Detecting non-causal artifacts in multivariate linear regression models

Dominik Janzing [1] **Bernhard Schölkopf** [2]

Abstract

We consider linear models where d potential causes $X_1, \ldots, X_d$ are correlated with one target quantity Y and propose a method to infer whether the association is causal or whether it is an artifact caused by overfitting or hidden common causes. We employ the idea that in the former case the vector of regression coefficients has 'generic' orientation relative to the covariance matrix Σ_{XX} of X. Using an ICA based model for confounding, we show that both confounding and overfitting yield regression vectors that concentrate mainly in the space of low eigenvalues of Σ_{XX}.

1. Introduction

Inferring causal relations from passive observations data has gained increasing interest in machine learning and statistics. Although reliable causal conclusions can only be drawn from interventional data, the idea of postulating assumptions that render causal inference from passive observations feasible becomes more and more accepted. In addition to the more 'traditional' *causal Markov condition* and *causal faithfulness assumption* (Spirtes et al., 1993; Pearl, 2000), researchers have also stated assumptions that admit causal inference when no conditional statistical independences hold, e.g., Kano & Shimizu (2003); Sun et al. (2006); Hoyer et al. (2009); Zhang & Hyvärinen (2009); Bloebaum et al. (2018); Marx & Vreeken (2017). Each of these method relies on idealized assumptions that rarely hold in practice; nevertheless they can be useful if the methods possess a degree of robustness regarding violation of model assumptions (Mooij et al., 2016). In a similar vein, the present work considers a causal inference problem that becomes solvable only under an idealized model assumption that is certainly debatable. However, it illustrates that high-dimensional observations contain a kind of causal information that has not been employed so far.

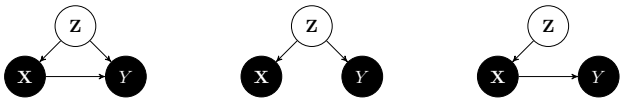

Figure 1. Generic scenario where the statistical relation between $\mathbf{X}$ and Y is due to an unobserved confounder $\mathbf{Z}$ and due to the influence of $\mathbf{X}$ on Y. The purely confounded (middle) and the purely causal (right) scenarios are obtained as limiting cases where one of the arrows is negligible.

We assume that we are given a scalar target variable Y that is potentially influenced by a multi-dimensional predictor variable $\mathbf{X} = (X_1, \ldots, X_d)$. Suppose that i.i.d. samples from $P_{\mathbf{X},Y}$ show that $\mathbf{X}$ and Y are significantly correlated, but it is unclear whether this is mainly due to the influence of $\mathbf{X}$ on Y or due to a common cause of $\mathbf{X}$ and Y (here we assume that prior knowledge excludes the case where Y causally influences $\mathbf{X}$, e.g, due to time order). Y may, for instance, be a quantitiave property of a material (e.g., electrical resistence) and X_j some features describing its chemical and physical structure. In biology, $\mathbf{X}$ and Y could represent information about genotype and phenotype, respectively. Note that conditional independences allow to decide which of the variables X_j influence Y, given that the association between $\mathbf{X}$ and Y is unconfounded. The question of unconfoundedness, which we address here, is therefore prior to the former problem.

Figure 1, left, visualizes the generic scenario that we consider throughout the paper, where the statistical dependences between $\mathbf{X}$ and Y are due to the influence of $\mathbf{X}$ on Y and due to the common cause $\mathbf{Z}$. It contains the purely confounded case (middle) as limiting case where the arrow from $\mathbf{X}$ to Y is arbitrarily weak. Likewise, the purely causal case is obtained when one of the arrows from $\mathbf{Z}$ gets weak (right).

Our confounder detection is based on observing 'non-generic' relations between $P_{\mathbf{X}}$ and $P_{Y|\mathbf{X}}$ (Janzing & Schölkopf, 2017). We thus follow the abstract principle of independent mechanisms (Janzing & Schölkopf, 2010; Lemeire & Janzing, 2012; Peters et al., 2017), stating that for the purely causal relation $X \to Y$ of two arbitrary variables

[1]Amazon Development Center, Tübingen, Germany [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany. This work has been done at the MPI before DJ joined Amazon. Correspondence to: Dominik Janzing <janzind@amazon.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

X, Y, the marginal P_X and the conditional $P_{Y|X}$ do not contain information about each other (where 'information' needs to be further specified). The present paper contains the following novel contributions:

• We allow for multi-dimensional confounders. In contrast, the entire analysis of Janzing & Schölkopf (2017) is restricted to the case of a one-dimensional confounder, and cannot be extended using the methods presented in that work.

