Dominik Janzing* and Bernhard Schölkopf

Detecting Confounding in Multivariate Linear Models via Spectral Analysis

<https://doi.org/10.1515/jci-2017-0013>

Received April 05, 2017; revised July 14, 2017; accepted: September 21, 2017

Abstract: We study a model where one target variable Y is correlated with a vector $\mathbf{X} := (X_1, \dots, X_d)$ of predictor variables being potential causes of Y . We describe a method that infers to what extent the statistical dependences between $\mathbf{X}$ and Y are due to the influence of $\mathbf{X}$ on Y and to what extent due to a hidden common cause (confounder) of $\mathbf{X}$ and Y . The method relies on concentration of measure results for large dimensions d and an independence assumption stating that, in the absence of confounding, the vector of regression coefficients describing the influence of each X_i on Y typically has ‘generic orientation’ relative to the eigenspaces of the covariance matrix of $\mathbf{X}$. For the special case of a scalar confounder we show that confounding typically spoils this generic orientation in a characteristic way that can be used to quantitatively estimate the amount of confounding (subject to our idealized model assumptions).

Keywords: confounding, independence of mechanisms, spectral analysis

1 Introduction and general model

Estimating the causal influence of some variables $X_1, \dots, X_d$ on a target variable Y is among the most important goals in statistical data analysis. However, drawing causal conclusions from observational data alone without intervening on the system is difficult. This is because the observed statistical dependences between Y and each X_j need not be due to an influence of X_j on Y . Instead, due to Reichenbach’s Principle of Common Cause [1], Y may also be the cause of X_j or there may be a common cause Z influencing both. In many applications, time order or other prior information excludes that Y influences X_j . For instance, if Y describes the health condition of a patient at time t and X_j some treatments at an earlier time, we ‘only’ need to decide to what extent the dependences between X_j and Y are due to X_j influencing Y and to what extent they are due to common causes (‘confounders’). Here we are not interested in the reason for dependences between the variables X_j themselves, we therefore merge them to a vector-valued variable $\mathbf{X}$. Moreover, we restrict the attention to the case where there is only one real-valued confounder Z . In the case of linear relations, the structural equations then read:

$$\mathbf{X} = \mathbf{b}Z + \mathbf{E} \quad (1)$$

$$Y = \langle \mathbf{a}, \mathbf{X} \rangle + cZ + F. \quad (2)$$

where $\mathbf{E}$ is a random vector with values in $\mathbb{R}^d$ and Z, F are scalar random variables. Here, $\mathbf{E}, Z, F$ are jointly independent, while the components of $\mathbf{E}$ may be dependent. Further, $\mathbf{a} \in \mathbb{R}^d$ is the vector of structure coefficients determining the influence of the d -dimensional variable $\mathbf{X}$ on the scalar target variable Y . Likewise, $\mathbf{b} \in \mathbb{R}^d$ is the vector determining the influence of Z on $\mathbf{X}$ and the scalar $c \in \mathbb{R}$ is the structure coefficient determining the influence of Z on Y . By rescaling $\mathbf{b}$ and c , we may assume Z to have unit variance without loss of generality. The corresponding DAG is shown in Figure 1. If all variables are centered Gaussian, the remaining model parameters are the vectors $\mathbf{a}, \mathbf{b}$, the covariance matrix $\Sigma_{\mathbf{E}\mathbf{E}}$ and the scalars c, σ_F , where c

*Corresponding author: Dominik Janzing, Department ‘Empirical Inference’, Max Planck Institute for Intelligent Systems, Spemannstr. 36, 70569 Tübingen, Germany, E-mail: dominik.janzing@tuebingen.mpg.de

Bernhard Schölkopf, Department ‘Empirical Inference’, Max Planck Institute for Intelligent Systems, Tübingen, Germany

Figure 1: Most general DAG considered in this paper. All scenarios discussed later refer to this DAG and differ only by the parameters $\mathbf{a}$, $\mathbf{b}$, σ_F , Σ_{EE} , c .

describes the strength of the influence of Z on Y and σ_F the standard deviation of F . Since this paper will be based on second-order statistics alone, we will treat these parameters as the only relevant ones. The following special cases can be obtained by appropriate choices of these parameters:

Purely causal: The case $\mathbf{X} \rightarrow Y$ with no confounding can easily be obtained by setting $\mathbf{b} = 0$ or $c = 0$. In the first case, Z takes the role of an additional independent noise term (apart from F), see Figure 2(a).

The structural equation then reads

$$Y = \langle \mathbf{a}, \mathbf{X} \rangle + cZ + F = \langle \mathbf{a}, \mathbf{X} \rangle + \tilde{F}, \quad (3)$$

with $\mathbf{E}, \tilde{F}$ being jointly independent. For fixed c , the limit of a deterministic influence of $\mathbf{X}$ on Y can be obtained by letting at least one component of the vector $\mathbf{a}$ grow to infinity. Then Y is dominated by the term $\langle \mathbf{a}, \mathbf{X} \rangle$.

Purely confounded: Setting $\mathbf{a} = 0$ turns the influence of $\mathbf{X}$ on Y off. Then the relation between $\mathbf{X}$ and Y is generated by the confounder Z only, see Figure 2(b). Depending on the remaining parameters Σ_{EE} , σ_F and $\mathbf{b}$, we can also obtain a scenario where $\mathbf{X}$ provides perfect knowledge about Z (when $\|\mathbf{b}\|^2 \rightarrow \infty$ or $\Sigma_{EE} \rightarrow 0$) or a scenario where Y provides perfect knowledge of Z ($\sigma_F \rightarrow 0$).

Purely anticausal: We have actually excluded a scenario where Y is the cause of $\mathbf{X}$. Nevertheless, if $\mathbf{a} = 0$ and $\sigma_F = 0$, we have $Y = Z$ almost surely and Z is the cause of $\mathbf{X}$. Hence, the scenario gets indistinguishable from an 'anticausal' scenario where Y is the cause as in Figure 3(b), although performing interventions on Y would still tell us that it is not the cause.

We now ask how to distinguish between these cases given joint observations of $\mathbf{X}$ and Y that are i.i.d. drawn from $P_{\mathbf{X},Y}$. Conditional statistical independences, which are usually employed for causal inference, [2, 3] may not help to distinguish between the above cases if there are no conditional independences in $P_{\mathbf{X},Y}$. Moreover, we assume that there are no observed causes of $\mathbf{X}$ that could act as so-called *instrumental* variables [4] which would enable the distinction between 'causal' and 'confounded'.

To see that the parameters $\mathbf{a}$, $\mathbf{b}$, Σ_{EE} , σ_F are heavily underdetermined in linear *Gaussian* models, just note that any multivariate Gaussian $P_{\mathbf{X},Y}$ can be explained by $\mathbf{X}$ being an unconfounded cause of Y according to the structural eq. (3) by replacing $\mathbf{a}$ with

$$\hat{\mathbf{a}} := \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}Y}. \quad (4)$$

Figure 2: (a) Purely causal case: setting $\mathbf{b} = 0$ renders Z independent of $\mathbf{X}$. Thus, we can consider $\tilde{F} := cZ + F$ formally as noise term and then obtain the simple DAG with nodes $\mathbf{X}$, Y only. (b) Purely confounded case: by setting $\mathbf{a} = 0$ the correlations between $\mathbf{X}$ and Y are only due to the confounder Z .

Figure 3: (a) A special case of the purely confounded case is obtained by setting the noise F of Y to 0 and $c = 1$. Then Y is an exact copy of Z , i.e., the cause of X . Since $P_{X,Y} = P_{X,Z}$, no statistical method relying purely on observational data is able to distinguish this case from the scenario in (b), although the difference certainly matters when interventions on Y are made.

The vector $\hat{\mathbf{a}}$ is the vector of regression coefficients obtained by regressing Y on $\mathbf{X}$ without caring about the true causal structure. Here we use the symbol $\hat{\mathbf{a}}$ instead of $\mathbf{a}$ to indicate that it differs from the vector $\mathbf{a}$ that appears in the structural eq. (2) which correctly describes the causal relation between $\mathbf{X}$ and Y . This way, we obtain a model that correctly describes the observed correlations, but not the *causal* relations since the impact of interventions is not modelled correctly. – Note that identifying $\mathbf{a}$ is typically the main goal of causal data analysis since a_j directly describes how changing X_j changes Y . Confusing $\mathbf{a}$ with $\hat{\mathbf{a}}$ would be the common fallacy of naively attributing all dependences to the causal influence of X_j on Y . To see the relation between $\hat{\mathbf{a}}$ and $\mathbf{a}$ we first find

$$\begin{aligned}\Sigma_{XY} &= \text{Cov}(\mathbf{X}, Y) = \text{Cov}(\mathbf{E} + \mathbf{b}Z, \langle \mathbf{a}, \mathbf{X} \rangle + cZ + F) \\ &= \text{Cov}(\mathbf{E} + \mathbf{b}Z, \langle \mathbf{a}, \mathbf{E} + \mathbf{b}Z \rangle + cZ + F) = (\Sigma_{EE} + \mathbf{b}\mathbf{b}^T)\mathbf{a} + c\mathbf{b},\end{aligned}$$

where we have used the joint independence of $\mathbf{E}, Z, F$ and that Z is normalized. Likewise,

$$\Sigma_{XX} = \text{Cov}(\mathbf{X}, \mathbf{X}) = \text{Cov}(\mathbf{E} + \mathbf{b}Z, \mathbf{E} + \mathbf{b}Z) = \Sigma_{EE} + \mathbf{b}\mathbf{b}^T.$$

Due to eq. (4) we thus obtain

$$\hat{\mathbf{a}} = \mathbf{a} + (\Sigma_{EE} + \mathbf{b}\mathbf{b}^T)^{-1}c\mathbf{b}. \quad (5)$$

Equation (5) shows that the vector $\hat{\mathbf{a}}$ obtained by standard regression consists of $\mathbf{a}$ (which defines the causal influence of $\mathbf{X}$ on Y) and a term that is due to confounding.

It is known that confounding can be detected in linear models with *non-Gaussian* variables [5]. This is because describing data, that has been generated by the model eqs. (1) and (2), by the structural equation

$$Y = \langle \hat{\mathbf{a}}, \mathbf{X} \rangle + F',$$

as in eq. (3), yields in the generic case a noise variable F' that is not statistically *independent* of $\mathbf{X}$, although it is *uncorrelated*. Other recent proposals to detect confounding using information beyond conditional statistical dependences rely on different model assumptions. Ref. [6], for instance assumes non-linear relations with additive noise, while Ref. [7] assumes a discrete confounder attaining a few values only.

Here we propose a method for distinguishing the purely causal from the confounded case in the linear scenario above that only relies on second order statistics and thus does not rely on non-Gaussianity of noise variables and higher-order statistical independence tests like [6], for instance. In other words, the relation must be linear but the distributions may be Gaussian or not. The paper is structured as follows. Section 2 defines the strength of confounding, which is the crucial target quantity to be estimated. Section 3 describes the idea of the underlying principle, defines it formally in terms of a spectral measure and justifies it by a toy model where parameters are randomly generated. Section 4 describes the method to estimate the strength and justifies it by intuitive arguments first and by theoretical results which are rigorously shown in Section 5.

Section 6 reports results with simulated causal structures according to the model assumptions, while Section 7 describes results from an experiment where the data has been generated by an electronic and optical setup with known causal structure. Section 8 reports some studies with observational data where the causal ground truth is only partially known.

2 Characterizing confounding by two parameters

2.1 Strength of confounding

We now define a strength of confounding that measures how much $\hat{\mathbf{a}}$ deviates from $\mathbf{a}$ (relative to the sum of the squared lengths of these vectors). The goal of this paper is to estimate this quantity from $P_{\mathbf{X},Y}$ alone.

Definition 1. (structural strength of confounding) *The structural strength of confounding is defined by*

$$\beta := \frac{\|c \cdot \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}\|^2}{\|\mathbf{a}\|^2 + \|c \cdot \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}\|^2} \in [0, 1]. \quad (6)$$

We will later see that, subject to the genericity assumptions stated in Section 3, the denominator can be approximated via

$$\|\mathbf{a}\|^2 + \|c \cdot \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}\|^2 \approx \|\mathbf{a} + c \cdot \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}\|^2 = \|\hat{\mathbf{a}}\|^2. \quad (7)$$

It should be emphasized that β does not measure to what extent the confounder changes $\Sigma_{\mathbf{X}Y}$. This could be quantified by a parameter that we call ‘correlative strength of confounding’, denoted by γ , and defined via

$$\gamma := \frac{\|\mathbf{c}\mathbf{b}\|^2}{\|\Sigma_{\mathbf{X}\mathbf{X}}\mathbf{a}\|^2 + \|\mathbf{c}\mathbf{b}\|^2} \in [0, 1]. \quad (8)$$

It is non-trivially related to β . Remarkably, β and γ can differ by orders of magnitudes¹. Without claiming that β would be the better measure², we focus on β because it is more relevant for causal statements: whenever β is large identifying $\hat{\mathbf{a}}$ with $\mathbf{a}$ yields significantly wrong causal conclusions even when γ is small. We have introduced γ only to show that whether confounding is negligible or not highly depends on how it is quantified.

