b}\mathbf{v}^\top \mathbf{v}\mathbf{b}^\top \mathbf{w}}{\mathbf{w}^\top (\mathbf{\Sigma}_{\psi(z)} + \mathbf{\Sigma}) \mathbf{w}} \\ &= \frac{\|\mathbf{v}\|_2^2}{\phi(x)^2} \gamma^2(\mathbf{w}). \end{aligned}$$

□

**Proposition B.6.** *Let $P(Y \mid x)$ be a probability distribution over $Y$ parameterised by $x \in \mathbb{R}^d$. Consider a small perturbation $\delta x$ such that $P(Y \mid x + \delta x)$ remains close to $P(Y \mid x)$. Then, the Kullback–Leibler divergence between these two distributions admits the following second-order expansion:*

$$D_{\mathrm{KL}}(P(Y \mid x) \,\|\, P(Y \mid x + \delta x)) = \frac{1}{2}\delta x^\top I(x)\delta x + O(\|\delta x\|^3),$$

*where $I(x)$ is the* Fisher information matrix*, given by:*

$$I_{\mathbf{w}}(x) = \mathbb{E}\left[U(x)U(x)^\top\right].$$

*With $U(x) = \nabla_x \log P(\mathbf{w}^\top Y \mid X = x)$ the informant (or score) function.*

*Proof sketch.* Assuming $P(Y \mid x)$ is smooth in $x$, we approximate it to its second order Taylor expansion:

$$\log P(Y \mid x + \delta x) = \log P(Y \mid x) + \delta x^\top \nabla_x \log P(Y \mid x) + \frac{1}{2}\delta x^\top \nabla_x^2 \log P(Y \mid x)\delta x + O(\|\delta x\|^3).$$

The KL divergence is defined as:

$$D_{\mathrm{KL}}(P(Y \mid x) \,\|\, P(Y \mid x + \delta x)) = \mathbb{E}_{Y \sim P(Y|x)}\left[\log \frac{P(Y \mid x)}{P(Y \mid x + \delta x)}\right].$$

Substituting into the KL divergence and using the property that $\mathbb{E}_{Y \sim P(Y|x)}[\nabla_x \log P(Y \mid x)] = 0$, the first-order term vanishes (see Lehmann and Casella, 2006, section 6), leaving:

$$D_{\mathrm{KL}}(P(Y \mid x) \,\|\, P(Y \mid x + \delta x)) = -\frac{1}{2}\mathbb{E}_{Y \sim P(Y|x)}\left[\delta x^\top \nabla_x^2 \log P(Y \mid x)\delta x\right] + O(\|\delta x\|^3).$$

Since the Fisher information matrix is defined as $I(x) = -\mathbb{E}[\nabla_x^2 \log P(Y \mid x)]$, we obtain:

$$D_{\mathrm{KL}}(P(Y \mid x) \,\|\, P(Y \mid x + \delta x)) = \frac{1}{2}\delta x^\top I(x)\delta x + O(\|\delta x\|^3).$$

□

## B.5 DISTRIBUTION OF LEADING EIGENVALUES UNDER CONDITIONAL INDEPENDENCE HYPOTHESIS

**Proposition B.7** (Distribution of $\lambda_F$ under conditional independence)**.** *Let the distribution $P$ be induced by the SCM in* (6) *with linear assignments and Gaussian noise, and assume $p = q = 1$. Under the null hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$, the largest root $\lambda_F$ is $F$-distributed such that $(dfn/dfd)\lambda_F \sim F(dfd, dfn)$ where $dfn = d$ and $dfd = n - p - r - 1$.*

*Proof sketch.* It can easily be shown that $\hat{R}^2_{\text{full}}$ and $\hat{R}^2_{\text{res}}$ follows $\chi^2$ distributions of respectively $d(n - p - r - 1)$ and $d(n - p - r - 1)$ degrees of freedom as they are computed as sums of squared Gaussian distributions. Their ratio can thus be shown to follow an F distribution with degrees of freedom $dfn = d$ and $dfd = n - p - r - 1$. As the weights related to $Z$ are frozen when getting $\hat{R}^2_{\text{noise}}$, we have that it follows a $\chi^2$ with $d(n - p - 1)$ degrees of freedom. Thus $\Lambda_D \sim F(p, n - p - 1)$. □

We refer reader to the distribution of Roy's largest root, the Chow test (Chow, 1960) or the generalised linear hypothesis test (see e.g. Anderson (1958) chapter 7) as similar problems have been widely studied in the multivariate statistics literature (Anderson, 2003; Bilodeau and Brenner, 1999).

**Proposition B.8** (Upper Bound on $\Lambda_D$ Under Conditional Independence)**.** *Under similar assumptions as in Prop 4.6 we have under the null hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$ that $P(\Lambda_D \geq \lambda_D | H_0) \leq P(\Lambda_F \geq \lambda_D | H_0)$.*

*Proof sketch.* As the conditioning set in the computation of root squared errors $R^2_{\text{noise}}$ is larger than of $R^2_{\text{full}}$, the empirical residuals $R^2_{\text{noise}}$ are always larger than $R^2_{\text{full}}$ thus we have that

$$\frac{R^2_{\text{res}} - R^2_{\text{full}}}{R^2_{\text{noise}}} \leq \frac{R^2_{\text{res}} - R^2_{\text{full}}}{R^2_{\text{full}}}.$$

Hence, we have for any $\lambda_D$ that $P(\Lambda_D \geq \lambda_D | H_0) \leq P(\Lambda_F \geq \lambda_D | H_0)$ under the null hypothesis $H_0 : X \perp\!\!\!\perp Y | Z$. □

Note that by using this upper bound we tend to lose power in the test procedure but we still control type I errors (we reject less than we would optimally do) and thus the test is valid. Further research should aim at discovering a better approximation for the distribution $\Lambda_D$.

### B.6 CONVERGENCE RATES

We first introduce an important theorem that will be useful for the proof.

**Theorem B.9** (Davis-Kahan theorem (Davis and Kahan, 1970))**.** *Let $\lambda^{(1)} - \lambda^{(2)} = \delta > 0$ where $\lambda^{(1)} > \lambda^{(2)} \geq \cdots \geq \lambda^{(d)}$ be the eigenvalues of $\mathbf{\Sigma}$ and $\hat{\lambda}^{(1)} - \hat{\lambda}^{(2)} = \hat{\delta} > 0$ where $\hat{\lambda}^{(1)} > \hat{\lambda}^{(2)} \geq \cdots \geq \hat{\lambda}^{(d)}$ be the eigenvalues of $\hat{\mathbf{\Sigma}}$ and let $\mathbf{W}$ and $\hat{\mathbf{W}}$ their corresponding eigenvectors. We have that*

$$\| \sin \Theta(\mathbf{W}, \hat{\mathbf{W}}) \|_{op} \leq \frac{\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{\mathbf{op}}}{\max_j(|\hat{\lambda}_{j-1} - \lambda_j|, |\hat{\lambda}_{j+1} - \lambda_j|)} \tag{17}$$

*where $\Theta$ is a distance between subspaces. Similarly, for any $j$ we have that $\|\hat{\mathbf{w}}_j - \mathbf{w}_j\| \leq \sqrt{2} \sin \Theta(\mathbf{w}_j, \hat{\mathbf{w}}_\mathbf{j})$.*

We show that under common assumptions, specifically that there are two unbiased estimators $\hat{g}_{\text{full}}$ and $\hat{g}_{\text{res}}$ with convergence rates $\kappa_1(n)$ and $\kappa_2(n)$, the estimators proposed in Eq. (3) are consistent with their population counterparts. Furthermore, we demonstrate that their convergence rate typically depends on the convergence rates $\kappa_1(n)$ and $\kappa_2(n)$.