• We show that the multivariate setting permits an analysis which is significantly simpler, and also the 'dependences' between $P_{\mathbf{X}}$ and $P_{Y|\mathbf{X}}$ become simple.

• We derive a statistical test for non-confounding based on our model assumptions.

• We show that for our model, overfitting generates the same kind of dependences between $P_{\mathbf{X}}$ and $P_{Y|\mathbf{X}}$ as confounding. This suggests a subtle link between regularization and the correction of confounding. One may conjecture, for instance, that models with 'independent' $P_{\mathbf{X}}$ and $P_{Y|\mathbf{X}}$ have better chances to generalize to future data points as well as to related data sets from other domains (including interventional data), cf. also Schölkopf et al. (2012).

It may sound counter-intuitive that the multivariate case can be simpler than the scalar case, but our derivations are based on a certain notion of genericity of the multivariate confounder which does not necessarily hold for the scalar case, although our experiments will also include data with scalar confounding.

2. Model for confounding with uncorrelated sources

Our model for the influence of the high-dimensional common cause $\mathbf{Z}$ on both $\mathbf{X}$ and Y is inspired by Independent Component Analysis (ICA) (Hyvärinen et al., 2001). Let $\mathbf{Z}$ consist of $\ell \geq d$ independent sources[1] $Z_1, \ldots, Z_\ell$, each having unit variance. They influence $\mathbf{X}$ via a mixing matrix M and Y via a mixing vector $\mathbf{c}$, as shown in Figure 2. Explicitly, the structural equations relating $\mathbf{Z}, \mathbf{X}, Y$ thus read:

$$\mathbf{X} = M\mathbf{Z} \quad (1)$$

$$Y = \mathbf{a}^T\mathbf{X} + \mathbf{c}^T\mathbf{Z}, \quad (2)$$

where M is a $d \times \ell$ matrix and $\mathbf{a}$ are $\mathbf{c}$ are vectors in $\mathbb{R}^d$ and $\mathbb{R}^\ell$, respectively. The model induces the following correlations of the observed variables $\mathbf{X}$ and Y:

$$\Sigma_{\mathbf{XX}} = M\mathbf{I}M^T = MM^T \quad (3)$$

$$\Sigma_{\mathbf{X}Y} = MM^T\mathbf{a} + M\mathbf{c}, \quad (4)$$

[1] In contrast to ICA, however, it is actually enough that the sources are uncorrelated.

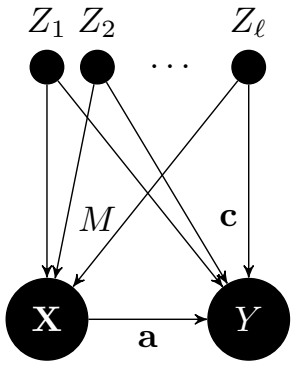

Figure 2. Model of a confounded influence of $\mathbf{X}$ on Y where the hidden common causes are independent sources that influence $\mathbf{X}$ and Y at the same time.

where $\mathbf{I}$ denotes the identity matrix. While $\mathbf{a}$ describes the *causal influence* of $\mathbf{X}$ on Y, formally regressing Y on $\mathbf{X}$ yields

$$\mathbf{a}' := \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}Y} = \mathbf{a} + M^{-T}\mathbf{c}, \quad (5)$$

where M^{-T} denotes the transpose of the pseudoinverse of M. The vector $\mathbf{a}'$ describes how the distribution of Y is shifted when one *observes* that $\mathbf{X}$ has attained a particular d-tuple, while $\mathbf{a}$ describes how it changes when $\mathbf{X}$ is *set* to some d-tuple by an *intervention*. In Pearl's leanguage (Pearl, 2000), $\mathbf{a}'$ vs. $\mathbf{a}$ describe the difference between $p(y|\mathbf{x})$ and $p(y|do(\mathbf{x}))$ for our particular linear model. Janzing & Schölkopf (2017) define the strength of confounding by

$$\beta := \frac{\|\mathbf{a}' - \mathbf{a}\|^2}{\|\mathbf{a}\|^2 + \|\mathbf{a}' - \mathbf{a}\|^2} \in [0, 1], \quad (6)$$

which is 0 for the purely causal case $\mathbf{a}' = \mathbf{a}$ and 1 for the purely confounded case $\mathbf{a} = 0$, which is already a nice property. To further justify this definition, they argue that the vectors $\mathbf{a}$ and $\mathbf{a}' - \mathbf{a} = M^{-T}\mathbf{c}$ are close to orthogonal in high dimensions if $\mathbf{a}$ is drawn independently from $M^{-T}\mathbf{c}$ from a rotation-invariant distribution. Thus, the denominator is close to $\|\mathbf{a}'\|^2$ and β is the fraction of squared length of $\mathbf{a}'$ that can be attributed to the confounder. Following Janzing & Schölkopf (2017) we define the estimation of β from $P_{\mathbf{X},Y}$ as our crucial task, since β quantifies the relative deviation of $\mathbf{a}'$ from $\mathbf{a}$ caused by confounding.