2.2 A second parameter characterizing confounding

The contribution of Z to the covariance between $\mathbf{X}$ and Y is given by the product $c \cdot \mathbf{b}$. Accordingly, rescaling $\mathbf{b}$ with some factor while rescaling c with the inverse factor preserves the *correlative* strength of confounding. Although *structural* confounding strength is affected by rescaling $\mathbf{b}$ and c in a more sophisticated way, β can also be unaffected by rescaling both $\mathbf{b}$ and c in an appropriate way. The regimes with small $\mathbf{b}$ and large c versus the one with large $\mathbf{b}$ and small c have simple interpretations: in the first case, the uncertainty of $\mathbf{X}$ is hardly reduced by knowing Z , while in the second case, knowing Z reduces most of the uncertainty of $\mathbf{X}$.

¹ This is an interesting phenomenon in high dimensions: a confounder may generate almost no observable covariance between $\mathbf{X}$ and Y and still perturb the vector of regression coefficients significantly.

² Note that quantifying causal influence in causal Bayesian networks is non-trivial and there exists no generally accepted measure [8].

To distinguish between these different regimes of confounding we introduce the second parameter η , which measures the explanatory power of Z for $\mathbf{X}$:

$$\eta := \text{tr}(\Sigma_{\mathbf{X}\mathbf{X}}) - \text{tr}(\Sigma_{\mathbf{X}\mathbf{X}|Z}) = \text{tr}(\Sigma_{\mathbf{X}\mathbf{X}}) - \text{tr}(\Sigma_{\mathbf{E}\mathbf{E}}) = \|\mathbf{b}\|^2 \leq \|\Sigma_{\mathbf{X}\mathbf{X}}\|.$$

For the entire Section 2 it is important to keep in mind that we always referred to the case where Z has unit variance.

3 Detecting confounders by the principle of generic orientation

3.1 Intuitive idea and background

The idea of our method is based on the recently stated *Principle of Independent Conditionals* [9, 10] in the context of causal inference. To introduce it, let G be a directed acyclic graph (DAG) formalizing the hypothetical causal relations among the random variables $Z_1, \dots, Z_n$. The distributions compatible with this causal structure factorize into

$$P_{Z_1, \dots, Z_n} = \prod_{j=1}^n P_{Z_j|PA_j},$$

where each $P_{Z_j|PA_j}$ denotes the conditional distribution of Z_j , given its parents [2]. Informally speaking, the Principle of Independent Conditionals states that, usually, each $P_{Z_j|PA_j}$ describes an independent mechanism of nature and therefore these objects are ‘independent’ and contain ‘no information’ about each other. [9, 10] formalized ‘no information’ by postulating that the description length of one $P_{Z_j|PA_j}$ does not get shorter when the description of the other $P_{Z_i|PA_i}$ for $i \neq j$ are given. Here, description length is defined via Kolmogorov complexity, which is, unfortunately, uncomputable [11]. To deal with this caveat, one can either *approximate* Kolmogorov complexity, or, as shown in [12], indirectly use the principle as a justification for new inference methods rather than as an inference method itself.

However, there are also other options to give a definite meaning to the term ‘independence’. To see this, consider some parametric model where each $P_{Z_j|PA_j}$ is taken from a set of possible conditionals $P_{Z_j|PA_j}^{\theta_j}$ where θ_j is taken from some parameter space Θ_j . Assume that, for a given distribution $P_{Z_1, \dots, Z_n}$, the parameters $\theta_1, \dots, \theta_n$ are related by an equation (i.e., one is the function of the others) that is not satisfied by *generic* n -tuples. One can then consider this as a hint that the mechanisms corresponding to the $P_{Z_j|PA_j}$ have not been generated independently and become skeptical about the causal hypothesis³. This philosophical argument is also the basis for Causal Faithfulness [3], that is, the principle of rejecting a causal DAG for which the joint distribution satisfies conditional independences that do not hold for *generic* vectors $(\theta_1, \dots, \theta_n)$ because it requires the vector to lie in a lower dimensional manifold. In its informal version, the Principle of Independent Conditionals generalizes this idea by excluding also other ‘non-generic’ relations between parameter vectors θ_j . More precisely, the assumption of faithfulness can be justified by generating models according to which unfaithful distributions occur with probability zero [13]. For practical purposes one needs to also exclude *almost* unfaithful distributions and assume that they occur only with small probability (although this is problematic [14]). Here we describe relations between parameters that are also non-generic in the sense that they occur with small probability according to our generating model.

After these general remarks, let us come back to the scenario where the causal hypothesis reads $\mathbf{X} \rightarrow Y$ (without confounding and within a linear model). Recalling that we consider $\Sigma_{\mathbf{X}\mathbf{X}}$ the crucial model parameter for $P_{\mathbf{X}}$ and the regression vector $\mathbf{a}$ the crucial parameter for $P_{Y|\mathbf{X}}$, we therefore postulate that $\mathbf{a}$ lies in a ‘generic’ orientation relative to $\Sigma_{\mathbf{X}\mathbf{X}}$ in a sense to be described in Subsection 3.2. To approach this idea

³ See [10], Theorem 3, for a detailed discussion of the conditions under which one should trust this argument.

first by intuition, note, for instance, that it is unlikely that $\mathbf{a}$ is close to being aligned with the eigenvector of $\Sigma_{\mathbf{X}\mathbf{X}}$ corresponding to the largest eigenvalue (i.e., the *first* principal component), given that $\mathbf{a}$ has been chosen ‘without knowing’ $\Sigma_{\mathbf{X}\mathbf{X}}$. Likewise, it is unlikely that it is approximately aligned with the *last* principal component.

For the more general DAG shown of Figure 1 we again assume that $\mathbf{a}$ has generic orientation with respect to the eigenspaces of $\Sigma_{\mathbf{X}\mathbf{X}}$ and, in addition, that $\mathbf{b}$ has generic orientation with respect to the eigenspaces of $\Sigma_{\mathbf{E}\mathbf{E}}$.

To provide a first intuition about why the orientation of the resulting vector $\hat{\mathbf{a}}$ of regression coefficients is no longer ‘generic’ relative to $\Sigma_{\mathbf{X}\mathbf{X}}$ as a result of confounding, we show two somehow opposite extreme cases where $\hat{\mathbf{a}}$ gets aligned with the first and the last principal component of $\mathbf{X}$, respectively. To this end, we consider the purely confounded case where $\mathbf{a} = 0$ and thus $\hat{\mathbf{a}} = c \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}$.

$\hat{\mathbf{a}}$ aligned with the first eigenvector of $\Sigma_{\mathbf{X}\mathbf{X}}$: Let $\Sigma_{\mathbf{E}\mathbf{E}}$ be the identity matrix $\mathbf{I}$. Since $\Sigma_{\mathbf{E}\mathbf{E}}$ is then rotation invariant, $\mathbf{b}$ has certainly ‘generic orientation’ relative to $\Sigma_{\mathbf{E}\mathbf{E}}$, according to any reasonable sense of ‘generic orientation’. Then $\mathbf{b}$ does not have generic orientation relative to $\Sigma_{\mathbf{X}\mathbf{X}} = \mathbf{I} + \mathbf{b}\mathbf{b}^T$, because it is the unique eigenvector of the latter with maximal eigenvalue. In other words, $\mathbf{b}$ is aligned with the first principal component of $\mathbf{X}$. Then, $\hat{\mathbf{a}}$ is also aligned with the same principal component since it is a multiple of $\mathbf{b}$ due to $\hat{\mathbf{a}} = \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b} = (\mathbf{I} + \mathbf{b}\mathbf{b}^T)^{-1} \mathbf{b}$. Note that one also gets close to this scenario when the spectral gaps of $\Sigma_{\mathbf{E}\mathbf{E}}$ are small⁴ compared to the norm of $\mathbf{b}\mathbf{b}^T$ and to the eigenvalues of $\Sigma_{\mathbf{E}\mathbf{E}}$.

$\hat{\mathbf{a}}$ close to the last eigenvector of $\Sigma_{\mathbf{X}\mathbf{X}}$: Let the spectral gaps between adjacent eigenvalues of $\Sigma_{\mathbf{E}\mathbf{E}}$ be much larger than the norm of $\mathbf{b}\mathbf{b}^T$. Then adding $\mathbf{b}\mathbf{b}^T$ changes the eigenspaces only slightly [15]. Hence, if $\mathbf{b}$ has a generic orientation relative to $\Sigma_{\mathbf{E}\mathbf{E}}$, it is still generic relative to $\Sigma_{\mathbf{X}\mathbf{X}}$. Multiplying $\mathbf{b}$ with $\Sigma_{\mathbf{X}\mathbf{X}}^{-1}$ then generates a vector that has stronger coefficients in the small eigenvalues of $\Sigma_{\mathbf{X}\mathbf{X}}$. If the smallest eigenvalue of $\Sigma_{\mathbf{X}\mathbf{X}}$ is much smaller than the others, the orientation of $\hat{\mathbf{a}}$ gets arbitrarily close to the smallest eigenvector.

For general $\Sigma_{\mathbf{E}\mathbf{E}}$, where the gaps between the eigenvalues are neither tiny nor huge compared to the norm of $\mathbf{b}\mathbf{b}^T$, the orientation of $\hat{\mathbf{a}}$ changes in a more sophisticated way that heavily depends on the structure of the spectrum of $\Sigma_{\mathbf{E}\mathbf{E}}$. This will be analyzed in Section 4.

3.2 Defining ‘generic orientation’ via the induced spectral measure

We start with some notation and terminology and formally introduce two measures which have quite simple intuitive meanings. For $d \times d$ matrices, we introduce the renormalized trace⁵

$$\tau(A) := \frac{1}{d} \text{tr}(A).$$

For notational convenience, we will assume that the spectra of all matrices are non-degenerate throughout the paper, i.e., all eigenvalues are different. Every symmetric matrix A thus admits a unique decomposition

$$A = \sum_{j=1}^d \lambda_j \phi_j \phi_j^T, \quad (9)$$

where $\lambda_1 > \dots > \lambda_d$ and $\phi_1, \dots, \phi_d$ denote the corresponding eigenvectors of unit length. Every A uniquely defines a measure on $\mathbb{R}$, namely the distribution of eigenvalues, formally given as follows:

Definition 2. (tracial spectral measure) Let A be a real symmetric matrix with non-degenerate spectrum. Then the tracial spectral measure μ_A^τ of A is the discrete measure on $\mathbb{R}$ given by the uniform distribution over its eigenvalues $\lambda_1, \dots, \lambda_d$, i.e.,

⁴ In this limit, the first eigenvector of $(\Sigma_{\mathbf{E}\mathbf{E}} + \mathbf{b}\mathbf{b}^T)$ is almost aligned with $\mathbf{b}$. On the other hand, the vector $(\Sigma_{\mathbf{E}\mathbf{E}} + \mathbf{b}\mathbf{b}^T)^{-1} \mathbf{b}$ is almost aligned with $\mathbf{b}$ because, as will be shown later, it is a multiple of $\Sigma_{\mathbf{E}\mathbf{E}}^{-1} \mathbf{b}$ which is almost a multiple of $\mathbf{b}$ whenever the spectral gaps of $\Sigma_{\mathbf{E}\mathbf{E}}$ are negligible compared to the spectral values of $\Sigma_{\mathbf{E}\mathbf{E}}^{-1}$.

⁵ Note that τ is known as ‘tracial state’ in the context of functional analysis [16].

$$\mu_A^\tau := \frac{1}{d} \sum_{j=1}^d \delta_{\lambda_j},$$

where δ_s denotes the point measure on s for some $s \in \mathbb{R}$.

By elementary spectral theory of symmetric matrices [17], we have:

Lemma 1. (expectation for tracial spectral measure) *The expectation of any function $f : \{\lambda_1, \dots, \lambda_d\} \rightarrow \mathbb{R}$ with respect to the tracial measure is given as follows:*

$$\int f(s) d\mu_A^\tau(s) = \tau(f(A)).$$

While the spectral measure is a property of a matrix alone, the following measure describes the relation between a matrix and a vector:

Definition 3. (vector-induced spectral measure) *Let A be a symmetric matrix and λ_j, ϕ_j be defined by eq. (9). For arbitrary $\psi \in \mathbb{R}^d$, the (unnormalized) spectral measure induced by ψ on A , denoted by $\mu_{A,\psi}$, is given by*

$$\mu_{A,\psi}(S) = \sum_{j \text{ with } \lambda_j \in S} \langle \psi, \phi_j \rangle^2,$$

for any measurable set $S \subset \mathbb{R}$.