**Proposition B.10** (Convergence Rate of F-Test Based Losses)**.** *Assume the following conditions hold:*

1. $\mathbb{E} \|\hat{g}_{full}(\mathbf{X}_i, \mathbf{Z}_i) - \mathbb{E}[\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{Z}_i]\|^2 = o_P(\kappa_1(n))$,
2. $\mathbb{E} \|\hat{g}_{res}(\mathbf{Z}_i) - \mathbb{E}[\mathbf{Y}_i \mid \mathbf{Z}_i]\|^2 = o_P(\kappa_2(n))$,
3. $\lambda_1^M - \lambda_2^M = \delta_M > 0$, *where* $\lambda_1^M > \lambda_2^M \geq \cdots \geq \lambda_d^M$ *are the eigenvalues of* $\mathbf{M}$,
4. $\lambda_1^N - \lambda_2^N = \delta_N > 0$, *where* $\lambda_1^N > \lambda_2^N \geq \cdots \geq \lambda_d^N$ *are the eigenvalues of* $\mathbf{N}$,
5. $\mathbb{E} \|Y - \mathbb{E}[Y \mid X, Z]\|^2 \leq N_{full}$ *and* $\mathbb{E} \|Y - \mathbb{E}[Y \mid Z]\|^2 \leq N_{res}$.

*Let $\mathbf{w}_1$ be the optimal solution to Eq.* (3)*, Eq.* (4)*, or Eq.* (5)*, and let $\hat{\mathbf{w}}$ be the empirical solution to their respective empirical estimators. Under the given conditions, we have the following convergence result:*

$$\mathbb{E}\left[\|\mathbf{w}_1 - \hat{\mathbf{w}}\|_2^2\right] = o\left(\sqrt{\kappa_1(n)} + \sqrt{\kappa_2(n)}\right). \tag{18}$$

*Proof.* Similar to what was done for the empirical estimators, the population loss can be written as an eigenvalue decomposition problem $\mathbf{N}^{-1}\mathbf{M}$, where $\mathbf{M} = \mathbf{\Sigma}_{\text{res}} - \mathbf{\Sigma}_{\text{full}}$, and $\mathbf{N}$ depends on the loss used. For simplicity, we consider $\mathbf{N} = \mathbf{I}$, which leads to the convergence result for the simple loss in Eq. (3). A similar reasoning can be applied to the convergence of the two other losses.

Let us first decompose $\hat{\mathbf{\Sigma}}_{\text{res}}$ as follows:

$$\hat{\mathbf{\Sigma}}_{\text{res}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y_i} - \hat{g}_{res}(\mathbf{Z}_i))(\mathbf{Y_i} - \hat{g}_{res}(\mathbf{Z}_i))^\top \tag{19}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{Y_i} - g_{res}(\mathbf{Z}_i) + g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i)\right)\left(\mathbf{Y_i} - g_{res}(\mathbf{Z}_i) + g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i)\right)^\top \tag{20}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{N}_{y,i}^\top\mathbf{N}_{y,i} + \frac{2}{n}\sum_{i=1}^{n}\mathbf{N}_{y,i}^\top\left(g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i)\right) + \frac{1}{n}\sum_{i=1}^{n}\left(g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i)\right)^\top\left(g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i)\right) \tag{21}$$

where $\mathbf{N}_{y,i}$ is the population residual (noise) of sample $i$.

We now aim to bound $\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_F$. Using the previous notation, we have:

$$\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_F \leq \|\mathbf{\Sigma} - \frac{1}{n}\sum_{i=1}^{n} \mathbf{N}_{y,i}^\top \mathbf{N}_{y,i}\|_F + \frac{2}{n}\sum_{i=1}^{n} \|\mathbf{N}_{y,i}^\top (g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i))\|_F \tag{22}$$

$$+ \|\frac{1}{n}\sum_{i=1}^{n} (g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i))^\top (g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i))\|_F \tag{23}$$

$$\leq A + B + C. \tag{24}$$

We first handle the term $C$:

$$\mathbb{E}[C] \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|(g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i))^\top (g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i))\|_F\right] \tag{25}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|(g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i))\|_2^2\right] \tag{26}$$

$$\leq C_1 \kappa_1(n) \qquad \text{by assumption } (1). \tag{27}$$

Next, for the term $B$:

$$\mathbb{E}[B] \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{N}_{y,i}^\top (g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i))\|_F\right] \tag{28}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{N}_{y,i}\|\right] \mathbb{E}\left[\|(g_{res}(\mathbf{Z}_i) - \hat{g}_{res}(\mathbf{Z}_i))\|_2\right] \tag{29}$$

$$\leq N_{res}\sqrt{C_1}\sqrt{\kappa_1(n)} \qquad \text{by assumptions } (1) \text{ and } (5). \tag{30}$$

For the term $A$, by the Strong Law of Large Numbers, there exists a constant $C_3$ such that:

$$\mathbb{E}[A] \leq \frac{C_3}{\sqrt{n}}.$$

Therefore, we obtain the bound:

$$\mathbb{E}\left[\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{op}\right] \leq N_{res}\sqrt{C_1}\sqrt{\kappa_1(n)} + C_1\kappa_1(n) + \frac{C_3}{\sqrt{n}}.$$

Similarly, an equivalent reasoning gives:

$$\mathbb{E}\left[\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{op}\right] \leq N_{full}\sqrt{C_2}\sqrt{\kappa_2(n)} + C_2\kappa_2(n) + \frac{C_4}{\sqrt{n}} \qquad \text{using assumptions } (2) \text{ and } (5).$$

Finally, applying the Davis-Kahan theorem, we have:

$$\mathbb{E}\left[\|\mathbf{w}_1 - \mathbf{w}_S\|_2^2\right] \leq \sqrt{2}\frac{\mathbb{E}\left[\|\mathbf{\Sigma}_{\text{res}} - \hat{\mathbf{\Sigma}}_{\text{res}}\|_F\right] + \mathbb{E}\left[\|\mathbf{\Sigma}_{\text{full}} - \hat{\mathbf{\Sigma}}_{\text{full}}\|_F\right]}{\delta_M} \tag{31}$$

$$\leq \sqrt{2}\frac{N_{full}\sqrt{C_2}\sqrt{\kappa_2(n)} + C_2\kappa_2(n) + \frac{C_4}{\sqrt{n}} + N_{res}\sqrt{C_1}\sqrt{\kappa_1(n)} + C_1\kappa_1(n) + \frac{C_3}{\sqrt{n}}}{\delta_M} \tag{32}$$

$$= o(\sqrt{\kappa_1(n)} + \sqrt{\kappa_2(n)}), \tag{33}$$

assuming that $\kappa_1(n)$ and $\kappa_2(n)$ decrease no faster than $o(1/n)$, which is typically the case for most of the regression algorithms. This conclude the proof. □