The essential assumption that we add now is that the vectors $\mathbf{c}$ and $\mathbf{a}$ are randomly drawn from a rotation invariant prior (that is, an arbitrary distribution that is invariant w.r.t. orthogonal transformations). One can already guess from (5) what kind of 'non-generic' relation the vector $\mathbf{a}'$ then satisfies together with $\Sigma_{\mathbf{XX}}$: whenever $\mathbf{a}'$ is dominated by the confounding term $M^{-T}\mathbf{c}$ it tends to be mainly located in the eigenspaces of $\Sigma_{\mathbf{XX}}$ corresponding to small eigenvalues. The formal analysis is detailed below, but intuitively speaking, multiplication with M^{-T} amplifies the components

corresponding to small singular values of M and thus to small eigenvalues of $\Sigma_{\mathbf{XX}} = MM^T$.

To formally explore this idea we first introduce the following generating model for $\mathbf{a}$ and $\mathbf{c}$ and hence for $\mathbf{a}'$:

Definition 1 (ICA based confounding model)**.** *First sample each component of* $\mathbf{a}$ *from a Gaussian with zero mean and standard deviation* σ_a*, then sample each component of* $\mathbf{c}$ *from a Gaussian with standard deviation* σ_c*. Compute* $\mathbf{a}'$ *as in* (5)*, where* M *is some given* $d \times \ell$*-matrix.*

Together with M, the parameters σ_a and σ_c determine the expected value of β, but actually only their ratio matters because β depends only on the relative squared lengths of vectors.

3. Estimating the ratio of σ_a and σ_c

We now describe how to infer the ratio of σ_a and σ_c as an intermediate step for inferring β. We could infer both parameters by maximizing the likelihood of $\mathbf{a}'$ given our generating model in Definition 1 if we knew M and ℓ. Unfortunately, we only know $MM^T = \Sigma_{\mathbf{XX}}$ and d. However, we can construct an equivalent generating model for $\mathbf{a}'$ that contains only these observed elements:

Definition 2 (alternative generating model for $\mathbf{a}'$)**.** *Generate* $\mathbf{b} \in \mathbb{R}^d$ *by drawing each component from a standard Gaussian. Set*

$$\mathbf{a}' := \sqrt{\sigma_a^2 \mathbf{I} + \sigma_c^2 \Sigma_{\mathbf{XX}}^{-1}}\mathbf{b}.$$

Theorem 1 (equivalence of models)**.** *The model in Definition 2 generates vectors* $\mathbf{a}'$ *with the same distribution as in Definition 1.*

Proof. First define the $d \times (d + \ell)$-matrix

$$K_{\sigma_a,\sigma_b} := \begin{pmatrix} \sigma_a \mathbf{I} & \sigma_c M^{-T} \end{pmatrix}.$$

We can then rewrite $\mathbf{a}'$ in Definition 1 as

$$\mathbf{a}' = K_{\sigma_a,\sigma_c}\mathbf{b}',$$

with

$$\mathbf{b}' := \begin{pmatrix} \mathbf{a}/\sigma_a \\ \mathbf{c}/\sigma_c \end{pmatrix}.$$

Let

$$K_{\sigma_a,\sigma_c} = \sqrt{K_{\sigma_a,\sigma_c} K_{\sigma_a,\sigma_c}^T} V_{\sigma_a,\sigma_c}$$

be the right polar decomposition of K_{σ_a,σ_c}, where V_{σ_a,σ_c} is a partial isometry from $\mathbb{R}^{d+\ell}$ to $\mathbb{R}^d$. It can be written as

$$V_{\sigma_a,\sigma_c} = W_{\sigma_a,\sigma_c} Q,$$

where W_{σ_a,σ_c} is an orthogonal $d \times d$-matrix and $Q : \mathbb{R}^{d+\ell} \to \mathbb{R}^d$ is the projection that annihilates the last ℓ components of a vector. We then get

$$\mathbf{a}' = \sqrt{K_{\sigma_a,\sigma_c} K_{\sigma_a,\sigma_c}^T} W_{\sigma_a,\sigma_c} Q \mathbf{b}'.$$