For each set of eigenvalues of A , the measure describes the squared length of the component of ψ that lies in the respective eigenspace of A . Accordingly, we have the following normalization condition:

$$\mu_{A,\psi}(\mathbb{R}) = \|\psi\|^2. \quad (10)$$

In analogy to Lemma 1 we obtain:

Lemma 2. (expectations for vector-induced spectral measure) *The expectation of any function f on the spectrum of A with respect to $\mu_{A,\psi}$ is given as follows:*

$$\int f(s) d\mu_{A,\psi}(s) = \langle \psi, f(A)\psi \rangle.$$

To deal with the above measures in numerical computations, each measure will be represented by two vectors: first, one vector $\lambda := (\lambda_1, \dots, \lambda_d)$ listing its support (with the convention $\lambda_1 \geq \dots \geq \lambda_d$) and second, the vector $w := (w_1, \dots, w_d)$ listing the corresponding weights. For the tracial spectral measures, λ is just the list of eigenvalues and $w = (1/d, \dots, 1/d)$ is the uniform distribution on these d points. For spectral measures induced by a vector, λ is still the list of eigenvalues but now w describes the squared coefficients of the vector with respect to the eigenvector decomposition.

To understand our algorithm described later, it is helpful to note that using the eigenvectors ϕ_j of a matrix A one can easily construct a vector ψ that induces the tracial measure:

$$\psi := \frac{1}{\sqrt{d}} \sum_{j=1}^d \phi_j.$$

Then we have

$$\mu_{A,\psi} = \mu_A^\tau. \quad (11)$$

The crucial idea behind our detection of confounding is the insight that $\mu_{A,\psi}$ is close to $\|\psi\|^2 \mu_A^\tau$ for any ‘typical’ choices of ψ , which is made precise by the following result shown in Section 5.

Lemma 3. Let $(A_d)_{d \in \mathbb{N}}$ with $\|A_d\| \leq a$ be a sequence of symmetric matrices whose spectral measure converges weakly to some probability measure μ^∞ , i.e.,

$$\mu_{A_d}^T \rightarrow \mu^\infty. \quad (12)$$

Let $(\mathbf{c}_d)_{d \in \mathbb{N}}$ with $\mathbf{c}_d \in \mathbb{R}^d$ be randomly drawn from the sphere of radius r . Then

$$\mu_{A_d, \mathbf{c}_d} \rightarrow r^2 \mu^\infty,$$

weakly in probability.

The idea of this paper is that we expect $\mu_{\Sigma_{XX}, \hat{\mathbf{a}}}$ to be close to $\mu_{\Sigma_{XX}, \mathbf{a}}$ in the absence of confounding. Moreover, we show that confounding perturbs $\mu_{\Sigma_{XX}, \hat{\mathbf{a}}}$ in a characteristic way, given that the model parameters are ‘typical’. To study properties of $\mu_{\Sigma_{XX}, \hat{\mathbf{a}}}$ for ‘typical’ choices of the model parameters $\mathbf{a}, \mathbf{b}$, we now define the following sequence of models for increasing dimension d :

Covariance matrix of the noise of X: Let $(\Sigma_{EE}^d)_{d \in \mathbb{N}}$ be a uniformly bounded sequence of positive semi-definite $d \times d$ -matrices such that their tracial spectral measures converge weakly to some probability measure μ^∞ (describing the asymptotic distribution of eigenvalues).

Vector of causal structure coefficients a: Let $(\mathbf{a}_d)_{d \in \mathbb{N}}$ be a sequence of vectors in $\mathbb{R}^d$ drawn uniformly at random from a sphere of fixed radius $r_{\mathbf{a}}$.

Vector of confounding structure coefficients b: Let $(\mathbf{b}_d)_{d \in \mathbb{N}}$ be a sequence of vectors in $\mathbb{R}^d$ drawn uniformly at random (independently of $\mathbf{a}_d$) from a sphere of fixed radius $r_{\mathbf{b}}$. Let c be fixed for all d .

Then $\Sigma_{XX}^d = \Sigma_{EE}^d + \mathbf{b}_d \mathbf{b}_d^T$ and $\hat{\mathbf{a}}_d = \mathbf{a}_d + c(\Sigma_{XX}^d)^{-1} \mathbf{b}_d$ and we have the following result that will be shown in Section 5:

Theorem 1. (asymptotic spectral measure for rotation-invariant generating model) For the above generating model the below approximations eqs. (16), (17), and (18) are asymptotically justified in the sense that

$$\mu_{\Sigma_{XX}^d, \mathbf{a}_d} \rightarrow r_{\mathbf{a}}^2 \mu^\infty \quad (\text{weakly in probability}) \quad (13)$$

$$\mu_{\Sigma_{EE}^d, \mathbf{b}_d} \rightarrow r_{\mathbf{b}}^2 \mu^\infty \quad (\text{weakly in probability}) \quad (14)$$

$$\mu_{\Sigma_{XX}^d, \mathbf{a}_d + c(\Sigma_{XX}^d)^{-1} \mathbf{b}_d} - (\mu_{\Sigma_{XX}^d, \mathbf{a}_d} + \mu_{\Sigma_{XX}^d, c(\Sigma_{XX}^d)^{-1} \mathbf{b}_d}) \rightarrow 0 \quad (\text{weakly in probability}) \quad (15)$$

Intuitively, eq. (13) just states that the causal part of $\mathbf{a}$ induces, together with Σ_{XX} , a spectral measure that is close to the tracial measure up to a normalization factor because $\mathbf{a}$ has ‘generic’ orientation relative to $\Sigma_{XX} = \Sigma_{EE} + \mathbf{b} \mathbf{b}^T$ since $\mathbf{a}$ is chosen independently of Σ_{EE} and $\mathbf{b}$. On the other hand, $\mathbf{b}$ is chosen independently of Σ_{EE} , and thus the spectral measure induced by Σ_{EE} and $\mathbf{b}$ is close to the corresponding tracial measure up to normalization, as shown by eq. (14). Equation (15) states that $\mu_{\Sigma_{XX}, \mathbf{a} + \Sigma_{XX}^{-1} \mathbf{b}}$ almost decomposes into $\mu_{\Sigma_{XX}, \mathbf{a}}$ plus $\mu_{\Sigma_{XX}, \Sigma_{XX}^{-1} \mathbf{b}}$, that is, the spectral measure induced by the causal vector $\mathbf{a}$ and its perturbation $\Sigma_{XX}^{-1} \mathbf{b}$. This is useful because eq. (13) describes the asymptotic behavior of $\mu_{\Sigma_{XX}, \mathbf{a}}$. To understand the asymptotic of $\mu_{\Sigma_{XX}, \Sigma_{XX}^{-1} \mathbf{b}}$, on the other hand, we will later use eq. (14).

Assuming a rotation invariant generating model may appear as a too strong assumption for practical purposes. It should therefore be noted that much weaker assumptions would probably also yield the same approximate identities for high dimensions. Possible quantitative versions of Lemma 3 for finite dimensions could show that the overwhelming majority of vectors on the surface of the unit sphere induce spectral measures that are close to the tracial one. Such quantitative versions could provide some insights on the robustness of our method regarding violation of the strong model assumptions (since also a large class of non-rotation invariant priors then yield the same conclusions), but this goes beyond the scope of this paper. After all, the performance on real data cannot be predicted from purely theoretical arguments.

Motivated by the above asymptotic results, we assume to be d large enough that left and right hand side of eqs. (13) to (15) are approximately satisfied:

Postulate 1(generic orientation of vectors) If eqs. (1) and (2) are structural equations corresponding to the causal DAG in Figure 1, and d is large, then:

(I) The vector $\mathbf{a}$ has generic orientation relative to $\Sigma_{\mathbf{X}\mathbf{X}}$ in the sense that

$$\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \mathbf{a}} \approx \mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^T \|\mathbf{a}\|^2. \quad (16)$$

(II) The vector $\mathbf{b}$ has generic orientation relative to $\Sigma_{\mathbf{E}\mathbf{E}}$ in the sense that

$$\mu_{\Sigma_{\mathbf{E}\mathbf{E}}, \mathbf{b}} \approx \mu_{\Sigma_{\mathbf{E}\mathbf{E}}}^T \|\mathbf{b}\|^2. \quad (17)$$

(III) The vector $\mathbf{a}$ is generic relative to $\mathbf{b}$, $\Sigma_{\mathbf{E}\mathbf{E}}$ in the sense that

$$\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \mathbf{a} + \mathbf{c} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}} \approx \mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \mathbf{a}} + \mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \mathbf{c} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}}. \quad (18)$$

We therefore built our algorithm in Section 5 upon the postulates only instead of directly using the generating model.

Note that eqs. (16), (17), and (18) only hold if the $\approx$ -sign is interpreted in a sufficiently loose sense. After all, the differences between both sides of the above equations converge *weakly* to zero. They are *not* close, for instance, with respect to total variation distance. Hence, the measures are similar in the same sense as two empirical distributions with large sample size are similar when they are independently sampled from the same distribution.

Related work: We first describe the relation of the above ideas to those underlying the so-called *Trace Method* [18, 19], which is, to the best of our knowledge, the work in the literature that is closest to the present one. Let $\mathbf{X}$ and $\mathbf{Y}$ be vector-valued variables with values in $\mathbb{R}^d$ and $\mathbb{R}^m$, respectively. Assume $\mathbf{X}$ influences $\mathbf{Y}$ via the linear model

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{F},$$

where $\mathbf{A}$ is an $m \times d$ matrix and $\mathbf{F}$ a noise variable of dimension m . Then $\mathbf{A}$ and $\Sigma_{\mathbf{X}\mathbf{X}}$ satisfy the trace condition

$$\frac{1}{m} \text{tr}(\mathbf{A} \Sigma_{\mathbf{X}\mathbf{X}} \mathbf{A}^T) \approx \frac{1}{d} \text{tr}(\Sigma_{\mathbf{X}\mathbf{X}}) \frac{1}{m} \text{tr}(\mathbf{A} \mathbf{A}^T). \quad (19)$$

For $m = 1$, the $1 \times d$ matrix $\mathbf{A}$ is given by the row vector $\mathbf{a}^T$. Then $\mathbf{A}\mathbf{X} = \langle \mathbf{a}, \mathbf{X} \rangle$ and eq. (19) turns into

$$\langle \mathbf{a}, \Sigma_{\mathbf{X}\mathbf{X}} \mathbf{a} \rangle \approx \frac{1}{d} \text{tr}(\Sigma_{\mathbf{X}\mathbf{X}}) \langle \mathbf{a}, \mathbf{a} \rangle. \quad (20)$$

In terms of the spectral measures, eq. (20) reads

$$\int s d\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \mathbf{a}}(s) \approx \int s d\mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^T(s) \|\mathbf{a}\|^2.$$

Hence, eq. (19) postulates that the *first moments* of two measures on the left and the right of eq. (16) almost coincide, while our method also accounts for *higher order* moments which the Trace Condition ignores. As already sketched in [19], the Trace Condition (19) is closely related to the concept of *free independence* in free probability theory [20]. In the appendix we will explain why eqs. (16), (17), (18) are also related to free independence in spirit, although there is no straightforward way to apply those concepts here.

We should also mention [21], which is remotely related to this paper. The authors study a large number d of observed variables which are related by a sparse graphical model as well as by some latent variables influencing a large subset of the observed variables and provide conditions under which the latent part and the

graphical model part can be distinguished. On a high level, this is related to our work because the methods also employs effects in high dimensions to detect latent variables.