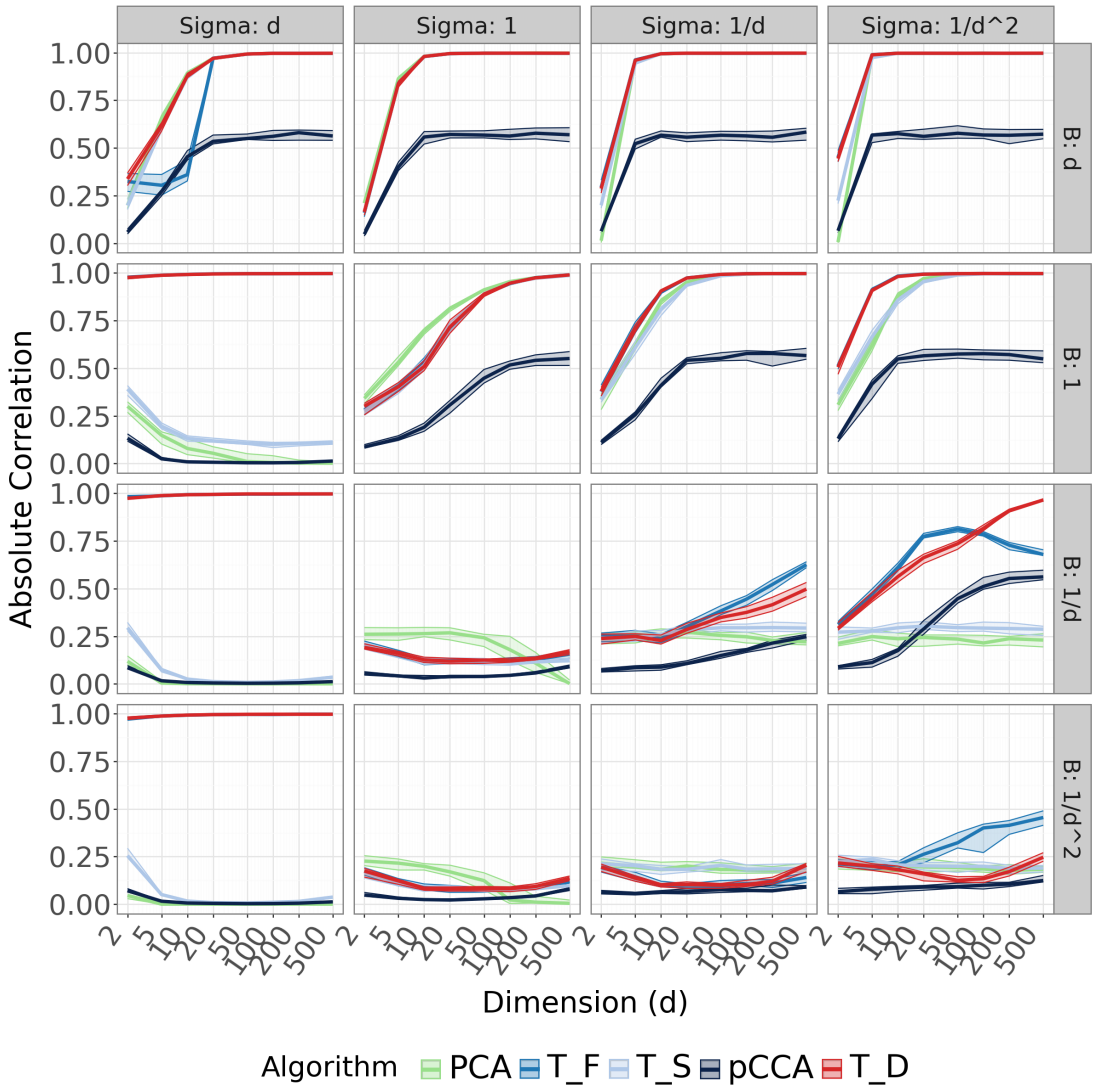



Figure 4: Correlation between $\mathbf{w}^\top Y$ and $\phi(X)$ as $d$ increases. $T_D$ consistently outperforms all methods, recovering $\phi(X)$ as $d$ grows, provided that $\mathbf{b}$ faster than $\mathbf{\Sigma}$. Columns are indexed by as A, B, C, D and rows by $1, 2, 3, 4$.

# C EXPERIMENTS

## C.1 SIMULATION EXPERIMENTS

The data are generated according to the following SCM:

$$
\begin{aligned}
N_x, N_z &\sim \mathcal{N}(0, \mathbf{I}), \\
N_y &\sim \mathcal{N}(0, \mathbf{\Sigma}), \\
Z &:= N_z, \\
X &:= f_a(\mathbf{C}^\top Z) + N_x, \\
Y &:= u\mathbf{b}^\top f_a(\Gamma^\top X) + v f_a(\mathbf{D}^\top Z) + w N_y.
\end{aligned} \tag{34}
$$

**Causal effect representation** Comparing how the different learning algorithms behave in different noise contexts seems relevant. Primarily, we can observe in the setting Strong_N_Y low_rank (Fig. 5) that the increase in performance using $T_F$

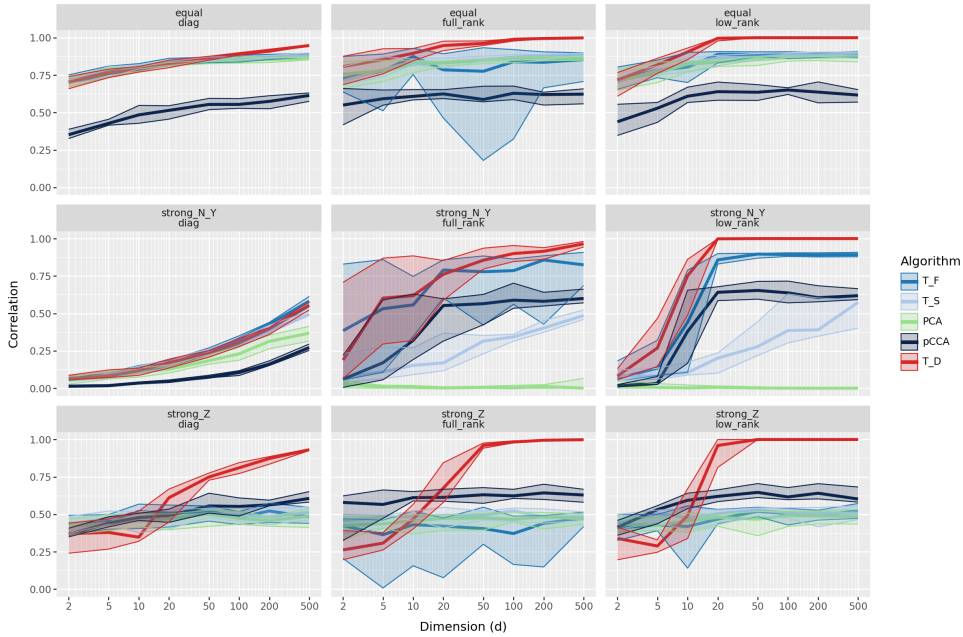



Figure 5: Experiments with different noise structure ($\mathbf{\Sigma}$ being diagonal, full rank and low rank) and scaling factors ($(u, v, w)$ as $(1/3, 1/3, 1/3)$, $(0.1, 0.1, 0.8)$, and $(0.1, 0.8, 0.1)$ for equal, Strong_N_Y and Strong_Z). Overall, learning algorithm $T_D$ performs better and tends to converge.

and $T_D$ is due to the low-rank structure of the noise. The overall better performance of $T_D$ over $T_F$ and pCCA is due to the correlation between $X$ and $Z$.

To better understand the behaviour of our algorithms, we conducted experiments in two additional settings:

1. **High-dimensional setting:** We conducted a similar experiment under different conditions on $\mathbf{b}$ and $\mathbf{\Sigma}$, increasing the dimensionality. In this case, however, we significantly reduced the sample size to $n = 100$, such that as $d$ grows, we obtain $n < d$.
2. **Nonlinear setting:** Again, we conducted a similar experiment with different conditions on $\mathbf{b}$ and $\mathbf{\Sigma}$. Still, here we applied a nonlinear mapping $f_a(z) := \exp(-z^2/2)\sin(az)$ with $a \in \{1, 2, 3\}$. We use a random forest algorithm with 100 trees as an estimator of the conditional expectation.
3. **$X$ independent of $Z$ setting:** We conducted experiments to clarify the discrepancy of performance between pCCA and $T_D$ by generating the data such that $X$ and $Y$ are independent.
4. **Non gaussian noises:** We conducted an experiment similar to the original experiment proposed in section 5 but considering uniform and exponential noises.

In the high-dimensional setting, as shown in Fig. 6, we observe results that are very similar to those in the large-sample setting, with one key difference: when $\mathbf{\Sigma}$ increases rapidly with $d$ (row 1), and $d > n$, the model performance drops significantly to near zero and when $\mathbf{b}$ is growing to slowly compared to $\sigma$ ($\sigma = [1, 1/d]$ and $\mathbf{b} = [1/d, 1/d^2]$). It would be interesting to further evaluate how the regularisation parameter (see Section A.3) might improve performance in this specific case. However, in this setting, ensuring algorithmic convergence is particularly challenging, as the signal strength is constrained by the number of available samples. As a result, the SNR is unlikely to grow unbounded unless the inherent noise in the data is minimal.

Interestingly, in the nonlinear setting, the learning algorithm $T_D$ is still able to recover $\phi(X)$ in most cases. In contrast, other algorithms show greater difficulty in achieving convergence under these conditions. This highlights the potential of learning algorithm $T_D$ to recover direct effects, even in complex nonlinear settings effectively.