Since the $d + \ell$ entries of $\mathbf{b}'$ are drawn from independent standard Gaussians, the d entries of $Q\mathbf{b}'$ are also standard Gaussians. This distribution of entries is invariant under orthogonal maps, hence the entries of

$$\mathbf{b} := W_{\sigma_a,\sigma_c} Q \mathbf{b}'$$

are also independent standard Gaussians. We have

$$\sqrt{K_{\sigma_a,\sigma_c} K_{\sigma_a,\sigma_c}^T} = \sqrt{\sigma_a^2 \mathbf{I} + \sigma_c^2 \Sigma_{\mathbf{XX}}^{-1}},$$

Hence,

$$\mathbf{a}' = \sqrt{\sigma_a^2 \mathbf{I} + \sigma_c^2 \Sigma_{\mathbf{XX}}^{-1}}\mathbf{b}.$$

□

Note that the length of $\mathbf{a}'$ is irrelevant for β. We thus consider $\mathbf{a}'/\|\mathbf{a}'\|$ and infer only the quotient $\theta := \sigma_c^2/\sigma_a^2$. We therefore introduce the matrix

$$R_\theta := \mathbf{I} + \theta \Sigma_{\mathbf{XX}}^{-1}, \tag{7}$$

and conclude that our generating models for $\mathbf{a}'$ induces a distribution for the directions $\mathbf{a}'/\|\mathbf{a}'\|$ that is the image of the uniform distribution on the unit sphere S^{d-1} (i.e. the Haar measure for the orthogonal group) under the map

$$\mathbf{b} \mapsto \frac{\sqrt{R_\theta}\mathbf{b}}{\|\sqrt{R_\theta}\mathbf{b}\|}.$$

To compute this distribution, we use the following result shown in the supplement:

Lemma 1 (distributions of directions induced by a matrix)**.** *Let* A *be an invertible real-valued* $d \times d$*-matrix. Define the map* $\Phi : S^{d-1} \to S^{d-1}$ *by*

$$\Phi(v) := \frac{1}{\|Av\|} Av.$$

Then the image of the uniform distribution on S^{d-1} *under* Φ *has the following density with respect to the uniform distribution:*

$$p(\tilde{v}) = \frac{1}{\det(A)\|A^{-1}\tilde{v}\|^d}. \tag{8}$$

We now apply Lemma 1 to $A := \sqrt{R_\theta}$ as defined by (7) and obtain

$$p_\theta(\tilde{v}) = \frac{1}{|\det \sqrt{R_\theta}| \left\| \sqrt{R_\theta^{-1}} \tilde{v} \right\|^d}. \tag{9}$$

Using

$$|\det \sqrt{R_\theta}| = \sqrt{\det R_\theta}$$

we can rewrite (9) as

$$p_\theta(\tilde{v}) = \frac{1}{\sqrt{\det R_\theta}\|\langle \tilde{v}, (1+\theta\Sigma_{\mathbf{XX}}^{-1})^{-1}\tilde{v}\rangle\|^{d/2}},$$

which proves the following theorem:

Theorem 2 (density of directions)**.** *The generating model in Definition 1 generates vectors* $\mathbf{a}'$ *whose distribution of unit vectors* $\tilde{v} := \mathbf{a}'/\|\mathbf{a}'\|$ *has the following log density with respect to the uniform distribution on the sphere:*

$$\log p_\theta(\tilde{v}) = \frac{1}{2}\left[\log\det R_\theta - d\log\langle \tilde{v}, R_\theta^{-1}\tilde{v}\rangle\right]. \qquad (10)$$

Given sufficiently many samples $\mathbf{a}'$ generated with the same θ, we can certainly infer θ by maximizing (10). Remarkably, we can infer the loglikelihood already from a single instance for large d under appropriate conditions:

Theorem 3 (concentration of measure)**.** *Let* $\tilde{v}$ *be drawn from* $p_{\theta'}$*. Then for sufficiently small* ϵ *we have*

$$\left|\log p_\theta(\tilde{v}) - \frac{1}{2}\left[\log\det R_\theta - \log\frac{\tau(R_{\theta'}R_\theta^{-1})}{\tau(R'_\theta)}\right]\right| \leq \epsilon$$

with probability at least

$$1 - \frac{8}{d\epsilon^2}\left[\tau(R_{\theta'}^2 R_\theta^{-2})\tau(R_{\theta'}R_\theta^{-1})^2 + \tau(R_{\theta'}^2)\tau(R_{\theta'})^2\right].$$

where $\tau() := \frac{1}{d}\mathrm{tr}(.)$ *denotes the renormalized trace.*

The proof can be found in the supplement. Whenever one assumes a limit for $d\to\infty$ in which the expressions with τ converge[2], the error thus tends to zero. Intuitively speaking, the reason is that drawing one vector from $p_{\theta'}$ in dimension d can be reduced to drawing d independent coefficients with respect to an appropriate basis, which finally reduces the above concentration of measure phenomenon problem to the usual law of large numbers.