4 Description of the method

4.1 Constructing typical spectral measures for given parameter values

The main result of this section states that asymptotically, $\mu_{\Sigma_{XX}, \hat{\mathbf{a}}}$ depends only on 3 parameters when Σ_{XX} is given: $\|\hat{\mathbf{a}}\|^2$, β , η . The first one is directly observable, hence we define a two-parametric family of normalized (i.e. probability) measures $\nu_{\beta, \eta}$ such that for large d with high probability

$$\mu_{\Sigma_{XX}, \hat{\mathbf{a}}} \approx \|\hat{\mathbf{a}}\|^2 \nu_{\beta, \eta},$$

for appropriate choices of β , η . We first describe the construction of $\nu_{\beta, \eta}$:

1. **Causal part:** this part describes the spectral measure that would be obtained in the absence of confounding. It is induced by $\mathbf{a}$ and Σ_{XX} . According to eq. (16), it is approximated by the uniform distribution over the spectrum of Σ_{XX} , i.e., the tracial measure introduced in Definition 2. We therefore define:

$$\nu^{\text{causal}} := \mu_{\Sigma_{XX}}^{\tau}.$$

2. **Confounding part:** we now approximate the spectral measure induced by the vector $\Sigma_{XX}^{-1}\mathbf{b}$ and Σ_{XX} . We will justify the construction later after we have described all its steps. We first define the matrix $M_X := \text{diag}(\nu_1^X, \dots, \nu_d^X)$, where ν_j^X are the eigenvalues of Σ_{XX} in decreasing order. Then we define a rank-one perturbation of M_X by

$$T := M_X + \eta \mathbf{g} \mathbf{g}^T, \quad (21)$$

where $\mathbf{g}$ is the vector $\mathbf{g} := (1, \dots, 1)^T / \sqrt{d}$. We then compute the spectral measure induced by the vector $T^{-1}\mathbf{g}$ and T and define

$$\nu_{\eta}^{\text{confounded}} := \frac{1}{\|T^{-1}\mathbf{g}\|^2} \mu_{T, T^{-1}\mathbf{g}}. \quad (22)$$

3. **Mixing both contributions:** Finally, $\nu_{\beta, \eta}$ is a convex sum of the causal and the confounded part where the mixing weight of the latter is given by the confounding strength:

$$\nu_{\beta, \eta} := (1 - \beta) \nu^{\text{causal}} + \beta \nu_{\eta}^{\text{confounded}}.$$

We now explain Step 2 in the above construction. According to Postulate 3.2, $\mathbf{b}$ has generic orientation relative to Σ_{EE} in the sense of eq. (17). With respect to its eigenbasis, Σ_{EE} reads $M_E := \text{diag}(\nu_1^E, \dots, \nu_d^E)$, where ν_j^E are the eigenvalues of Σ_{EE} . Of course we don't know the vector $\mathbf{b}$, neither do we know the coordinates of $\mathbf{b}$ with respect to this basis. Remarkably, it turns out that knowing that $\mathbf{b}$ is generic relative to Σ_{EE} is enough because then we can replace $\mathbf{b}$ with a vector that is 'particularly generic', namely $\mathbf{g}$. This vector satisfies

$$\mu_{M_E, \mathbf{g}} = \mu_{M_E}^{\tau}, \quad (23)$$

(see eq. (11)) while asymptotically the overwhelming majority of unit vectors satisfy eq. (23) *approximately*. Therefore, an appropriate multiple of $\mathbf{g}$ nicely mimics the behavior of generic vectors. Accordingly, we can approximate the spectral measure induced by the vector $(\Sigma_{EE} + \mathbf{b}\mathbf{b}^T)^{-1}\mathbf{b}$ and the matrix $(\Sigma_{EE} + \mathbf{b}\mathbf{b}^T)$ by the spectral measure induced by $(M_E + \eta \mathbf{g} \mathbf{g}^T)^{-1}\sqrt{\eta} \mathbf{g}$ and $M_E + \eta \mathbf{g} \mathbf{g}^T$. Unfortunately, this construction would need the eigenvalues of Σ_{EE} , which cannot be computed from observing $\mathbf{X}$, $\mathbf{Y}$ alone. Asymptotically, however, the

difference between the spectra of $\Sigma_{\mathbf{X}\mathbf{X}}$ and $\Sigma_{\mathbf{E}\mathbf{E}}$ do not matter and we can replace M_E with M_X . This step will be justified later using the fact that for large d we have

$$\mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^\tau \approx \mu_{\Sigma_{\mathbf{E}\mathbf{E}}}^\tau, \quad (24)$$

which is made more precise by the following result:

Lemma 4. (tracial measures are close) *For any interval $[r, l]$ we have*

$$|\mu_X^\tau[r, l] - \mu_E^\tau[r, l]| \leq d.$$

Proof. If $v_1^E > \dots > v_d^E$ denote the eigenvalues of $\Sigma_{\mathbf{E}\mathbf{E}}$, then the eigenvalues v_j^X of $\Sigma_{\mathbf{X}\mathbf{X}} = \Sigma_{\mathbf{E}\mathbf{E}} + \mathbf{b}\mathbf{b}^T$ satisfy

$$v_j^X \in [v_j^E, v_{j-1}^E] \quad \forall j \geq 2,$$

by Theorem 10.2 in [22]. Hence the number of eigenvalues in a given interval can differ by 1 at most. $\square$

We now describe the main theoretical result of this article:

Theorem 2. (convergence to two-parametric family) *Let $(\Sigma_{\mathbf{X}\mathbf{X}}^d)$ be a sequence of covariance matrices whose distribution of eigenvalues (i.e. the tracial spectral measure) converges weakly to some probability measure μ^∞ supported by a compact interval in $\mathbb{R}_0^+$. Assume, moreover, we are given sequences of vectors $(\mathbf{a}_d)$ and $(\mathbf{b}_d)$ with $\|\mathbf{a}_d\| = r_{\mathbf{a}}$ and $\|\mathbf{b}_d\| = r_{\mathbf{b}}$ (with $r_{\mathbf{a}}, r_{\mathbf{b}}$ fixed) and fixed c , such that eqs. (13)–(15) hold (recall $\Sigma_{\mathbf{E}\mathbf{E}}^d = \Sigma_{\mathbf{X}\mathbf{X}}^d - \mathbf{b}_d\mathbf{b}_d^T$). Then $v_{\beta, \eta}$ approximates $\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \hat{\mathbf{a}}}$ up to normalization in the sense that*

$$\frac{1}{\|\hat{\mathbf{a}}_d\|^2} \mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \hat{\mathbf{a}}_d}^d - v_{\beta, \eta}^d \rightarrow 0 \quad (\text{weakly in probability}),$$

where $\eta := r_{\mathbf{b}}^2$ and the confounding strength β can be derived from μ^∞ , $r_{\mathbf{b}}$, and c by equations specified later.

Hence, the theorem states that whenever Postulate 3.2 holds with sufficient accuracy and for sufficiently high dimension, then $v_{\beta, \eta}$ is a good approximation for the induced spectral measure. We will prove Theorem 2 in Section 5.

Apart from this weak convergence result we also know that the measures $\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \hat{\mathbf{a}}_d}^d$ and $v_{\beta, \eta}^d$ have precisely the same support for any d because, by construction, $v_{\beta, \eta}^d$ is also supported by the spectrum of $\Sigma_{\mathbf{X}\mathbf{X}}$. This enables to conveniently represent both measures by vectors whose entries describe the weight of the corresponding eigenvalue.

Two limiting cases as examples: To get a further intuition about $v_{\beta, \eta}$ let it describe for the two motivating limiting cases in Subsection 3.1, both with pure confounding, i.e., $\beta = 1$. Following Theorem 2, we have to choose $\eta = \|\mathbf{b}\|^2$ to mimic $\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \hat{\mathbf{a}}}$ in the limit of $d \rightarrow \infty$. The case where $\hat{\mathbf{a}}$ is almost aligned with the first eigenvector corresponds to $v_{\beta, \eta}$ having the majority of its weight on the first eigenvalue, while the second case corresponds to $v_{\beta, \eta}$ for which the weights for small eigenvectors are more dominant, as we will see now.

For the first case $\Sigma_{\mathbf{E}\mathbf{E}} = \mathbf{I}$, the spectrum of $\Sigma_{\mathbf{X}\mathbf{X}} = \mathbf{I} + \mathbf{b}\mathbf{b}^T$ reads $1 + \eta, 1, \dots, 1, 1$, that is, $M_X = \text{diag}(1 + \eta, 1, \dots, 1)$. We thus have

$$T = \mathbf{I} + \eta(\mathbf{e}_1\mathbf{e}_1^T + \mathbf{g}\mathbf{g}^T),$$

where $\mathbf{e}_1 := (1, 0, \dots, 0)^T$. The subspace spanned by $\mathbf{e}_1$ and $\mathbf{g}$ is T -invariant. The restriction T_2 of T to this 2-dimensional subspace is asymptotically (for d to infinity) given by $(\eta + 1)\mathbf{I}$ because $\mathbf{e}_1$ and $\mathbf{b}$ become orthogonal. Therefore,

$$v_{\beta, \eta} = \frac{1}{\|T^{-1}\mathbf{g}\|^2} \mu_{T, T^{-1}\mathbf{g}} = \frac{1}{\|T_2^{-1}\mathbf{g}\|^2} \mu_{T_2, T_2^{-1}\mathbf{g}} \xrightarrow{d \rightarrow \infty} \delta_{\eta+1},$$

that is, the weights of $T^{-1}\mathbf{g}$ tend to zero for all eigenvectors other than $\mathbf{v}_1^X$, in agreement with our expectation when we recall the remarks in Subsection 3.1

In the second case, let all eigenvalues $\nu_1^X, \dots, \nu_d^X$ of $\Sigma_{\mathbf{X}\mathbf{X}}$ be distinct and η be small compared to the spectral gaps $\nu_j^X - \nu_{j+1}^X$. The perturbation of $M^X = \text{diag}(\nu_1^X, \dots, \nu_d^X)$ with $\eta \mathbf{g}\mathbf{g}^T$ changes the eigenvectors and eigenvalues only slightly, that is, $T \approx M^X$ and thus

$$\mu_{T, T^{-1}\mathbf{g}} \approx \mu_{M^X, (M^X)^{-1}\mathbf{g}} = \frac{1}{d} \sum_{j=1}^d (\nu_j^X)^{-2} \delta_{\nu_j^X}. \quad (25)$$

Accounting for the normalization factor then yields

$$\nu_{\beta, \eta} = \frac{1}{\|T^{-1}\mathbf{g}\|^2} \mu_{T, T^{-1}\mathbf{g}} \approx \frac{1}{\sum_{j=1}^d (\nu_j^X)^{-2}} \sum_{j=1}^d (\nu_j^X)^{-2} \delta_{\nu_j^X}.$$

Hence, as expected due to the remarks in Subsection 3.1, small eigenvalues get higher weights than the higher ones (due to the weighting factor $(\nu_j^X)^{-2}$).

4.2 Description of the algorithm

To estimate the confounding parameters we just take the element in the family $(\nu_{\beta, \eta})_{\beta \in [0, 1], \eta \in [0, \nu_1^X]}$ that is closest to $\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \hat{\mathbf{a}}}$, but we have to choose an appropriate distance measure. Since Theorem 2 only guarantees weak convergence, l_1 or l_2 distances between the weight vectors would be inappropriate. Instead, we have decided to smoothen the measures by a Gaussian kernel and then compare the l_1 distance. As kernel bandwidth we have worked with $\sigma := 0.2(\nu_1^X - \nu_d^X)$. This heuristic choice is motivated by the idea that the kernel should smoothen over a significant fraction of the eigenvalues. For dimension $d = 10$, for instance, the bandwidth 0.2 thus is two times the average distance between adjacent eigenvalues.

Accordingly, we define a distance between two weight vectors w and w' by

$$D(w, w') := \|K(w - w')\|_1, \quad (26)$$

where K denotes the kernel smoothing matrix with entries

$$K(i, j) := e^{-\frac{(\nu_i^X - \nu_j^X)^2}{2\sigma^2}}.$$

Based on these findings, we describe how to estimate β in Algorithm 4.2. The code as well as all the data sets used in this paper and instructions on how to reproduce the results can be found at <http://webdav.tuebingen.mpg.de/causality/>.

Algorithm 1 Estimating the strength of confounding

- 1: **Input:** I.i.d. samples from $P(\mathbf{X}, Y)$.
 - 2: Compute the empirical covariance matrices $\Sigma_{\mathbf{X}\mathbf{X}}$ and $\Sigma_{\mathbf{X}Y}$
 - 3: Compute the regression vector $\hat{\mathbf{a}} := \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}Y}$
 - 4: PHASE 1: Compute the spectral measure $\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \hat{\mathbf{a}}}$
 - 5: Compute eigenvalues $\nu_1^X > \dots > \nu_d^X$ and the corresponding eigenvectors $\phi_1, \dots, \phi_d$ of $\Sigma_{\mathbf{X}\mathbf{X}}$
 - 6: Compute the weights $w'_j = \langle \hat{\mathbf{a}}, \phi_j \rangle^2$ and then the normalized weights $w_j := w'_j / \sum_j w'_j$.
 - 7: PHASE 2: find the parameter values $\hat{\beta}, \hat{\eta}$ that minimize the distance $D(w, w^{\beta, \eta})$ with D defined by eq. (26), where $w^{\beta, \eta}$ denotes the weight vector of the measure $\nu_{\beta, \eta}$.
 - 8: **Output:** Estimated confounding strength $\hat{\beta}$
-

Since the pseudocode does not describe how to compute the weight vector $w^{\beta,\eta}$ we provide this missing detail now. First compute the matrix T as defined by eq. (21) and compute its eigenvectors $\psi_1, \dots, \psi_d$. Then we compute the vector $\mathbf{v} := T^{-1}\mathbf{g}/\|T^{-1}\mathbf{g}\|$. The squared coefficients of $\mathbf{v}$ with respect to the eigenvector basis $\psi_1, \dots, \psi_d$ describe the weights of $v_\eta^{\text{confounded}}$, see eq. (22). To obtain the weights of $v_{\beta,\eta}$ we need to add the contribution of v^{causal} and finally obtain the weights

$$w_j^{\beta,\eta} := \frac{1}{d}(1 - \beta) + \beta \langle \mathbf{v}, \phi_j \rangle^2.$$

4.3 Remark on normalization

So far we have ignored the case where the variables $X_1, \dots, X_d$ refer to quantities that are measured in different units. If, for instance, X_1 denotes the temperature and X_2 the traffic density (which both influence the NO_x concentration in the air), the relative scale of their numeric values depend on the units one choses. Another related issue is that all X_j refer to the same unit, but the variance of one of the variables is overwhelmingly larger than the variance of the others, which results in a covariance matrix whose rank is basically one. A pragmatic and straightforward solution for both issues is to normalize all variables X_j as preprocessing step. Actually, this is obviously in conflict with the justification of the method because normalization jointly changes Σ_{XX} and $\mathbf{a}$, which spoils the idea of ‘independence’. In our limited simulation studies, however, the results turned out to be reasonably robust with respect to normalizing all X_j . Here, robustness is only meant in the sense that the performance over a large number of runs looked almost the same. For every single experiment, however, the estimated values $\hat{\beta}$ can significantly differ by the amount of uncertainty that is inherent to our method anyway. Due to the lack of theoretical justification, we recommend to avoid normalization if possible and remain skeptical about the results with normalized data.