We also conducted an experiment where $X$ and $Z$ were generated as independent variables to highlight the potential advantages of our learning algorithms. In this case, we observed that partial CCA (pCCA) could recover the latent structure

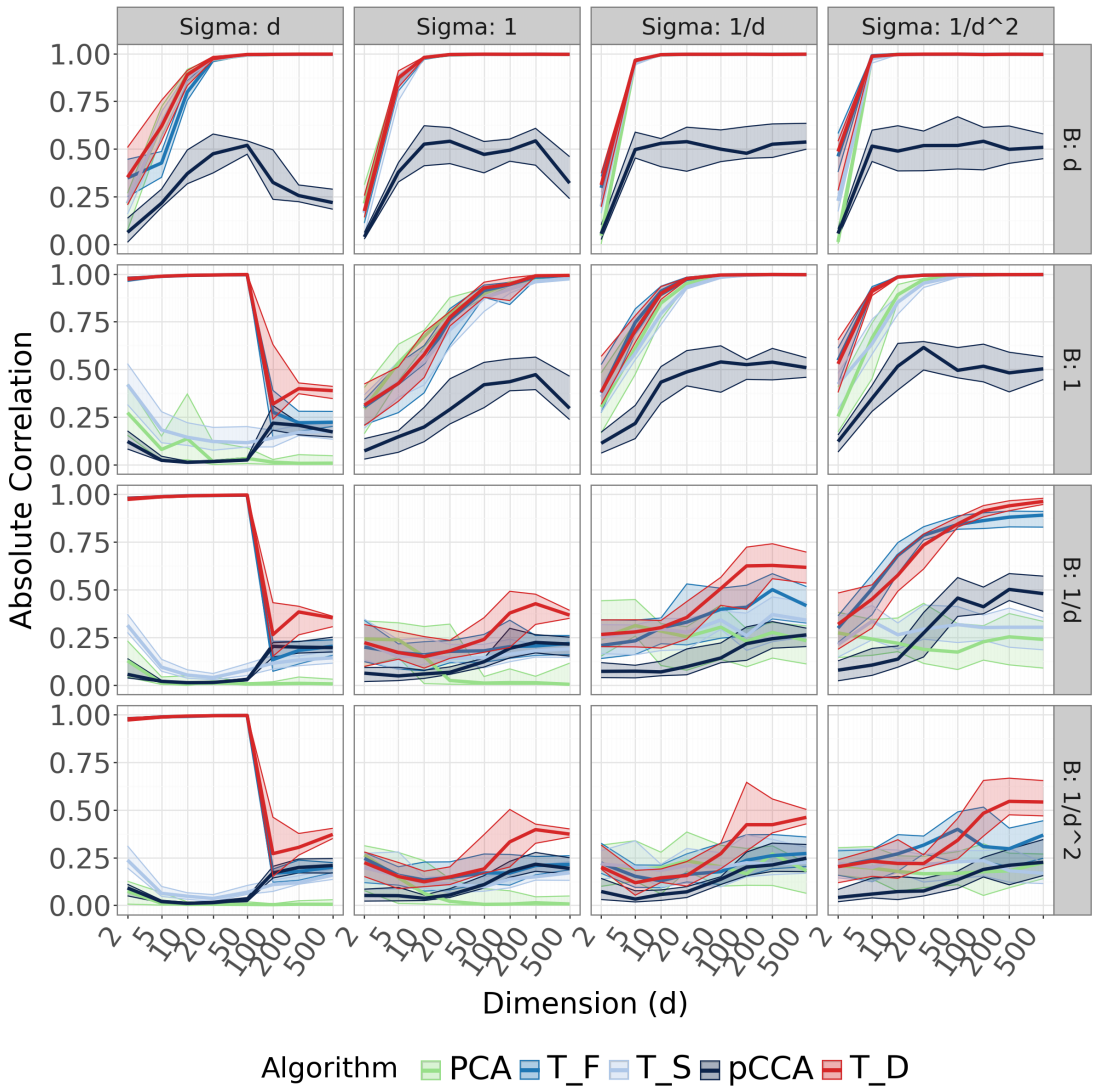



Figure 6: High dimensional experiment using 100 samples for training. As observed in the last row, all algorithms fail to recover the signal when the noise variance increases rapidly and the number of samples falls below the outcome dimension.

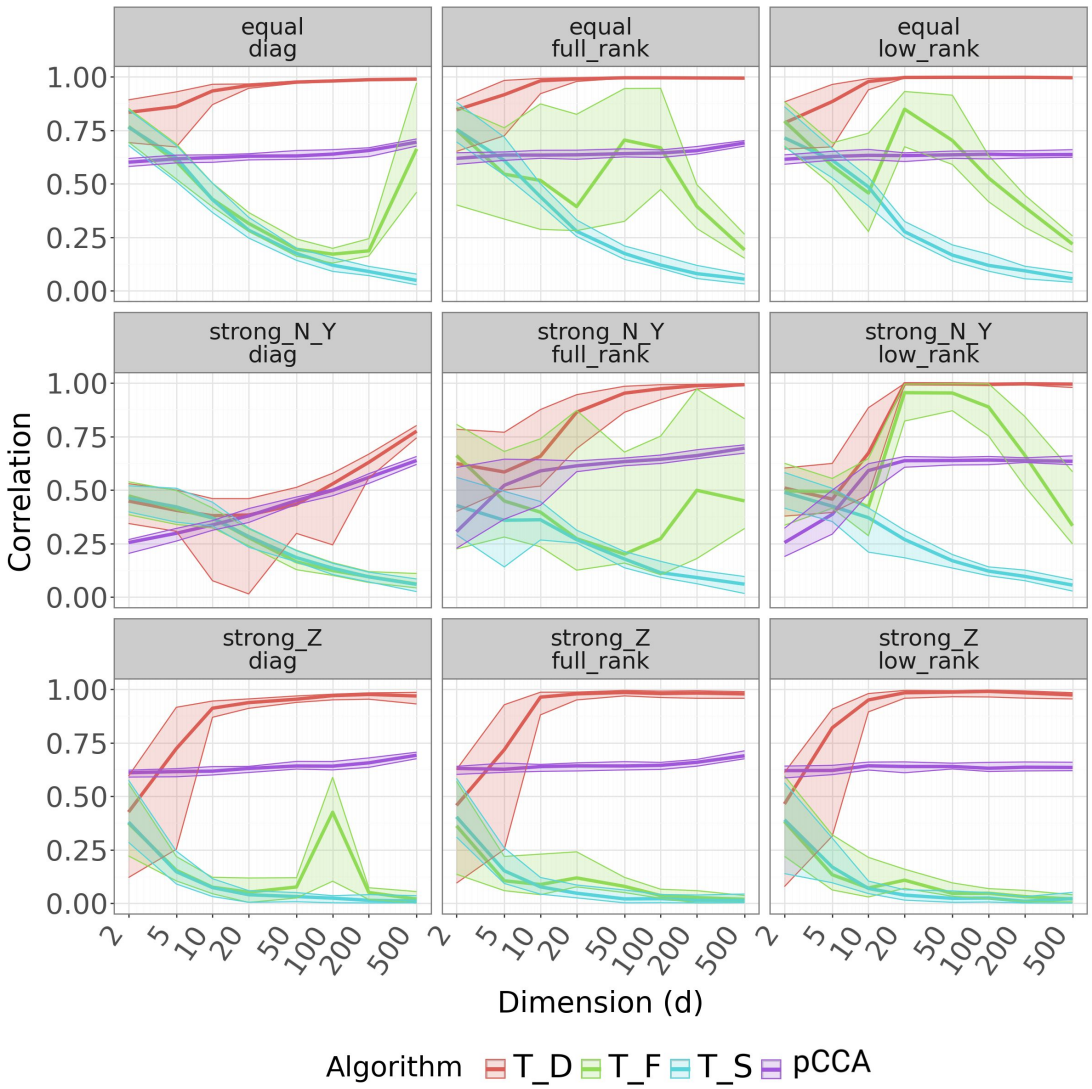



Figure 7: Experiments with nonlinear map $f_a$ and with different noise structure ($\mathbf{\Sigma}$ being diagonal, full rank and low rank) and scaling factors ($(u, v, w)$ as $(1/3, 1/3, 1/3)$, $(0.1, 0.1, 0.8)$, and $(0.1, 0.8, 0.1)$ for equal, Strong_N_Y and Strong_Z). Overall, the learning algorithm $T_D$ performs better and tends to converge.