4. Estimating confounding strength β

To infer β (which we defined as our crucial task) from θ we need some approximations that hold for large d. First we use $\|\mathbf{a}\|^2/d \approx \sigma_a^2$ which is justified by the law of large numbers. Moreover we can estimate the length of $M^{-T}\mathbf{c}$ using the trace of the concentration matrix of $\mathbf{X}$:

$$\begin{aligned}\frac{1}{d}\|\mathbf{a}'-\mathbf{a}\|^2 &= \frac{1}{d}\|M^{-T}\mathbf{c}\|^2 = \frac{1}{d}\langle \mathbf{c}, M^{-1}M^{-T}\mathbf{c}\rangle \\ &\approx \sigma_c\tau(M^{-1}M^{-T}) = \sigma_c\tau(M^{-T}M^{-1}) \\ &= \sigma_c\tau(\Sigma_{\mathbf{XX}}^{-1}),\end{aligned}$$

where the approximation uses also the law of large numbers since we can generate $\mathbf{c}$ by drawing its coefficients with respect to the eigenbasis of $M^{-1}M^{-T}$ from independent Gaussians of standard deviation σ_c. Thus we obtain

$$\beta \approx \frac{\tau(\Sigma_{\mathbf{XX}}^{-1})\sigma_c^2}{\tau(\Sigma_{\mathbf{XX}}^{-1})\sigma_c^2 + \sigma_a^2} = \frac{\tau(\Sigma_{\mathbf{XX}}^{-1})\theta}{\tau(\Sigma_{\mathbf{XX}}^{-1})\theta + 1}. \qquad (11)$$

Putting everything together, we obtain the following procedure for estimating β from $(\mathbf{X}, Y)$ samples:

1. Compute the empirical covariance matrices $\widehat{\Sigma_{\mathbf{XX}}}$ and $\widehat{\Sigma_{\mathbf{X}Y}}$.
2. Estimate $\mathbf{a}'$ via $\widehat{\mathbf{a}'} := \widehat{\Sigma_{\mathbf{XX}}}^{-1}\widehat{\Sigma_{\mathbf{X}Y}}$.
3. Infer θ via maximizing the likelihood $\log p_\theta(\widehat{\mathbf{a}'}/\|\widehat{\mathbf{a}'}\|)$ defined by (10).
4. Compute β from the estimated value of θ via (11).

Here we have neglected finite sample issues completely. We will discuss them in section 6.

5. Test for non-confounding

To test the null hypothesis $\theta = 0$, that is $\mathbf{a}' = \mathbf{a}$, we define the test statistics (applied to a *single* instance $\tilde{v} = \mathbf{a}'/\|\mathbf{a}'\|$)

$$T(\tilde{v}) := \frac{1}{\sqrt{d}}\left\{\langle \tilde{v}, \Sigma_{\mathbf{XX}}^{-1}\tilde{v}\rangle - \tau(\Sigma_{\mathbf{XX}}^{-1})\right\}. \qquad (12)$$

One can easily show that its expectation is zero when $\tilde{v}$ is drawn uniformly at random from the unit sphere, which we assumed for the unconfounded case. Intuitively, the definition of T is motivated by the idea to detect overpopulation of eigenspaces with small eigenvalues, which we expect for confounding. As a further justification, we observed that T coincides, up to an additive constant and a scaling factor, with the score function

$$\frac{\partial \log p_\theta(\tilde{v})}{\partial\theta},$$

at $\theta = 0$. This is a natural candidate for detecting changes of θ because score functions occur in the construction of optimal estimators whenever there exist unbiased estimators attaining the Cramér Rao bound (Cramér, 1946).

[2] This holds, for instance, for any sequence $\Sigma_{\mathbf{XX}}^{(d)}$ for which the eigenvalues are constrained by some interval $[l,u]$ with $0 < l < u < \infty$ and the distribution of eigenvalues converges weakly to some measure μ. Then, $\tau\left(f(\Sigma_{\mathbf{XX}}^{(d)})\right)$ converges to $\int f d\mu$ for any continuous function $f : [l,u] \to \mathbb{R}$.

To derive a simple approximation for the null distribution of T we think of $\tilde{v} = \mathbf{a}'/|\mathbf{a}'|$ as being generated by drawing its coefficients a_j with respect to the eigenbasis of $\Sigma_{\mathbf{XX}}^{-1}$ from $\mathcal{N}(0, 1/\sqrt{d})$ followed by renormalization:

$$\begin{aligned} T(\tilde{v}) &= \frac{1}{\sqrt{d}} \left(\frac{\sum_{j=1}^d a_j^2 s_j}{\sum_{j=1}^d s_j^2} - \tau(\Sigma_{\mathbf{XX}}^{-1}) \right) \\ &\approx \frac{1}{\sqrt{d}} \left(\sum_{j=1}^d a_j^2 s_j - \tau(\Sigma_{\mathbf{XX}}^{-1}) \right), \end{aligned}$$

where s_j denotes the eigenvalues of $\Sigma_{\mathbf{XX}}^{-1}$. Already for moderate size of d, we can thus get a good approximation for the null distribution of T by a weighted sum of squared Gaussian, i.e., it approximately follows a mixed χ^2-distribution.