5 Proofs of asymptotic statements

5.1 Proof of Lemma 3

It is sufficient to show the statement for $r = 1$ because the measure obviously scales quadratically in r . Since the support of $\mu_{A_d, \mathbf{c}_d}$ is contained in the compact interval $[-a, a]$, it is sufficient to show convergence of all moments, i.e., that for every $k \in \mathbb{N}$

$$\int s^k d\mu_{A_d, \mathbf{c}_d}(s) \rightarrow \int s^k d\mu^\infty(s),$$

in probability. To this end, we drop most indices d and consider fixed dimension d . To generate a random unit vector $\mathbf{c}$, we first take d independent Gaussian random variables $C_j \sim N(0, 1/\sqrt{d})$ and define the j th coefficient of $\mathbf{c}$ by $C_j / \sum_{i=1}^d C_i^2$. Let $A = \text{diag}(\lambda_1, \dots, \lambda_d)$ without loss of generality. Then the k th moment reads:

$$\int s^k d\mu_{A, \mathbf{c}}(s) = \frac{\sum_{j=1}^d \lambda_j^k C_j^2}{\sum_{j=1}^d C_j^2} =: \frac{\Lambda_d}{\Gamma_d}. \quad (27)$$

One easily checks

$$\mathbf{E}[\Lambda_d] = \sum_{j=1}^d \frac{\lambda_j^k}{d} = \int s^k d\mu_{A_d}^r(s).$$

Moreover, since all C_j are independent and because squared standard Gaussians have variance 2, we have

$$\text{Var}[\Lambda_d] = \sum_{j=1}^d \lambda_j^{2k} \text{Var}[C_j^2] = \frac{2}{d^2} \sum_{j=1}^d \lambda_j^{2k} \leq \frac{2}{d} a^{2k},$$

where we used $\lambda_j \leq a$. By Chebyshev's inequality, the probability for large deviations from the mean can be bounded by

$$\Pr \left\{ \left| \Lambda_d - \int s^k d\mu_A^\tau(s) \right| \geq \epsilon \right\} \leq \frac{2}{d\epsilon^2} a^{2k}. \quad (28)$$

Then we get

$$\begin{aligned} & \int s^k d\mu_{A_d, c_d}(s) - \int s^k d\mu^\infty(s) = \frac{\Lambda_d}{\Gamma_d} - \int s^k d\mu^\infty(s) \\ &= \frac{\Lambda_d - \int s^k d\mu_{A_d}^\tau(s)}{\Gamma_d} + \left(\frac{\int s^k d\mu_{A_d}^\tau(s)}{\Gamma_d} - \int s^k d\mu^\infty(s) \right). \end{aligned} \quad (29)$$

We have $\Gamma_d \rightarrow 1$ almost surely by the strong law of large numbers. Due to eq. (28), the quotient on the left of expression (29) converges to zero in probability. Thus, expression (29) converges to zero in probability due to the assumption $\int s^k d\mu_{A_d}^\tau(s) \rightarrow \int s^k d\mu^\infty(s)$.

5.2 Proof of Theorem 1

The difference between the left and the right hand sides of eq. (16) converges weakly to zero in probability due to Lemma 3 because $\mathbf{a}_d$ is drawn independently of $(\Sigma_{\mathbf{X}\mathbf{X}})_d = \Sigma_{EE}^d + b_d b_d^T$, since $\mathbf{a}_d$ and $\mathbf{a}_d$ are drawn independently by assumption. Statement (17) follows even more directly from Lemma 3 because $\mathbf{b}_d$ is drawn independently of $(\Sigma_{EE})_d$.

To show that the difference between the left and the right hand side of eq. (18) converges weakly to zero in probability, we recall that it is sufficient to show that expectations of bounded continuous functions converge. For any measurable function $g : \mathbb{R} \rightarrow \mathbb{R}$ the difference of expectations reads:

$$\begin{aligned} & \int g d\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \mathbf{a} + \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}} - \int g d\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \mathbf{a}} - \int g d\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}} \\ &= \langle \mathbf{a} + \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}, g(\Sigma_{\mathbf{X}\mathbf{X}})(\mathbf{a} + \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}) \rangle - \langle \mathbf{a}, g(\Sigma_{\mathbf{X}\mathbf{X}}) \mathbf{a} \rangle - \langle \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b}, g(\Sigma_{\mathbf{X}\mathbf{X}}) \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b} \rangle \\ &= 2 \langle \mathbf{a}, g(\Sigma_{\mathbf{X}\mathbf{X}}) \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b} \rangle. \end{aligned}$$

Hence, we only have to show that

$$\langle \mathbf{a}, g(\Sigma_{\mathbf{X}\mathbf{X}}) \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{b} \rangle \rightarrow 0 \quad (30)$$

in probability. Note that this already follows from the fact that $\mathbf{a}$ is chosen independently from the vector on the right hand side in the inner product eq. (30), due to the following elementary result:

Lemma 5. (asymptotic orthogonality) Let $\mathbf{v}_d \in \mathbb{R}^d$ be a sequence of vectors. Let $\mathbf{c}_d \in \mathbb{R}^d$ be drawn uniformly at random from the unit sphere. Then,

$$\langle \mathbf{v}_d, \mathbf{c}_d \rangle^2 \rightarrow 0,$$

almost surely.

Proof. Without loss of generality, assume $\mathbf{v} = (v, 0, \dots, 0)^T$ with $v \in \mathbb{R}$. Generate the entries c_j of $\mathbf{c}$ by first taking independent standard Gaussians C_j and renormalizing afterwards. Then

$$\lim_{d \rightarrow \infty} \langle \mathbf{v}_d, \mathbf{c}_d \rangle^2 = \lim_{d \rightarrow \infty} \frac{1}{d} \frac{C_1^2}{\sum_{j=1}^d C_j^2} = 0,$$

because $\frac{1}{d} \sum_{j=1}^d C_j^2$ converges to $\mathbf{E}[C_j^2] = 1$ almost surely due to the law of large numbers. $\square$

5.3 Proof of Theorem 2

We first need some definitions and tools. The following one generalizes Definition 3 to infinite-dimensional Hilbert spaces (see [17] for spectral theory of self-adjoint operators):

Definition 4. (vector-induced spectral measure) Let $\mathcal{H}$ be a Hilbert space and $A : \mathcal{H} \rightarrow \mathcal{H}$ a self-adjoint operator with spectral decomposition $A = \int \lambda dE_\lambda$, where $(E_\lambda)_{\lambda \in \mathbb{R}}$ denotes the spectral family of A (that is, E_λ projects onto the spectral subspace corresponding to all spectral values not larger than λ). For any $\psi \in \mathcal{H}$, let $\mu_{A,\psi}$ be defined by

$$\int f d\mu_{A,\psi} = \int f d\langle \psi, E_\lambda \psi \rangle,$$

for all integrable functions $f : \mathbb{R} \rightarrow \mathbb{R}$.

We then define a map on the space of measures that will be a convenient tool for the proof:

Definition 5. (rank one perturbation for general measures) Let ν be a (not necessarily normalized) finite Borel measure on $\mathbb{R}$. Let $L^2(\nu, \mathbb{R})$ denote the Hilbert space of real square integrable functions on $\mathbb{R}$. Define an operator B_ν on $L^2(\nu, \mathbb{R})$ by

$$B_\nu := id + \mathbf{1} \langle \cdot, \cdot \rangle, \quad (31)$$

where id denotes identical map $s \mapsto s$ and $\mathbf{1}$ the constant function attaining always 1. Then we define the rank one perturbation of ν by

$$R(\nu) := \mu_{B_\nu, \mathbf{1}}.$$

The name ‘rank one perturbation’ is justified because R describes how the spectral measure induced by an operator A and a vector ψ changes by replacing A with its rank-one perturbation $A + \psi\psi^T$. To see this, let $\nu = \mu_{A,\psi}$. Assume, without loss of generality, that ψ is a cyclic vector for A , i.e., that the span of $\{A^k \psi\}_{k \in \mathbb{N}}$ is dense in $\mathcal{H}$ (otherwise we restrict A to the completion of this span). By standard spectral theory of operators [17], there is a unitary map $U : \mathcal{H} \rightarrow L^2(\mathbb{R}, \mu_{A,\psi})$ ‘diagonalizing’ A in the sense that $A = U^{-1} id U$ and $U\psi = \mathbf{1}$. Therefore,

$$\mu_{id + \mathbf{1}\mathbf{1}^T, \mathbf{1}} = \mu_{A + \psi\psi^T, \psi}.$$

We do not have a more explicit description of R , but the relation between the Cauchy transforms of ν and $R(\nu)$ is remarkably simple. To describe the relation, we first introduce Cauchy transforms [23]:

Definition 6. (Cauchy transform) Let ν be a not necessarily normalized measure. Then the Cauchy transform of ν is defined⁶ as the complex-valued function from $\mathbb{C}^+$ (that is, the set of complex numbers with positive imaginary part) to $\mathbb{C}$ given by

⁶ Note that some authors define the Cauchy transform as the negative of the below definition.

$$F_V(z) := \int (z - t)^{-1} d\nu(t).$$

Then we find:

Lemma 6. (spectral measure for rank-one perturbation) *Let A be a self-adjoint operator on some Hilbert space $\mathcal{H}$ and let $\mathbf{c} \in \mathcal{H}$ be some vector. Define the rank one perturbation $A_{\mathbf{c}} := A + \mathbf{c}\langle \mathbf{c}, \cdot \rangle$. Set $\nu := \mu_{A_{\mathbf{c}}, \mathbf{c}}$. Then*

$$F_V = \frac{F_{\mu_{A, \mathbf{c}}}}{1 - F_{\mu_{A, \mathbf{c}}}}. \quad (32)$$

Proof. Equation (32) is a special case of the so-called Aronszajn-Krein formula [24–28]. It can be easily seen as follows. Set $A_z := A - z\mathbf{I}$. Moreover, by slightly abusing notation define the linear form $\mathbf{c}^T := \langle \mathbf{c}, \cdot \rangle$. Using the Sherman-Morrison formula [29]

$$(A_z + \mathbf{c}\mathbf{c}^T)^{-1} = A_z^{-1} - \frac{A_z^{-1}\mathbf{c}\mathbf{c}^T A_z^{-1}}{1 + \langle \mathbf{c}, A_z^{-1}\mathbf{c} \rangle},$$

one easily obtains

$$\langle \mathbf{c}, (A_z + \mathbf{c}\mathbf{c}^T)^{-1}\mathbf{c} \rangle = \frac{\langle \mathbf{c}, A_z^{-1}\mathbf{c} \rangle}{1 + \langle \mathbf{c}, A_z^{-1}\mathbf{c} \rangle}.$$

Then the statement follows using

$$\begin{aligned} F_{\mu_{A, \mathbf{c}}}(z) &= -\langle \mathbf{c}, A_z^{-1}\mathbf{c} \rangle \\ F_{A + \mathbf{c}\mathbf{c}^T, \mathbf{c}}(z) &= -\langle \mathbf{c}, (A_z + \mathbf{c}\mathbf{c}^T)^{-1}\mathbf{c} \rangle. \end{aligned}$$

□

By applying Lemma 6 to the operator B_V in eq. (31) we obtain:

Corollary 1. (Cauchy transform of rank one perturbation) *For any finite Borel measure ν on $\mathbb{R}$, the Cauchy transforms of ν and $R(\nu)$ are related by*

$$F_{R(\nu)} = \frac{F_V}{1 - F_V}.$$

Moreover, we will need the following map:

Definition 7. (multiplication map) *If ν denotes a Borel measure on $\mathbb{R}$, we define $M(\nu)$ by*

$$\int f(\lambda) dM(\nu)(\lambda) = \int f(\lambda) \lambda^{-2} d\nu(\lambda),$$

for every measurable function f on $\mathbb{R}$.

The transformation M describes how the spectral measure induced by any matrix A and any vector $\mathbf{c}$ changes when $\mathbf{c}$ is replaced with $A^{-1}\mathbf{c}$, that is

$$M(\mu_{A, \mathbf{c}}) = \mu_{A, A^{-1}\mathbf{c}}.$$

This is also easily verified by diagonalizing A to a multiplication operator on $L^2(\mathbb{R}, \mu_{A, \mathbf{c}})$ as in the remarks after Definition 5.