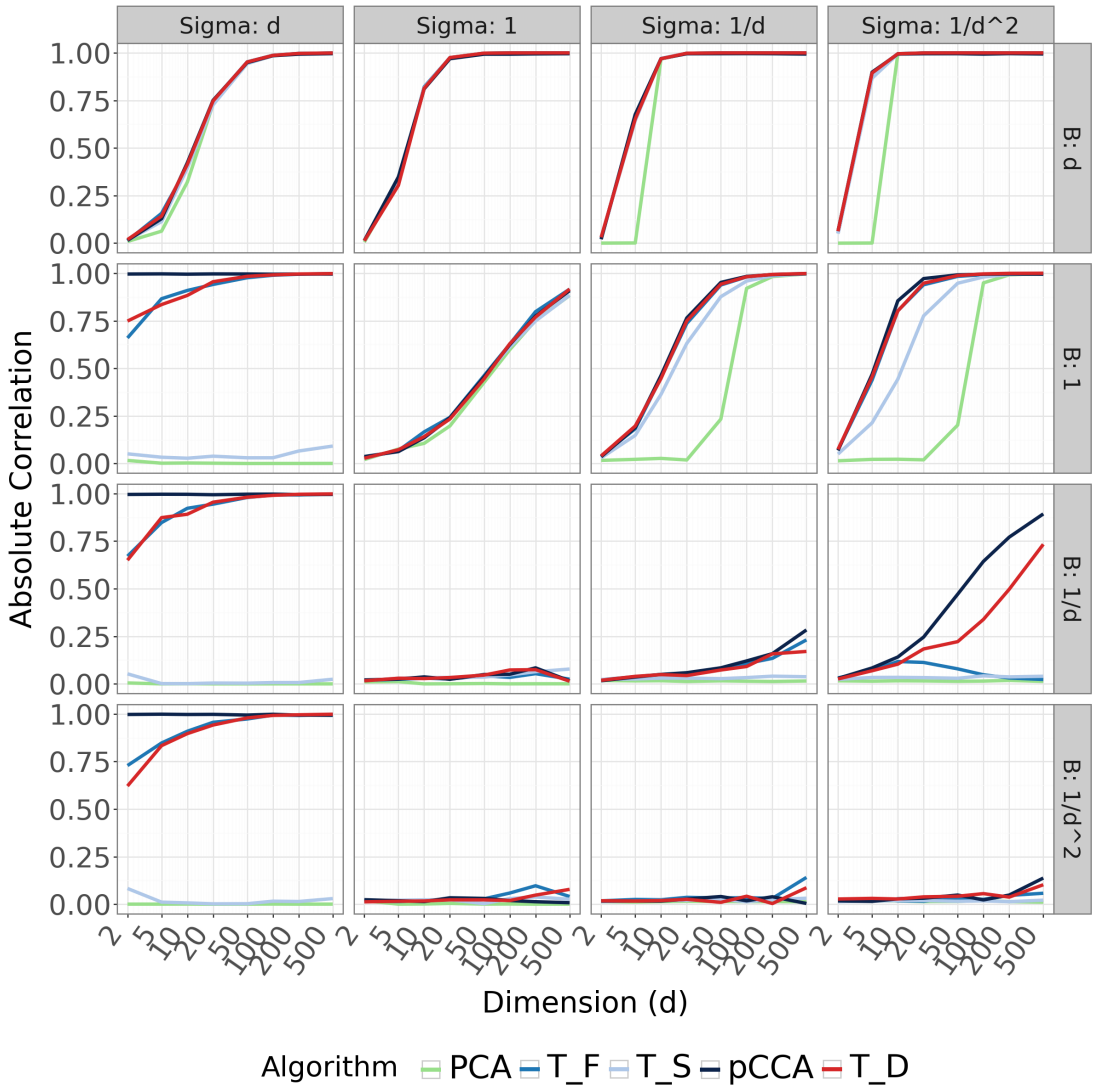



Figure 8: Experiment where $X \perp\!\!\!\perp Z$. When the dependence between $X$ and $Z$ is removed, the pCCA algorithm performs similarly to $T_D$. This highlights the effectiveness of our approach in scenarios with confounding or mediation effects.

similarly to the other algorithms. This result highlights the robustness of $T_D$, as it remains stable across different structural relationships between $X$ and $Y$, reinforcing its applicability in both confounded and mediated settings.

The experiments with both exponential and uniform noise, shown in Figure 9, demonstrate that the convergence behavior of the algorithms remains consistent across Gaussian, exponential, and uniform noise types. This aligns with our theoretical convergence guarantees presented in Section 4.1, which are noise-agnostic by design.

**Hypothesis testing** For a test to be valid, it must control the Type I error rate. Specifically, if we test at level $\alpha$, then under the null hypothesis $H_0$, the probability of rejecting $H_0$ should be less than or equal to $\alpha$. In Fig. 11, we observe that all the methods control the Type I error reasonably accurately.

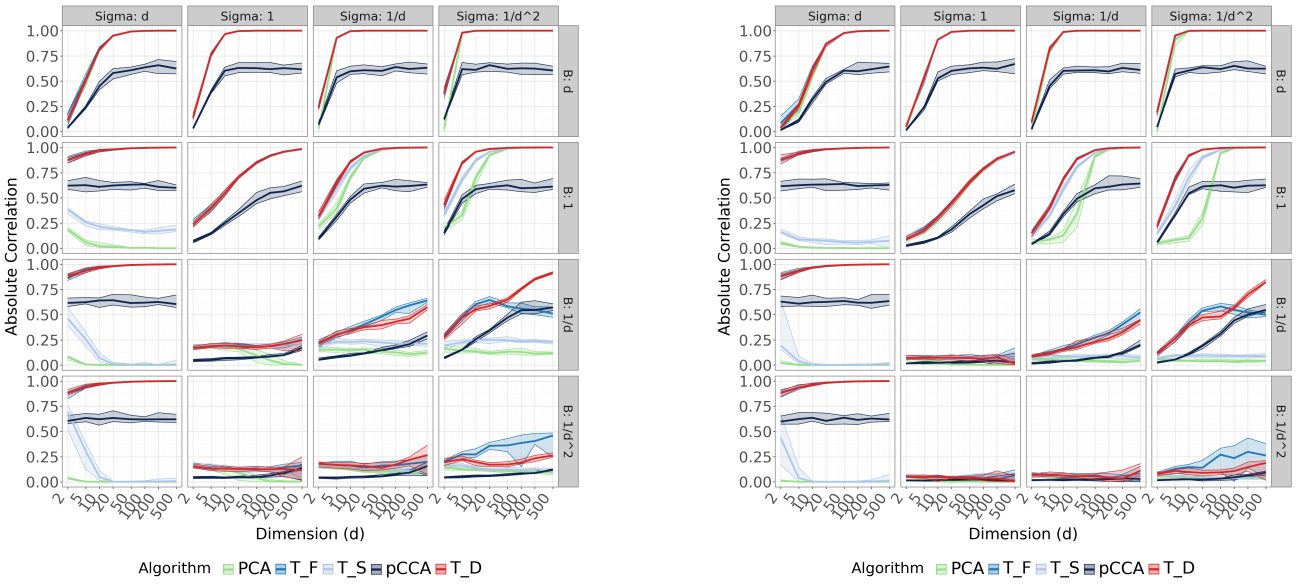



Figure 9: Uniform noise (left) and exponential noise (right).

## C.2 REAL-WORLD EXPERIMENTS

**Separating internal climate variability from the externally forced response.** This experiment aims to assess the performance of our learning algorithms in disentangling internal climate variability from the forced response to external factors, such as greenhouse gas (GHG) emissions or solar radiation. For this analysis, we focus on temperature fields. We use $M = 50$ members from the CESM2 historical climate simulations (Danabasoglu et al., 2020), covering 1880 to 2014. The variables under consideration are Sea Level Pressure (SLP) and Temperature (T), with monthly data yielding 1669 samples per member. The detrended SLP field is treated as a proxy for internal variability ($Z \in \mathbb{R}^{648}$). In contrast, the temperature field ($Y \in \mathbb{R}^{648}$) serves as the response variable of interest. The temperature for member $i$, at location $j$, and time $t$, is denoted by $Y_j^{(i)}(t)$.

As a proxy for climate external forcing, we use a smoothed version (5-year moving average) of the Global Mean Temperature (GMT), which is computed as a spatial average of the temperature field ($X \in \mathbb{R}$):

$$X(t) = \frac{1}{\text{years} \times 12} \sum_{\tau=1}^{\text{years}\times 12} \frac{1}{d} \sum_{j=1}^{d} Y_j(t-\tau) \tag{35}$$

where $d$ represents the number of spatial locations.