6. Overfitting

So far we have completely ignored finite sampling issues. High-dimensional regression requires regularization which could spoil our model assumptions, e.g., if they enforce sparsity which is not compatible with our rotation invariant prior on $\mathbf{a}$. Therefore, the method should only be applied if the sample size is sufficiently high for the respective dimension (see section 7) to avoid overfitting. Remarkably, overfitting generates the same kind of 'dependences' between the estimator of $\mathbf{a}'$ and the estimator of $\Sigma_{\mathbf{XX}}$ as confounding generated for the true objects $\mathbf{a}'$ and $\Sigma_{\mathbf{XX}}$ themselves.

To show this, assume that Y is independent of $\mathbf{X}$ and let $(x_1^j, \ldots, x_d^j, y^j)$ for $j = 1, \ldots, n$ be samples independently drawn from $P_{\mathbf{X}} P_Y$, where $P_{\mathbf{X}}$ is arbitrary and P_Y is Gaussian. Define the matrix

$$\mathbf{x} := \left(x_j^i - \bar{x}_j\right)_{i=1,\ldots,n, j=1,\ldots,d},$$

where $\bar{x}_j := \frac{1}{n} \sum_{i=1}^n x_j^i$ denotes the empirical average of the respective component. Likewise, define the vector $y := (y^1, \ldots, y^n)^T - \bar{y}(1, \ldots, 1)^T$. Then, we obtain $\Sigma_{\mathbf{XX}} = \mathbf{x}^T\mathbf{x}$ and $\Sigma_{\mathbf{X}Y} = \mathbf{x}^T y$ (where we have skipped the symbol $\widehat{\ }$ for better readability). Since y is the projection of $(y_1, \ldots, y_n)^T$ onto the orthogonal complement of $\mathbf{1} := (1, \ldots, 1)^T$, its distribution is isotropic in the $n-1$-dimensional subspace defined by the orthogonal complement $\mathbf{1}^\perp$ of $\mathbf{1}$. Let V be an $(n-1) \times n$ matrix that rotates $\mathbf{1}^\perp$ onto $\mathbb{R}^{n-1}$. Then we may write

$$\Sigma_{\mathbf{XX}} = \mathbf{x}^T V^T V \mathbf{x},$$

because the image of $\mathbf{x}$ is contained in the image of the projection $V^T V$. Moreover,

$$\Sigma_{\mathbf{X}Y} = \mathbf{x}^T V^T V y.$$

To show the formal analogy to the mixing scenario above we now set $M := V\mathbf{x}$ and $y' := Vy$. Then we can write $\Sigma_{\mathbf{XX}} = M^T M$ and $\Sigma_{\mathbf{X}Y} = M^T y'$, and thus obtain

$$\hat{\mathbf{a}} = M^{-T} y',$$

where y' is isotropically chosen from R^{n-1}. The generating model for $\hat{\mathbf{a}}$ thus coincides with the model above with $\ell = n - 1$ for the case of pure confounding.

Computing an unregularized regression for $\mathbf{X}$ and Y being independent thus yields a regression vector $\mathbf{a}'$ that is also mainly located in the low eigenvalue eigenspace of $\Sigma_{\mathbf{XX}}$. We expect the same behavior if $\mathbf{X}$ influences Y without confounder when the sample size is so small that the observed correlations are dominated by statistical fluctuations rather than by the true causal influence.

On the one hand one may regret that confounding and overfitting becomes indistinguishable. On the other hand, the method thus provides an unified approach to detect that a regression vector $\widehat{\mathbf{a}'}$ does not show the true causal influence; either because $\widehat{\mathbf{a}'} \neq \mathbf{a}'$ or because $\mathbf{a}' \neq \mathbf{a}$ due to confounding. There is a simple reason why both cases generate similar dependences between $\Sigma_{\mathbf{XX}}$ and $\mathbf{a}'$: Whenever $\Sigma_{\mathbf{X}Y}$ is a vector that has been generated independently of $\Sigma_{\mathbf{XX}}$, the vector $\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}Y}$ tends to live mainly in the small eigenvalue subspace of $\Sigma_{\mathbf{XX}}$. Only if $\Sigma_{\mathbf{X}Y}$ is not drawn independently of $\Sigma_{\mathbf{XX}}$, for instance, because it is generated by $\Sigma_{\mathbf{XX}}\mathbf{a}$ (where $\mathbf{a}$ is drawn independently of $\Sigma_{\mathbf{XX}}$), this overpopulation of small eigenvalues does not happen.