Using the transformations M and R , we obtain the following concise form for the spectral measure induced by the confounding vector:

$$\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, c\Sigma_{\mathbf{X}\mathbf{X}}^{-1}\mathbf{b}} = c^2 \mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \Sigma_{\mathbf{X}\mathbf{X}}^{-1}\mathbf{b}} = c^2 \mu_{(\Sigma_{\mathbf{E}\mathbf{E}} + \mathbf{b}\mathbf{b}^T), (\Sigma_{\mathbf{E}\mathbf{E}} + \mathbf{b}\mathbf{b}^T)^{-1}\mathbf{b}} = c^2 M[R[\mu_{\Sigma_{\mathbf{E}\mathbf{E}}, \mathbf{b}}]].$$

We will also need the following result:

Lemma 7. (weak continuity of R and M) Let ν_d be a sequence of measures with common support $[l, r]$ with $r > l > 0$. If $\nu_d \rightarrow \nu$ weakly then $R(\nu_d) \rightarrow R(\nu)$ and $M(\nu_d) \rightarrow M(\nu)$ weakly.

Proof. Since ν_d converges weakly to ν , F_{ν_d} converges pointwise to F_ν . Thus, $F_{R(\nu_d)}$ converges pointwise to $F_{R(\nu)}$ for all $z \in \mathbb{C}^+$. Due to Section 3, Theorem 13 in [30], $F_{R(\nu)}$ is the Cauchy transform of the limit of $R(\nu_d)$, see also [31], Section 5.

Weak continuity of M is immediate since M is the multiplication with a function that is bounded by $1/r^2$ and $1/l^2$. □

Since we observe Σ_{XX} and not Σ_{EE} it is important for our purpose that the tracial measure of both matrices asymptotically coincide. The asymptotic version of Lemma 4 reads:

Lemma 8. (tracial measures coincide) If $\mu_{\Sigma_{XX}}^\tau \rightarrow \mu^\infty$ weakly then $\mu_{\Sigma_{EE}}^\tau \rightarrow \mu^\infty$ weakly, too.

Proof. We have for every interval $[r, l]$:

$$\begin{aligned} & \lim_{d \rightarrow \infty} |\mu_{\Sigma_{XX}}^\tau[r, l] - \mu^\infty[r, l]| \leq \\ & \lim_{d \rightarrow \infty} |\mu_{\Sigma_{XX}}^\tau[r, l] - \mu_{\Sigma_{EE}, \tau}^\tau| + \lim_{d \rightarrow \infty} |\mu_{\Sigma_{EE}}^\tau[r, l] - \mu^\infty[r, l]|. \end{aligned}$$

The first term is zero due to Lemma 4 and the second one by assumption. Since the intervals generate the entire Borel sigma algebra the statement follows. □

For each d , the confounding strength β_d depends on the length of $\Sigma_{XX}^{-1}\mathbf{b}$, which, in turn, depends on the specific random choice of $\mathbf{b}_d$. We will see, however, that asymptotically, one always obtains

$$\beta = \frac{c^2 M[R[r_{\mathbf{b}}^2 \mu^\infty]](\mathbb{R})}{r_{\mathbf{a}}^2 + c^2 M[R[r_{\mathbf{b}}^2 \mu^\infty]](\mathbb{R})},$$

since it turns out that the below proof is true for this value of β .

We are now prepared to prove Theorem 2. Due to Theorem 1 we have

$$\begin{aligned} \lim_{d \rightarrow \infty} \mu_{\Sigma_{XX}, \mathbf{a}_d + c(\Sigma_{XX}^d)^{-1}\mathbf{b}_d}^{\Sigma_{XX}^d} &= \lim_{d \rightarrow \infty} \mu_{\Sigma_{XX}, \mathbf{a}_d}^{\Sigma_{XX}^d} + \lim_{d \rightarrow \infty} \mu_{\Sigma_{XX}, c(\Sigma_{XX}^d)^{-1}\mathbf{b}_d}^{\Sigma_{XX}^d} = r_{\mathbf{a}}^2 \mu^\infty + c^2 \mu_{\Sigma_{XX}, (\Sigma_{XX}^d)^{-1}\mathbf{b}_d}^{\Sigma_{XX}^d} \\ &= r_{\mathbf{a}}^2 \mu^\infty + c^2 \lim_{d \rightarrow \infty} M[R[\mu_{\Sigma_{EE}, \mathbf{b}_d}^{\Sigma_{XX}^d}]], \end{aligned}$$

where the last step is due to

$$\mu_{\Sigma_{XX}, (\Sigma_{XX}^d)^{-1}\mathbf{b}_d}^{\Sigma_{XX}^d} = \mu_{\Sigma_{EE}^d + \mathbf{b}_d \mathbf{b}_d^T, (\Sigma_{EE}^d + \mathbf{b}_d \mathbf{b}_d^T)^{-1}\mathbf{b}_d}^{\Sigma_{XX}^d} = M[R[\mu_{\Sigma_{EE}, \mathbf{b}_d}^{\Sigma_{XX}^d}]],$$

by definition of R and M .

Due to Lemma 7 and Theorem 1 we have

$$\lim_{d \rightarrow \infty} M[R[\mu_{\Sigma_{EE}, \mathbf{b}_d}^{\Sigma_{XX}^d}]] = M[R[r_{\mathbf{b}}^2 \mu^\infty]].$$

Hence we obtain

$$\lim_{d \rightarrow \infty} \frac{\mu_{\Sigma_{XX}, \hat{\mathbf{a}}_d}^{\Sigma_{XX}^d}}{\|\hat{\mathbf{a}}_d\|^2} = \frac{r_{\mathbf{a}}^2 \mu^\infty + c^2 M[R[r_{\mathbf{b}}^2 \mu^\infty]]}{\lim_{d \rightarrow \infty} \|\hat{\mathbf{a}}_d\|^2}. \quad (33)$$

To evaluate the denominator on the right hand side, we employ eq. (10) and obtain:

$$\lim_{d \rightarrow \infty} \|\hat{\mathbf{a}}_d\|^2 = \lim_{d \rightarrow \infty} \mu_{\Sigma_{\mathbf{X}\mathbf{X}}^d, \mathbf{a}_d + c(\Sigma_{\mathbf{X}\mathbf{X}}^d)^{-1}\mathbf{b}_d}(\mathbb{R}).$$

Using the asymptotic decomposition of measures in eq. (15) in Theorem 1, we thus have

$$\lim_{d \rightarrow \infty} \|\hat{\mathbf{a}}_d\|^2 = \lim_{d \rightarrow \infty} \mu_{\Sigma_{\mathbf{X}\mathbf{X}}^d, \mathbf{a}_d}(\mathbb{R}) + \lim_{d \rightarrow \infty} \mu_{\Sigma_{\mathbf{X}\mathbf{X}}^d, c(\Sigma_{\mathbf{X}\mathbf{X}}^d)^{-1}\mathbf{b}_d}(\mathbb{R}) \quad (34)$$

$$= r_{\mathbf{a}}^2 \mu^\infty(\mathbb{R}) + \lim_{d \rightarrow \infty} c^2 M[R[\mu_{\Sigma_{\mathbf{E}\mathbf{E}}^d, \mathbf{b}_d}]](\mathbb{R}) = r_{\mathbf{a}}^2 + c^2 M[R[r_{\mathbf{b}}^2 \mu^\infty]](\mathbb{R}), \quad (35)$$

where the second equality is due to eq. (13) and the definition of R and M . The third equality uses the normalization $\mu^\infty(\mathbb{R}) = 1$ (since μ^∞ is a probability measure) and eq. (14). Inserting eq. (35) into eq. (33) yields:

$$\begin{aligned} \lim_{d \rightarrow \infty} \frac{\mu_{\Sigma_{\mathbf{X}\mathbf{X}}^d, \hat{\mathbf{a}}_d}}{\|\hat{\mathbf{a}}_d\|^2} &= \frac{r_{\mathbf{a}}^2}{r_{\mathbf{a}}^2 + c^2 M[R[r_{\mathbf{b}}^2 \mu^\infty]](\mathbb{R})} \mu^\infty + \frac{c^2}{r_{\mathbf{a}}^2 + c^2 M[R[r_{\mathbf{b}}^2 \mu^\infty]](\mathbb{R})} M[R[r_{\mathbf{b}}^2 \mu^\infty]] \\ &= (1 - \beta) \mu^\infty + \beta \frac{M[R[r_{\mathbf{b}}^2 \mu^\infty]]}{M[R[r_{\mathbf{b}}^2 \mu^\infty]](\mathbb{R})} = (1 - \beta) \mu^\infty + \beta \frac{M[R[\eta \mu^\infty]]}{M[R[\eta \mu^\infty]](\mathbb{R})}. \end{aligned} \quad (36)$$

On the other hand, recalling the construction of $v_{\beta, \eta}$ in Section 4, for fixed d (which we drop first) we have

$$\begin{aligned} v_{\beta, \eta} &= (1 - \beta) \mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^\tau + \beta \frac{\mu_{T, T^{-1}(\sqrt{\eta} \mathbf{g})}}{\|T^{-1}(\sqrt{\eta} \mathbf{g})\|^2} = (1 - \beta) \mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^\tau + \beta \frac{M[\mu_{T, \mathbf{g}}]}{M[\mu_{T, \mathbf{g}}]}(\mathbb{R}) \\ &= (1 - \beta) \mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^\tau + \beta \frac{M[R[\mu_{M_X, \sqrt{\eta} \mathbf{g}}]]}{M[R[\mu_{M_X, \sqrt{\eta} \mathbf{g}}]](\mathbb{R})} = (1 - \beta) \mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^\tau + \beta \frac{M[R[\eta \mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^\tau]]}{M[R[\eta \mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^\tau]](\mathbb{R})}, \end{aligned}$$

where we have used $\mu_{M_X, \mathbf{g}} = \mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^\tau$ by construction of $\mathbf{g}$. Hence we obtain:

$$\lim_{d \rightarrow \infty} v_{\beta, \eta}^d = \lim_{d \rightarrow \infty} (1 - \beta) \mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^\tau + \lim_{d \rightarrow \infty} \beta \frac{M[R[\eta \mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^d]]}{M[R[\eta \mu_{\Sigma_{\mathbf{X}\mathbf{X}}}^d]](\mathbb{R})} = (1 - \beta) \mu^\infty + \beta \frac{M[R[\eta \mu^\infty]]}{M[R[\eta \mu^\infty]](\mathbb{R})},$$

which coincides with eq. (36).

6 Experiments with simulated data

6.1 Estimation of strength of confounding

We first ran experiments where the data has been generated according to our model assumptions: First, both the influence of $\mathbf{X}$ on Y and the influence of Z on $\mathbf{X}$ and Y is linear. Second, the vectors $\mathbf{a}$ and $\mathbf{b}$ are randomly drawn from a uniform distribution over the sphere whose radius is randomly chosen. More specifically, the data generating process reads as follows:

- **Generate $\mathbf{E}$:** first generate n samples of a d -dimensional vector valued Gaussian random variable $\tilde{\mathbf{E}}$ with mean zero and covariance matrix $\mathbf{I}$. Then generate a $d \times d$ random matrix G whose entries are independent standard Gaussians and set $\mathbf{E} := G\tilde{\mathbf{E}}$.
- **Generate scalar random variables Z and F** by drawing n samples of each independently from a standard Gaussian distribution.
- **Draw scalar model parameters $c, r_{\mathbf{a}}, r_{\mathbf{b}}$** by independent draws from the uniform distribution on the unit interval.
- **Draw vectors $\mathbf{a}, \mathbf{b}$** independently from a sphere of radius $r_{\mathbf{a}}$ and $r_{\mathbf{b}}$, respectively.
- **Compute $\mathbf{X}$ and Y** via the structural eqs. (1) and (2).

Note that for the above generating process the computation of the true confounding strength β involves only model parameters that are exactly known, even $\Sigma_{\mathbf{X}\mathbf{X}}$ need not be estimated from the data matrix because it is simply given by $\mathbf{G}\mathbf{G}^T + \mathbf{b}\mathbf{b}^T$.

Figure 4 shows the results for dimensions $d = 5; 10; 20$ and sample sizes $n = 100; 1000; 10,000; 100,000$. They indicate that the sample size is even more critical for the performance than the dimension. It seems that the required sample sizes grow so quickly with the dimension that data sets with sample sizes under 1000 should only be considered if the dimension is not larger than about 10.

Figure 4: Simulation results: true value β versus estimated value $\hat{\beta}$ for different dimensions d and sample sizes n . The performance increases with higher dimensions provided that the sample size is large enough.

This observation for moderate dimensions is somehow in contrast to the asymptotic statement that the sample size required for non-regularized estimation of covariance matrices (given some fixed error bound in terms of the operator norm) grows only with $O(d \log d)$ [32, 33] for any distribution with finite second moment, a result that may be relevant in the ‘big data regime’ with huge n and d .

Although the results for dimension 5 look quite bad, it should be noted that β and $\hat{\beta}$ are already significantly correlated, the correlations coefficients varied in the range between 0.65 and 0.8 for the different sample sizes with p -values below 10^{-10} .