The climate-forced response, $Y_{\text{forced}}$, is calculated as the ensemble mean over all simulation members, $Y^{(i)}$:

$$Y_{\text{forced},j}(t) = \frac{1}{M} \sum_{i=1}^{M} Y_j^{(i)}(t) \quad \text{and} \quad Y_{\text{internal},j}^{(i)}(t) = Y_j^{(i)}(t) - Y_{\text{forced},j}(t), \tag{36}$$

where $Y_{\text{internal}}$ represents the true internal variability of $Y$ after removing the climate-forced component.

The Direct Effect (DEA) algorithm (employing $T_D$) is applied as follows: We train DEA using the triplet $\{GMT(t), T(t), SLP(t)\}_{t=1}^{years\times 12}$ as realisations of $(X, Y, Z)$, where $GMT$ serves as the climate external forcing proxy ($X$), $T$ represents the temperature response variable, and $SLP$ is a proxy for the internal climate variability ($Z$). Once the model is trained, we project the data onto the null space of the vector $\mathbf{b}$, denoted as $\mathbf{b}^{\perp}$, to recover the internal variability component $\hat{Y}_{\text{internal}}$. This projection isolates the portion of the temperature field that is not correlated with the external forcing, allowing us to separate the forced and internal components effectively. Finally, we compute the climate-forced response as $\hat{Y}_{\text{forced}} = Y - \hat{Y}_{\text{internal}}$, which provides an estimate of the temperature response attributed to external forcing alone.

We compare our learning algorithm with two common approaches used in climate science:

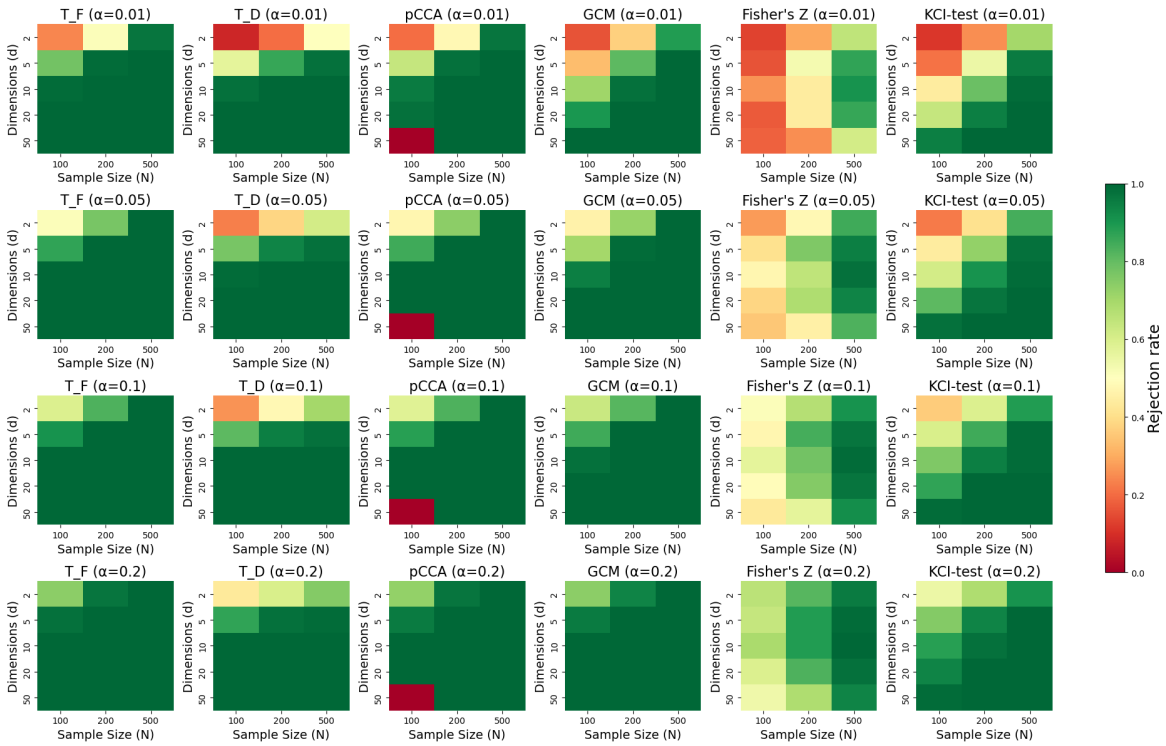



Figure 10: Power of the different methods. $T_D$, $T_F$, and pCCA generally exhibit better performance compared to the other approaches. This is partly because they rely on linear Gaussian models, which constrain the alternative hypotheses, thereby improving the power of these tests.

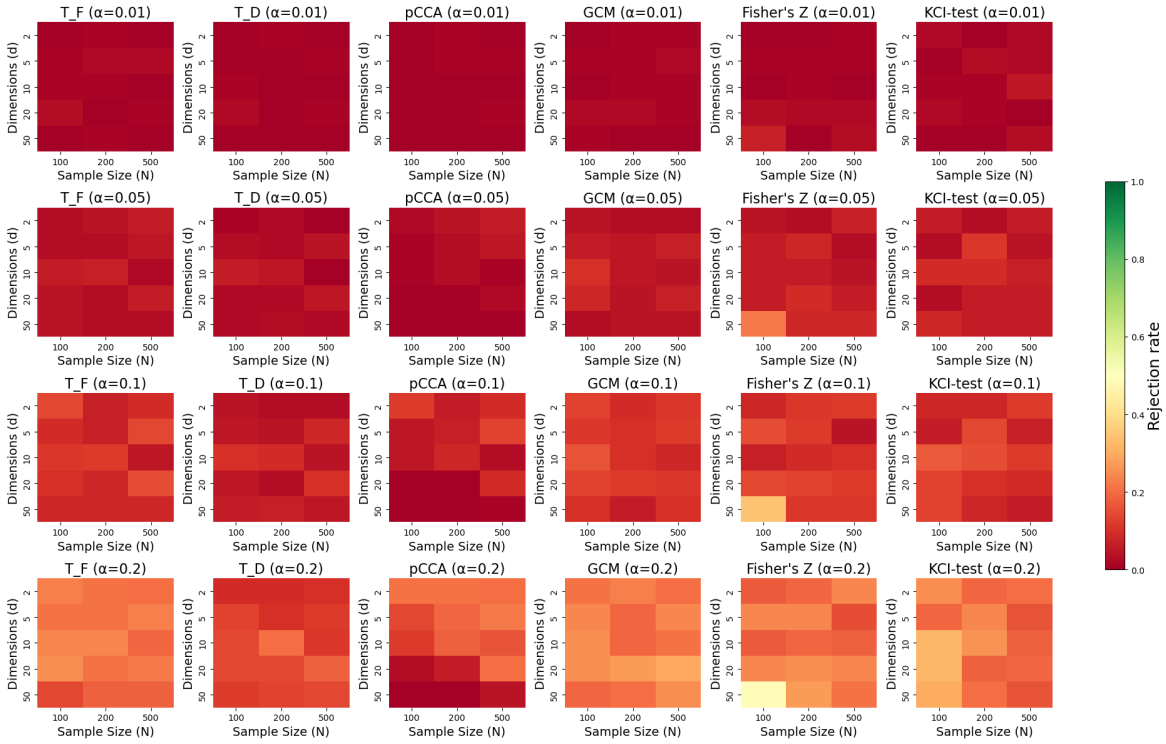
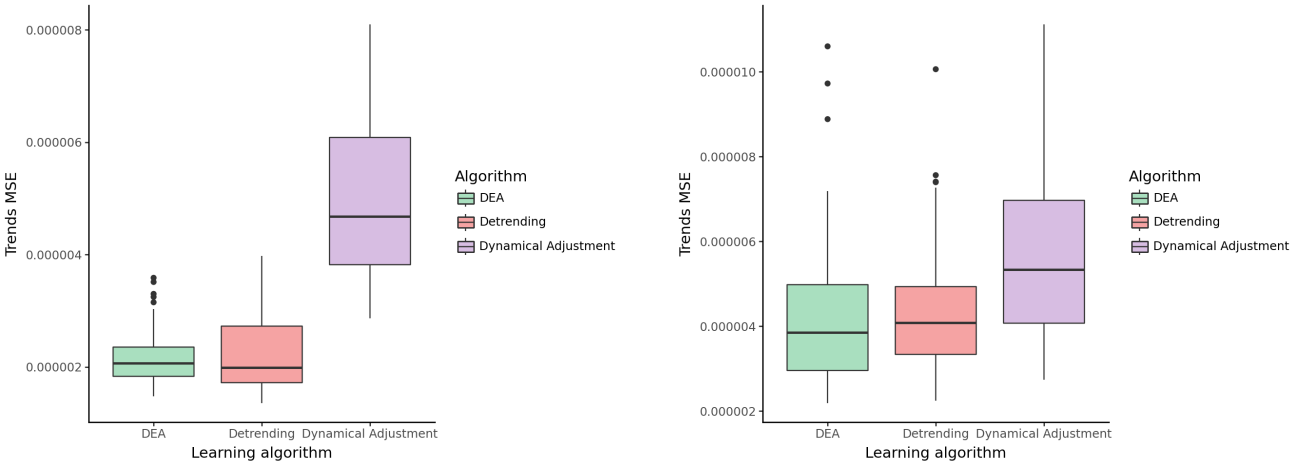