7. Experiments with simulated data

We generated models as follows:

1. We have drawn n samples of each $Z_1, \ldots, Z_\ell$ as independent standard Gaussians
2. We have drawn the entries of M by independent standard Gaussians
3. We have drawn the parameters σ_a, σ_c from the uniform distribution on $[0, 1]$
4. We have drawn each coefficient of $\mathbf{a} \in \mathbb{R}^d$ and $\mathbf{c} \in \mathbb{R}^\ell$ from Gaussians of standard deviation σ_a and σ_c, respectively.
5. We computed samples $(\mathbf{X}, Y)$ via the structural equations $\mathbf{X} = M\mathbf{Z}$ and $Y = \mathbf{a}^T\mathbf{X} + \mathbf{c}^T\mathbf{Z}$.

Knowing the above parameters, we can easily compute the exact confounding strength using

$$\beta = \frac{\|M^{-T}\mathbf{c}\|^2}{\|\mathbf{a}\|^2 + \|M^{-T}\mathbf{c}\|^2}.$$

7.1. Estimating β

We have estimated β as described at the end of section 4 for $d = \ell = 10, 20, 50, 100$ with sample size $10,000$. The scatter plots in Figure 3 show the relation between the true values β and the estimated values $\hat{\beta}$. One can see that β and $\hat{\beta}$ are clearly correlated and that the performance increases (although slowly) for higher dimension. The estimation is reasonably good in the regions where β is close to 0 or 1, which suggests that one should rather trust in the qualitative statement about whether there is confounding or not than in the exact value of $\hat{\beta}$.

Since our theory has shown that ℓ is completely irrelevant in our idealized scenario provided that it is not smaller than d (see the generating model in Definition 2) it would be pointless to explore the case $\ell > d$ here.

7.2. Test for non-confounding

For the simulated data described above we have applied the test for unconfoundedness described in section 5 by drawing 1000 samples from the null distribution of T and comparing them to the observed value $T(\hat{\mathbf{a}}'/\|\hat{\mathbf{a}}'\|)$. Figure 4 visualizes the joint distribution of p-values with β.

One can see that for $\beta > 0.5$ the p-values begin to be mostly close to zero. Figure 5 shows how the fraction of rejections increases when β increases for the two cases where the confidence level α is set to 0.1 (left) or 0.05 (right). Here we have chosen a one-sided test because confounding increases T due to the overpopulation of subspaces with small eigenvalues of $\Sigma_{\mathbf{XX}}$. The results show that for those confidence levels unconfoundedness is mostly rejected for models with $\beta > 0.6$.

7.3. Overfitting

We generated $d+1$-tuples of $\mathbf{X}, Y$ by first drawing $\mathbf{X}$ via a random mixing matrix and then Y by $Y = \mathbf{a}^T\mathbf{X} + E$, where E is $\mathcal{N}(0,1)$ distributed and $\mathbf{a}$ is a random vector whose entries are randomly drawn from $\mathcal{N}(0,1)$.

Figure 6 shows the distribution of p-values of the test for unconfoundedness for different sample sizes n. As one can see, for $n = 20$ one gets mostly small p-values although the model is actually unconfounded (in agreement with our theoretical insights saying that overfitting yields the same type of untypical regression vectors as confounding). For $n = 100$ and $n = 1000$, small p-values are still overrepresented and only for $n = 10,000$ the distribution of p-values is close to uniform. This suggests that dimension 10 already requires sample sizes of the order $10,000$ if one wants to avoid too many false rejections (when focusing on confounding rather than on overfitting).