We found the true and estimated value of η to be quite uncorrelated, it seems that η is hard to estimate using our method. Since our focus is on the confounding strength, we will not explore this any further.

7 Experiments with real data under controlled conditions

It is hard to find real data where the strength of confounding is known. This is because there are usually unknown confounders in addition to the ones that are obvious for observers with some domain knowledge. For this reason, we have designed an experiment where the variables are observables of technical devices among which the causal structure is known by construction of the experimental setup.

7.1 Setup for a confounded causal influence

To obtain a causal relation where $\mathbf{X}$ influences Y and there is, at the same time, a confounder Z influencing both $\mathbf{X}$ and Y , we have chosen the setup shown in Figure 5.⁷ The cause $\mathbf{X}$ is a 9-dimensional pixel vector generated by extremely reducing the resolution of an image taken by a webcam to 3×3 pixels. The effect Y is the intensity measured at a light sensor in front of a laptop screen that displays the 3×3 image, amplified to a size of about 10×10 centimeter. The sensor is located at a distance of about 10 centimeter from the screen. To confound the causal relation by a common cause Z , we have generated an independent random voltage that controls the brightness of two LEDs: one influencing $\mathbf{X}$ because it is placed in front of the webcam and one that is placed in front of the light sensor. To ensure that $\mathbf{X}$ is not entirely determined by the LED, we have placed the webcam in front of a TV. This way, the image taken by the webcam is influenced by both the LED and the TV signal – the latter plays the role of $\mathbf{E}$ in our structural eq. (1). To avoid that fluctuations of daylight is an additional confounder we have covered the pair sensor and laptop screen by a towel. Since we have measured Z (the random value of the voltage which determines the brightness of the LEDs), we are able to compute the strength β of confounding up to an extent where the estimations of $\Sigma_{\mathbf{X}\mathbf{X}}$, $\mathbf{b}$, $\mathbf{a}$ from empirical data coincide with their true counterparts. We will denote this value by β' to emphasize that it may still deviate from the true value β when the sample size is not sufficient. Figure 6 shows β' and $\hat{\beta}$ for experiments with sample size 1000. For this setup, the algorithm tends to underestimate confounding, but shows qualitatively the right tendency since β' and $\hat{\beta}$ clearly correlate.

After inspecting some spectral measures for the above scenario we believe that the algorithm underestimates confounding for the following reason: The vector $\mathbf{a}$ describing the influence of the images on the sensor is not in generic orientation relative to $\Sigma_{\mathbf{X}\mathbf{X}}$. This is because the pixels are usually *positively* correlated and each pixel has *positive* influence on the total intensity measured by the sensor. This way, $\mathbf{a}$ has stronger weights for high eigenvalues. Since confounding often increases the weights of low eigenvalues (note, however, that this depends also on our second confounding parameter η), the ‘non-genericness’ of the vector $\mathbf{a}$ tends to compensate the effect of confounding on the spectral measure induced by $\hat{\mathbf{a}}$. It is likely that such an underestimation of confounding occurs for many other scenarios as well. This is because it is not uncommon

⁷ The dataset and all others used in the paper can be found at <http://webdav.tuebingen.mpg.de/causality/>.

Figure 5: Setup for the generic causal relation where X influences Y and Z influences both X and Y , see text. Note that we used the symbol for light bulbs to represent the LEDs in order to simplify the drawing.

Figure 6: Confounding strength estimated using observations of Z vs. confounding strength estimated by our algorithm for sample size 1000.

that all variables in a vector are positively correlated⁸ and that they all have a positive influence on some target variable.

⁸ Note also the concept of multivariate total positivity of order two (MTP2) [34], which implies positive correlations between all variables and occurs in many applications [35] such as Markov random fields that consist only of attractive interaction terms.

The above setting contained the purely confounded and the purely causal scenario as limiting cases: the confounded one by putting the LEDs close to the sensor and the webcam, respectively, and setting the voltage to the maximal values, and the purely causal one by setting the voltage to zero.

To get further support for the hypothesis that the algorithms tends to behave qualitatively in the right way even when the estimated strength of confounding deviates from its true value, we also tested modifications of the above scenario that are purely confounded or purely causal *by construction* and not only by variations of parameters. This is described in Sections 7.2 and 7.3.

7.2 Purely causal scenarios

To build a scenario without confounding, $\mathbf{X}$ is the pixel vector of grey values of an image section (consisting of 3×3 pixel) randomly drawn from a fixed image. The image sections are displayed by the screen of a laptop after rescaling them to a size of about 10 cm $\times$ 10 cm. A light sensor, placed in front of the screen with a distance of about 10 cm, measures the light intensity, which is our variable Y . Clearly, $\mathbf{X}$ influences Y in an unconfounded way because the selection of the images is perfectly randomized. Fluctuations of the brightness caused by the environment certainly influence Y , but count as independent noise since they do not influence $\mathbf{X}$.

We tried this experiment with sample size 1000, where the estimated confounding strength was $\hat{\beta} = 0$. We should also mention that we obtained $\beta' = 6.8 \cdot 10^{-7}$ in agreement with our statement that the experimental setup is unconfounded because the image sections are drawn randomly. The extremely low value of β' also shows that β' is indeed very close to the true value β , which justifies to identify them.

7.3 Purely confounded scenario

The setup in Figure 5 can be easily modified to a causal structure where the relation between the pixel vector $\mathbf{X}$ and the light intensity Y is purely confounded by a one-dimensional variable Z : we just need to put the light sensor to a place where it neither sees the TV nor the screen of the laptop. If we, again, ensure that the light sensor is not influenced by the same fluctuations of daylight as the webcam (e.g. by covering the sensor by a towel), the statistical dependence between $\mathbf{X}$ and Y is due to Z , that is, the fluctuations of the random light signal from the LEDs alone.

We have performed this experiment with sample size 1000 and obtained $\hat{\beta} = 0.68$ and $\beta' = 0.998$, which again is consistent with our previous observation that confounding is underestimated.

8 Experiments with real data with partially known causal structure

The experiments in this section refer to real data where the causal structure is not known with certainty. For each data set, however, we will briefly discuss the plausibility of the results in light of our limited domain knowledge. The main purpose of the section is to show that the estimated values of confounding strength indeed spread over the whole interval $[0, 1]$. A priori, we could not be sure whether empirical data follow probability distributions that are so different from our model assumptions that only small or only large values of confounding were estimated.

8.1 Taste of wine

This dataset [36] describes the dependence between Y , the scores on the taste between 0 and 10 (given by human subjects) of red wine, and 11 different ingredients: X_1 : fixed acidity, X_2 : volatile acidity, X_3 : citric acid, X_4 : residual sugar, X_5 : chlorides, X_6 : free sulfur dioxide, X_7 : total sulfur dioxide, X_8 : density, X_9 : pH, X_{10} :

sulphates, X_{11} : alcohol. It turned out that the largest eigenvalue of the covariance matrix is by orders of magnitude larger than the others. We therefore normalized the X_j to unit variance and obtained the covariance matrix

$$\Sigma_{\mathbf{X}\mathbf{X}} = \begin{pmatrix} 1 & -0.26 & 0.67 & 0.11 & 0.09 & -0.15 & -0.11 & 0.67 & -0.68 & 0.18 & -0.06 \\ -0.26 & 1.00 & -0.55 & 0.00 & 0.06 & -0.01 & 0.08 & 0.02 & 0.23 & -0.26 & -0.20 \\ 0.67 & -0.55 & 1 & 0.14 & 0.20 & -0.06 & 0.04 & 0.36 & -0.54 & 0.31 & 0.11 \\ 0.11 & 0.00 & 0.14 & 1 & 0.06 & 0.19 & 0.20 & 0.36 & -0.09 & 0.01 & 0.04 \\ 0.09 & 0.06 & 0.20 & 0.06 & 1 & 0.01 & 0.05 & 0.20 & -0.27 & 0.37 & -0.22 \\ -0.15 & -0.01 & -0.06 & 0.19 & 0.01 & 1 & 0.67 & -0.02 & 0.07 & 0.05 & -0.07 \\ -0.11 & 0.08 & 0.04 & 0.20 & 0.05 & 0.67 & 1 & 0.07 & -0.07 & 0.04 & -0.21 \\ 0.67 & 0.02 & 0.36 & 0.36 & 0.20 & -0.02 & 0.07 & 1 & -0.34 & 0.15 & -0.50 \\ -0.68 & 0.23 & -0.54 & -0.09 & -0.27 & 0.07 & -0.07 & -0.34 & 1 & -0.20 & 0.21 \\ 0.18 & -0.26 & 0.31 & 0.01 & 0.37 & 0.05 & 0.04 & 0.15 & -0.20 & 1 & 0.09 \\ -0.06 & -0.20 & 0.11 & 0.04 & -0.22 & -0.07 & -0.21 & -0.50 & 0.21 & 0.09 & 1 \end{pmatrix}$$

We observe several correlation coefficients around ± 0.5 and ± 0.6 , hence $\Sigma_{\mathbf{X}\mathbf{X}}$ significantly differs from the identity, which is important because $\Sigma_{\mathbf{X}\mathbf{X}} = \mathbf{I}$ would render the method pointless. The vector of regression coefficients reads:

$$\hat{\mathbf{a}} = (0.044, -0.194, -0.036, 0.023, -0.088, 0.046, -0.107, -0.034, -0.064, 0.155, 0.294),$$

showing that alcohol has by far the strongest association with taste. According to common experience, alcohol indeed has a significant influence on the taste. Also the other associations are likely to be mostly causal and not due to a confounder. We estimated confounding for this data set and obtained $\hat{\beta} = 0$.

Since the above experiments suggest that the set of ingredients influence the target variable taste essentially in an unconfounded way, we now explore what happens when we exclude one of the variables $\mathbf{X} := (X_1, \dots, X_{11})$. Since this variable X_j will typically be correlated with the remaining ones and since it, at the same time, influences Y , this will typically confound the causal relation between $\mathbf{X} \setminus X_j$ and Y . For each $j \in \{1, \dots, 11\}$ we have therefore estimated the structural confounding strength and obtained the following results: for all $j < 11$ the algorithm estimated the confounding strength 0.0 (the lowest possible value), while it estimated 0.55 for $j = 11$. Since X_{11} has the strongest influence on Y , this result is remarkable because it is plausible that dropping it corresponds to strong confounding.

Figure 7, left visualizes the weights of the spectral measure for the case where all 11 variables are included and compares it to the one obtained when alcohol is dropped (right). In the latter case, one clearly sees that the weights decrease towards large eigenvalues, which may indicate confounding.

Figure 7: Spectral measure $\mu_{\Sigma_{\mathbf{X}\mathbf{X}}, \hat{\mathbf{a}}}$ for the taste of wine for the case where all ingredients are included (left) versus the case where X_{11} (alcohol) has been dropped (right). In the latter case, the weights decrease for the larger eigenvalues.

8.2 Chicago crime data

This dataset [37] reports⁹ the number of crimes for each of 77 community areas in Chicago, USA, and some potential factors influencing the crime rate [37]. Here, Y denotes the assaults (homicides) and X consists of the following 6 features: X_1 : below poverty level, X_2 : crowded housing, X_3 : dependency, X_4 : no high-school diploma, X_5 : per capita income, X_6 : unemployment. After normalization we obtain the following estimated vector of structure coefficients:

$$\hat{\mathbf{a}} = (3.3, 3.5, 2.8, -7.7, -2.6, 9.7).$$

It seems reasonable that the unemployment rate has the strongest influence. It is, however, surprising that ‘no high-school diploma’ should have a negative influence on the number of crimes. This is probably due to a confounding effect. The estimated confounding strength reads $\hat{\beta} = 0.0$.

8.3 Compressive strength and ingredients of concrete

This experiment considers the data set ‘concrete and compressive strength’ [38, 39] in the machine learning repository Y is the compressive strength in megapascals and X_1 to X_7 are the following components, measured in kg/m^3 . X_1 : cement, X_2 : blast furnace, X_3 : fly ash, X_4 : water, X_5 : superplasticizer, X_6 : coarse aggregate, X_7 : fine aggregate. X_8 is the age in days. After normalization, the estimated vector of structure coefficients reads

$$\hat{\mathbf{a}} = (12.5, 9.0, 5.6, -3.2, 1.7, 1.4, 1.6, 7.2)$$

The amount of superplasticizer seems to have the strongest influence, followed by cement. The estimated confounding strength reads $\hat{\beta} = 0.36$, but it is hard to speculate about possible confounders here.

9 Discussion

We have described a method that estimates the strength of a potential one-dimensional common cause Z that confounds the causal relation between a d -dimensional cause $\mathbf{X}$ (with ‘large’ d) and a one-dimensional effect Y . The presence of Z can, to some extent, be detected from the joint statistics of $\mathbf{X}$ and Y when the vector $\mathbf{a}$ of regression coefficients (after regressing Y on $\mathbf{X}$) is decomposed into the eigenvectors of the covariance matrix $\Sigma_{\mathbf{X}\mathbf{X}}$ of $\mathbf{X}$. This is because generically, without confounding, the weights of this decomposition are roughly uniformly spread over the principal values of $\mathbf{X}$, while the presence of Z will typically modify the weights in a way that is characteristic for the corresponding confounding scenario.