Figure 11: Type $I$ error control. We can observe that all methods have a good control of type $I$ error except the Fisher $Z$ test, which has a poor control in low sample high dimensional settings.




(a) Forced response trends

(b) Internal variability trends

Figure 12: Mean Squared Error (MSE) of different algorithms in reconstructing (a) internal climate variability trends and (b) forced response trends. The Direct Effect Analysis (DEA) algorithm, using the basis $(\mathbf{\Sigma}^{-1}\mathbf{b}, \mathbf{b}^{\perp})$, i.e. $T_D$ algorithm, is compared to Detrending and Dynamical Adjustment (two most common approaches for separating internal from external climate variability). Overall, DEA and Detrending perform better. DEA outperforms Detrending for internal variability trend estimation but has a higher median MSE for forced trend reconstruction. However, DEA provides better worst-case control in this case.

1. **Detrending:** A simple linear model predicts $Y$ from $X$, providing an estimate of the climate-forced response, $\hat{Y}_{\text{forced}}$, and the dynamical component as $\hat{Y}_{\text{internal}} = Y - \hat{Y}_{\text{forced}}$.
2. **Dynamical Adjustment** (Sippel et al., 2019): A model is trained using both $X$ (GMT) and $Z$ (SLP). Predictions are made by setting $Z$ to zero, isolating the dynamical component.

We compute trends for the dynamical components ($Y_{\text{internal}}$, $\hat{Y}_{\text{internal}}$, $Y_{\text{forced}}$, and $\hat{Y}_{\text{forced}}$) over 20-year periods. The performance of the methods is evaluated using the following metrics:

- **20-year trends MSEs:** MSEs for the trends of the three methods are shown in Fig. 12 (a) for forced trends and (b) for internal variability trends. The boxplots display the MSE distributions across different simulation members.
- **20-year trends maps (internal variability):** Internal variability trends are compared spatially in Fig. 13 for DEA and Fig. 14 for Detrending to better understand model biases.
- **20-year trends time series (forced response):** Forced response trends are compared over time, with time series for DEA and Detrending plotted for randomly selected locations in Fig. 15.

We train the algorithms (DEA, Detrending, Dynamical Adjustment), extract latent structures $\hat{Y}_{\text{forced}} = \mathbf{w}^{\top} Y$, and compare the trends of the last 20 years of $\hat{Y}_{\text{forced}}$ and $Y_{\text{forced}}$. Figure 12 shows that DEA performs similarly to Detrending but outperforms Dynamical Adjustment.

A qualitative evaluation of the trend maps generated by DEA (Figure 13) shows that the algorithm captures the warming and cooling patterns. However, both DEA and Detrending tend to underestimate trends in polar regions where temperature trends are generally stronger.

Figure 15 shows that both Detrending and DEA effectively capture the forced response trends. However, in regions where the forced response exhibits high variability (e.g., at $d = 30$ or $d = 493$, typically located in polar regions), both methods struggle to fully capture this variability. This may be due to the smoothing of GMT in the external forcing, but this observation warrants further investigation, as these regions of high variability may also reflect model artefacts. Further exploration of these phenomena is needed.

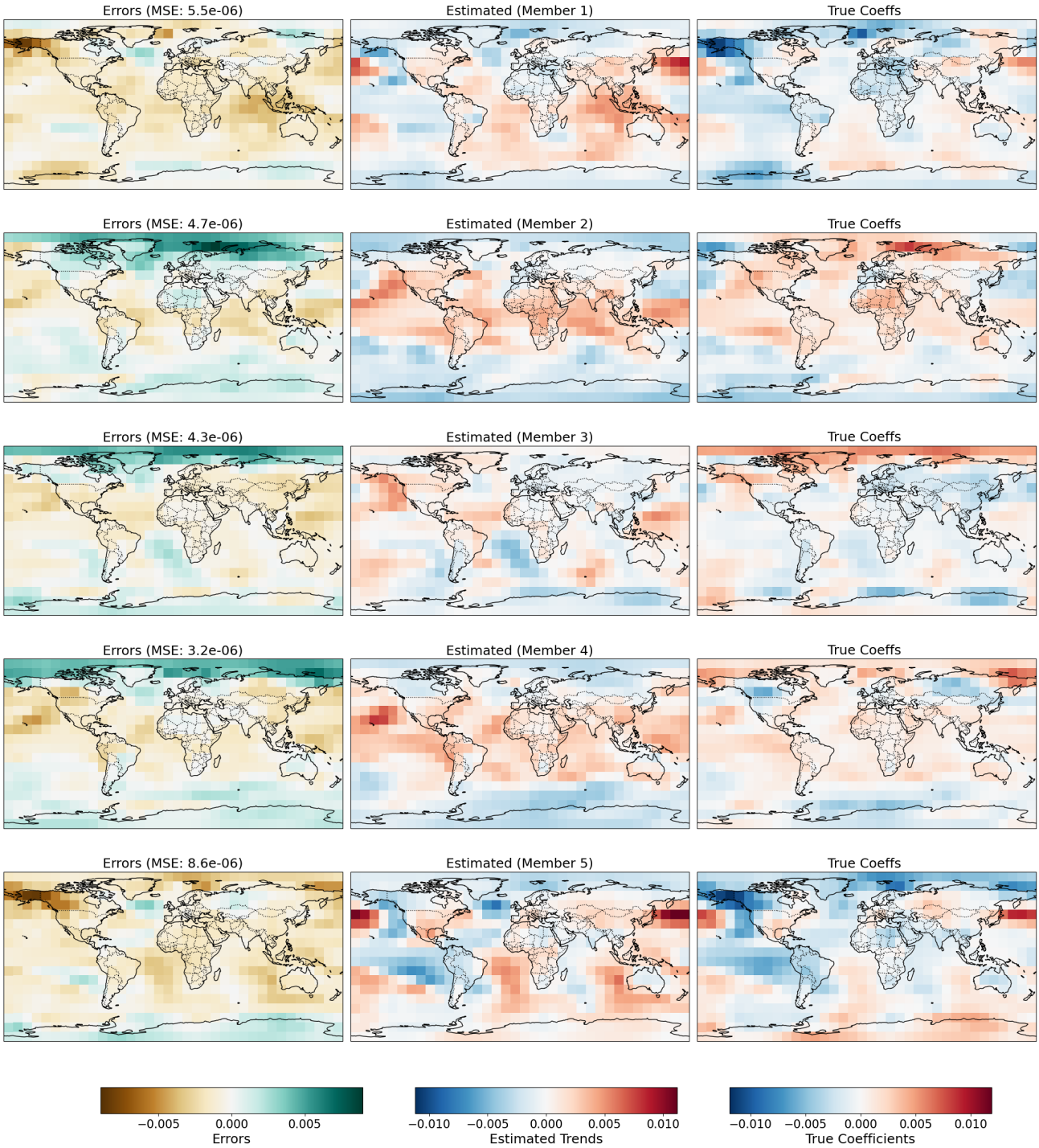



Figure 13: Trends of the reconstructed internal climate variability using the DEA algorithm. The algorithm captures general warming and cooling patterns but underestimates trends in the North Pole and overestimates them in Western America.