The method is based on the assumption that the vector $\mathbf{a}$ has, in a certain sense, ‘generic orientation’ with respect to $\Sigma_{\mathbf{X}\mathbf{X}}$. The justification of our method relies on a highly idealized model where the vectors of model parameters are randomly generated from a rotation invariant prior. This yields to several concentration of measure phenomena for high dimensions d which the method employs. There is some hope that empirical data show similar concentration of measure phenomena although our model assumptions are probably significantly violated.

Given the difficulty of the enterprise of inferring causal relations from observational data, one should not expect that any method is able to detect the presence of confounders with certainty. Following this modest

⁹ This site provides applications using data that has been modified for use from its original source, www.cityofchicago.org, the official website of the City of Chicago. The City of Chicago makes no claims as to the content, accuracy, timeliness, or completeness of any of the data provided at this site. The data provided at this site is subject to change at any time. It is understood that the data provided at this site is being used at ones own risk.

attitude, the results can be considered encouraging; after all the joint distribution of $\mathbf{X}$ and Y seems to contain some hints on whether or not their causal relation is confounded.

Although the theoretical justification of the method (using asymptotic for dimension to infinity) suggests that the methods should only be applied to large dimension, it should be emphasized that we have so far computed the regression without regularization which quickly requires prohibitively high sample sizes. Future work may apply the method following regularized regression but then one has to make sure that the regularizer does not spoil the method by violating our symmetry assumptions.

To generalize the ideas of this paper to non-linear causal relations one may consider kernel methods [40]. Then the covariance matrix in feature space captures also non-linear relations [41] between the components X_j and the regression vector also non-linear relations between $\mathbf{X}$ and Y . It is unclear, however, how the idea of ‘generic relative orientation’ translates into this setting given that a rotation invariant prior is impossible in an infinite-dimensional space.

Regardless of whether there is a sufficiently number of real-world problems for which the method developed here is applicable, it may help understand three things: first, it shows a possible formalization of the informal idea of ‘independence’ between P_{cause} and $P_{\text{effect}|\text{cause}}$ for the common scenario of multivariate linear models. Second, it shows what kind of sophisticated dependences are generated by confounding in such a scenario, and third, that employing the dependences for confounder detection may be challenging even when the idealized model assumptions hold. This way, the scenario discussed in the present paper may help develop novel practical methods for confounder detection.

Acknowledgment: We would like to thank Uli Wannek for helping us with the implementation of the video experiment. Many thanks also to Roland Speicher and his group in Saarbrücken for helpful discussions about free probability theory and to Steffen Lauritzen for pointing out that real data sometimes show the MTP2 property, which may be an issue here.

Appendix: Relation to free independence

Free probability theory [20, 30] defines a notion of independence that is asymptotically satisfied for independently chosen high-dimensional random matrices. To sketch the idea, we start with a model of generating independent random matrices considered in [42]:

Let $(F_d)_{d \in \mathbb{N}}$ and $(G'_d)_{d \in \mathbb{N}}$ be sequences of $d \times d$ matrices whose tracial spectral measures converge weakly. Let $(U_d)_{d \in \mathbb{N}}$ be random orthogonal matrices drawn from $O(d)$ according to the Haar measure. We start with the simplest case of the independence conditions: F_d and $G_d := U_d G'_d U_d^T$ satisfy asymptotically the Trace Condition [18, 19], i.e.,

$$|\tau(F_d G_d) - \tau(F_d) \tau(G_d)| \rightarrow 0$$

in probability, where τ denotes again the renormalized trace. It is then convenient to introduce limit objects F, G in a C^* -algebra [16] and a functional ϕ , expressing the limit of renormalized traces, for which the Trace Condition then holds exactly:

$$\phi(FG) = \phi(F)\phi(G). \quad (37)$$

However, eq. (37) is only the simplest one of an infinity of independence statements. First, one obtains statements on higher moments like

$$\phi(F^k G^l) = \phi(F^k) \phi(G^l),$$

which is analog to $\mathbf{E}[X^k Y^l] = \mathbf{E}[X^k] \mathbf{E}[Y^l]$ for independent random variables X and Y . But the model above also yields independence statements like $\phi(FGFG) = 0$ whenever $\phi(F) = \phi(G) = 0$, which have no counterpart

with classical random variables. F and G are also considered ‘non-abelian random variables’ and free independence as a stronger version of usual statistical independence which can only hold because the variables do not commute.

The following difficulty arises when we try to apply the above ideas to our generating model used in Theorem 1. Our sequence $(\Sigma_{\mathbf{X}\mathbf{X}}^d)_{d \in \mathbb{N}}$ may take the role of $(F_d)_{d \in \mathbb{N}}$. To draw $\mathbf{a}_d$ uniformly from the unit sphere, we may define an arbitrary sequence $(\mathbf{a}'_d)$ of unit vectors and set $\mathbf{a}_d := U_d \mathbf{a}'_d$ with U_d being a random rotation as above. Then one could naively argue that $\mathbf{a}'_d (\mathbf{a}')_d^T$ takes the role of G'_d and one also expects

$$\tau((\Sigma_{\mathbf{X}\mathbf{X}}^d)^k \mathbf{a}_d \mathbf{a}_d^T) \approx \tau((\Sigma_{\mathbf{X}\mathbf{X}}^d)^k) \tau(\mathbf{a}_d \mathbf{a}_d^T),$$

which is equivalent to

$$\langle \mathbf{a}_d, (\Sigma_{\mathbf{X}\mathbf{X}}^d)^k \mathbf{a}_d \rangle \approx \tau((\Sigma_{\mathbf{X}\mathbf{X}}^d)^k) \|\mathbf{a}_d\|^2.$$

Hence,

$$\int s^k d\mu_{\Sigma_{\mathbf{X}\mathbf{X}}^d, \mathbf{a}_d}(s) \approx \int s^k d\mu_{\Sigma_{\mathbf{X}\mathbf{X}}^d}^\tau(s),$$

for all $k \in \mathbb{N}$ for large d . Thus, all moments of $\mu_{\Sigma_{\mathbf{X}\mathbf{X}}^d, \mathbf{a}_d}$ coincide almost with $\mu_{\Sigma_{\mathbf{X}\mathbf{X}}^d}^\tau \|\mathbf{a}_d\|^2$. This argument, however, blurs the fact that $\tau(\mathbf{a}_d \mathbf{a}_d^T)$ converges to zero while $\text{tr}(\mathbf{a}_d \mathbf{a}_d^T)$ is constant, i.e., we need to consider the asymptotic of $d \cdot \tau(F_d G_d)$ instead of $\tau(F_d G_d)$, which is not covered by free probability theory to the best of our knowledge. Theorem 1 is thus close to the above statements although we do not see any straightforward way to derive it from existing work.

References

- [1] Reichenbach H. The direction of time. Berkeley: University of California Press, 1956.
- [2] Pearl J. Causality: Models, reasoning, and inference. Cambridge University Press, 2000.
- [3] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search (Lecture notes in statistics). New York, NY: Springer-Verlag, 1993.
- [4] Bowden R, Turkington D. Instrumental variables. Cambridge: Cambridge University Press, 1984.
- [5] Hoyer P, Shimizu S, Kerminen A, Palviainen M. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *Int J Approx Reason.* 2008;49:362–378.
- [6] Janzing D, Peters J, Mooij J, Schölkopf B. Identifying latent confounders using additive noise models. In: Ng A, Bilmes J, editor. Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009). Corvallis, OR, USA: AUAI Press, 2009:249–257.
- [7] Janzing D, Sgouritsa E, Stegle O, Peters P, Schölkopf B. Detecting low-complexity unobserved causes. In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011). Available at: <http://uai.sis.pitt.edu/papers/11/p383-janzing.pdf>.
- [8] Janzing D, Balduzzi D, Grosse-Wentrup M, Schölkopf B. Quantifying causal influences. *Ann Stat.* 2013;41:2324–2358.
- [9] Janzing D, Schölkopf B. Causal inference using the algorithmic Markov condition. *IEEE Trans Inf Theo.* 2010;56:5168–5194.
- [10] Lemeire J, Janzing D. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds Mach.* 2012;23:227–249.
- [11] Li M, Vitányi P. An Introduction to Kolmogorov Complexity and its Applications. New York: Springer, 1997 (3rd edition: 2008).
- [12] Janzing D, Steudel B. Justifying additive-noise-based causal discovery via algorithmic information theory. *Open Syst Inf Dynam.* 2010;17:189–212.
- [13] Meek C. Strong completeness and faithfulness in Bayesian networks. In: Proceedings of 11th Uncertainty in Artificial Intelligence (UAI). Montreal, Canada: Morgan Kaufmann, 1995:411–418.
- [14] Uhler C, Raskutti G, Bühlmann P, Yu B. Geometry of the faithfulness assumption in causal inference. *Ann Stat.* 2013;41:436–463.
- [15] Kato T. Perturbation theory for linear operators. Berlin: Springer, 1996.
- [16] Murphy G. C^* -algebras and operator theory. Boston: Academic Press, 1990.
- [17] Reed M, Simon B. Functional Analysis. San Diego, California: Academic Press, 1980.

- [18] Janzing D, Hoyer P, Schölkopf B. Telling cause from effect based on high-dimensional observations. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), Haifa, Israel, 06, 2010:479–486.
- [19] Zscheischler J, Janzing D, Zhang K. Testing whether linear equations are causal: A free probability theory approach. In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011), 2011. Available at: <http://uai.sis.pitt.edu/papers/11/p839-zscheischler.pdf>.
- [20] Voiculescu D, editor. Free probability theory, volume 12 of Fields Institute Communications. American Mathematical Society, 1997.
- [21] Chandrasekaran V, Parrilo P, Willsky A. Latent variable graphical model selection via convex optimization. *Ann Stat.* 2012;40:1935–1967.
- [22] Datta BN. Numerical Linear Algebra and Applications. Philadelphia, USA: Society for Industrial and Applied Mathematics, 2010.
- [23] Cima J, Matheson A, Ross W. The Cauchy Transform. Mathematical Surveys and Monographs 125. American Mathematical Society, 2006.
- [24] Simon B. Spectral analysis of rank one perturbations and applications. Lectur given at the Vancouver Summer School in Mathematical Physics (1993). Available at: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.9138>, 1994.
- [25] Simon B. Trace ideals and their applications. Providence, RI: American Mathematical Society, 2005.
- [26] Kiselev A, Simon B. Rank one perturbations with infinitesimal coupling. *J Funct Anal.* 1995;130:345–356.
- [27] Albeverio S, Konstantinov A, Koshmanenko V. The Aronszajn-Donoghue theory for rank one perturbations of the H_{-2} -class. *Integral Equ Operat Theo.* 2004;50:1–8.
- [28] Albeverio S, Kurasov P. Rank one perturbations, approximations, and selfadjoint extensions. *J Func Anal.* 1997;148:152–169.
- [29] Bartlett MS. An inverse matrix adjustment arising in discriminant analysis. *Ann. Math. Statist.* 1951;22:107–111.
- [30] Mingo J, Speicher R. Free probability and random matrices. New York: Springer, 2017.
- [31] Bercovici H, Voiculescu D. Free convolution of measures with unbounded supports. *Ind Univ Math J.* 1993;42:733–773.
- [32] Rudelson M. Random vectors in the isotropic position. *J Func Anal.* 1999;164:60–72.
- [33] R. Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *J Theo Probab.* 2012;25:655–686.
- [34] Karlin S, Rinott Y. Classes of orderings of measures and related correlation inequalities. I. multivariate totally positive distributions. *J Multiv Anal.* 1980;10:467–498.
- [35] Fallat S, Lauritzen S, Sadeghi K, Uhler C, Wermuth N, Zwiernik P. Total positivity in markov structures. To appear in *Annals of Statistics*, 2016.
- [36] Lichman M. UCI machine learning repository. Available at: <http://archive.ics.uci.edu/ml>, 2013.
- [37] City of Chicago. Data portal: Chicago poverty and crime. Available at: <https://data.cityofchicago.org/Health-Human-Services/Chicago-poverty-and-crime/fwns-pcmk>.
- [38] Yeh C. Concrete compressive strength data set. <https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength>.
- [39] Yeh I-C. Modeling of strength of high performance concrete using artificial neural networks. *Cement Concrete Res.* 1998.
- [40] Schölkopf B, Smola A. Learning with kernels. Cambridge, MA: MIT Press, 2002.
- [41] Gretton A, Herbrich R, Smola A, Bousquet O, Schölkopf B. Kernel methods for measuring independence. *J Mach Learn Res.* 2005;6:2075–2129.
- [42] Speicher R. Free probability theory and non-crossing partitions. *LOTHAR. COMB.* 1997;39.