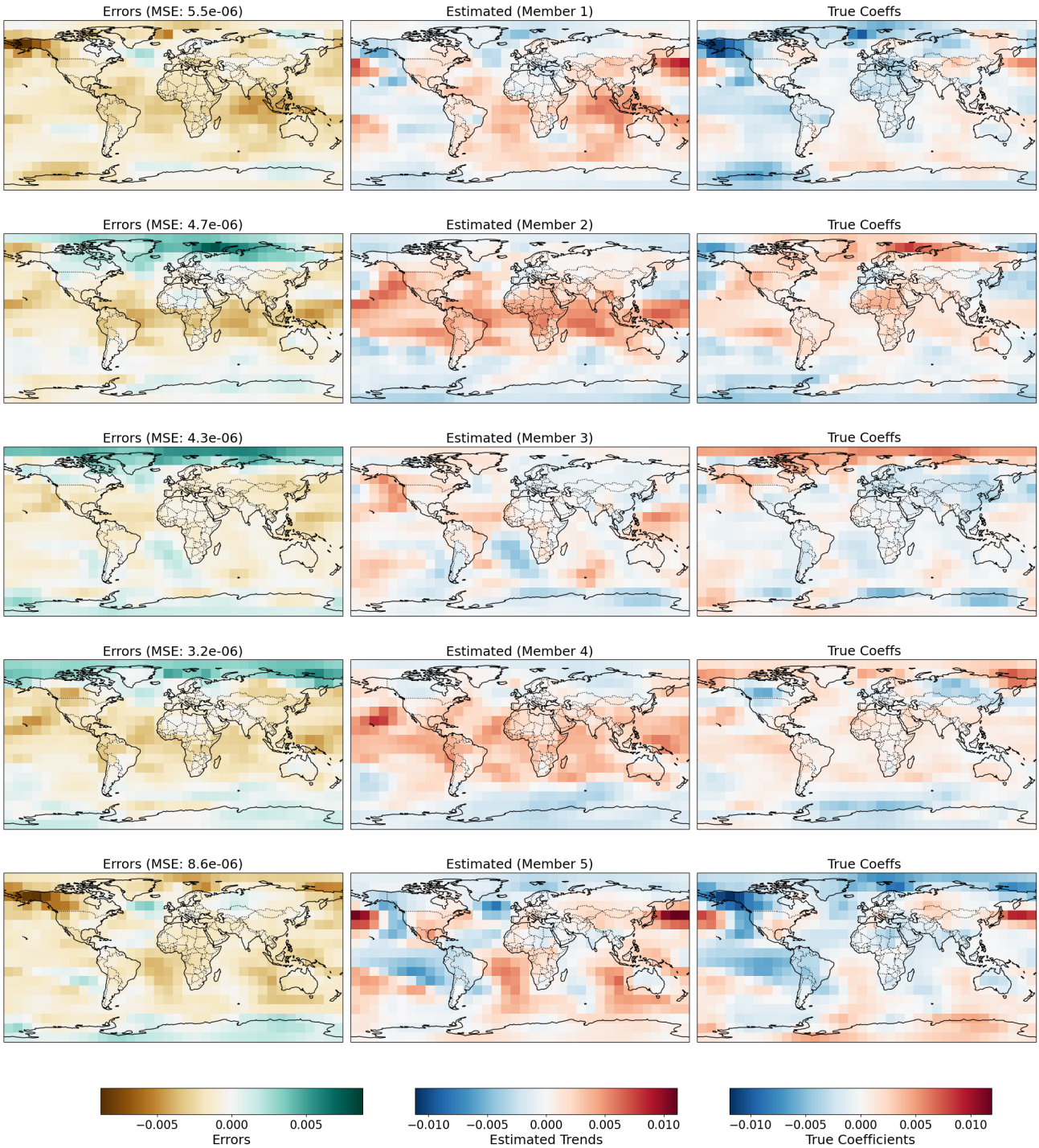



Figure 14: Trends of the reconstructed internal climate variability using the Detrending algorithm. The algorithm captures general warming and cooling patterns but underestimates trends in the poles and overestimates trends in Western America and Indonesia.

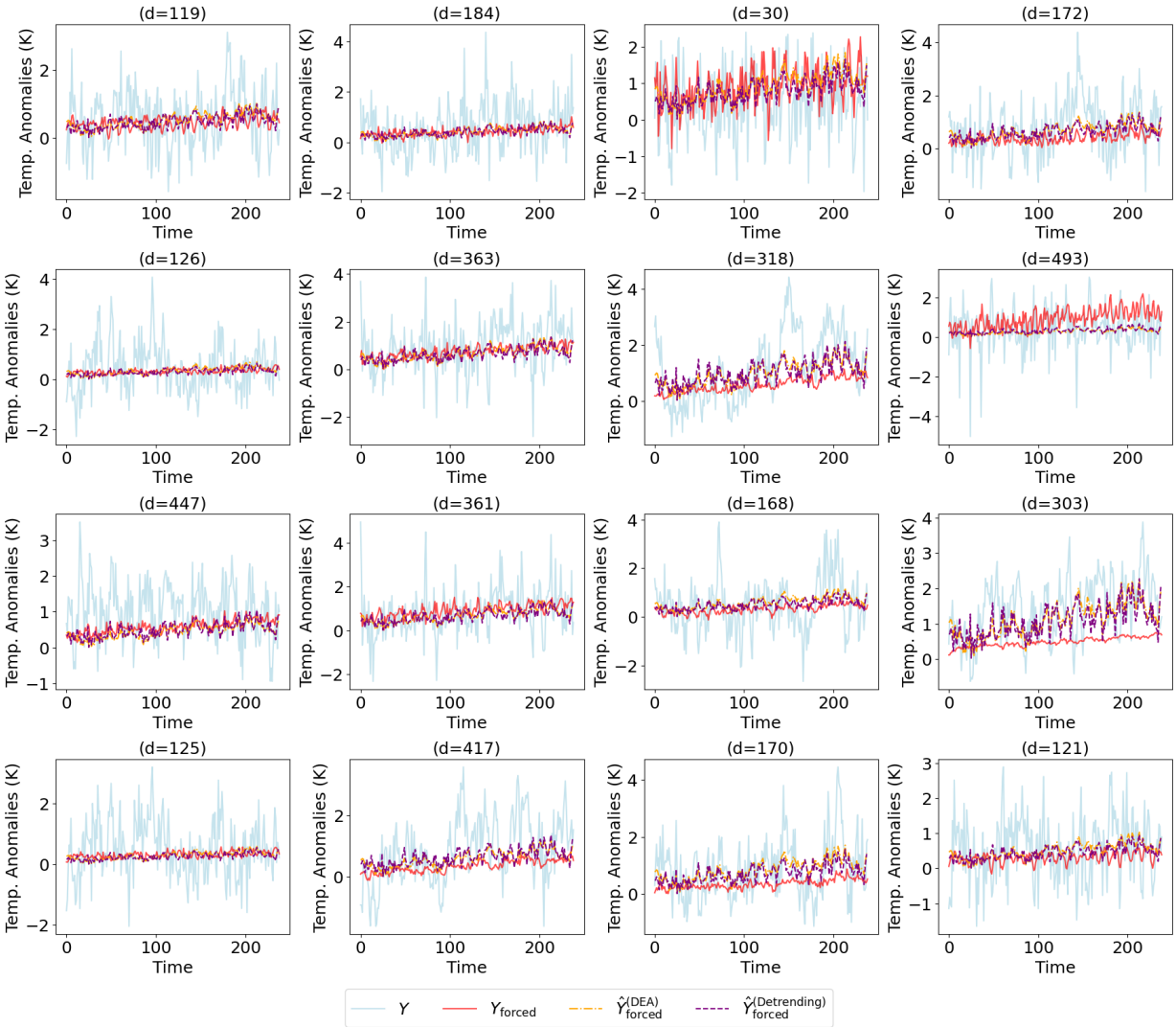



Figure 15: Comparison of original observations $Y$ and the reconstructed climate-forced response $\hat{Y}_{\text{forced}}$ at 16 randomly selected locations for both DEA and Detrending.