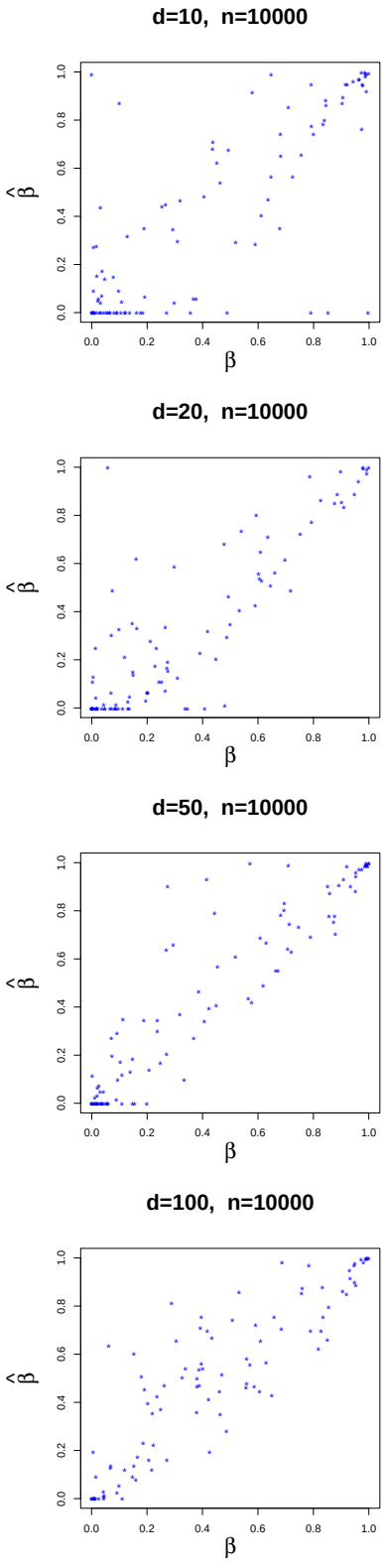

Figure 3. Simulation results: true value β versus estimated value $\hat{\beta}$ for different dimensions d and sample size $n = 10,000$.

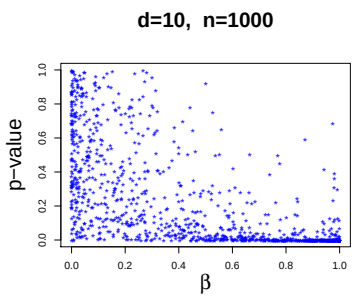

Figure 4. p-values obtained in the test for non-confounding for different values of the confounding parameters β. It can be seen that the p-values get close to zero when β tends to 1.

8. Experiments with real data

Since it is hard to get data where the confounding strength β is known we can mostly only discuss plausibility except for the data set in the following subsection.

8.1. Data from an optical device

Janzing & Schölkopf (2017) describe an optical device where the causal structure and β is known by construction. The variable $\mathbf{X}$ is a low-resolution image (3×3 pixel) shown on the screen of a laptop and Y is the brightness measured by a photodiode at some distance in front of the screen. The image $\mathbf{X}$ is generated by a webcam placed in front of a TV. As confounder Z (which is one-dimensional following the assumptions of Janzing & Schölkopf (2017)), an LED in front of the photodiode and another LED in front of the webcam is controlled by a random noise. Since Z is known, an approximation of β' for β can be directly computed from the observed covariances ($\beta \neq \beta'$ only due to finite sample issues). Among all data sets provided by Janzing & Schölkopf (2017), we first tried those 11 sets [3] that were generated with variable confounding and obtained the results displayed in Figure 7. The results are quite similar to those from Janzing & Schölkopf (2017) although the scenario matches the very specific one-dimensional confounding scenario there while our model is more general. Also here the results are qualitatively right (β' and $\hat{\beta}$ are significantly correlated) but with a clear tendency to underestimate confounding, which has already been discussed by Janzing & Schölkopf (2017).

We also tested the two data sets from Janzing & Schölkopf (2017) where one is purely confounded ($\beta = 1$) and one completely unconfounded ($\beta = 0$) and obtained $\hat{\beta} = 0.768$ and $\hat{\beta} = 0$, respectively.

[3] The data sets and the code are available at `http://webdav.tuebingen.mpg.de/causality/`

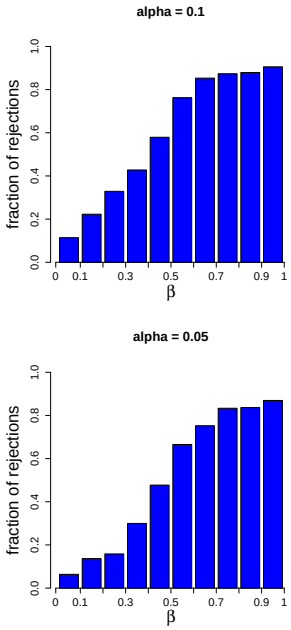

Figure 5. Fraction of rejections when the confidence is chosen to be 0.1 (top) and 0.05 (bottom) for different values of β in 5000 runs.

9. Taste of wine

This dataset (Lichman, 2013) describes the dependence between 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. Taking the taste Y as target variable we obtained $\hat{\beta} = 0$ (after we normalized all X_j to unit variance since their scale where incompatible) which is plausible to some extent given that the crucial ingerndients are considered in the data set.

After dropping alcohol, which one can easily check to have the most dominant influence on taste (given that the relation between the full variable $\mathbf{X}$ and Y has been unconfounded), we obtained $\hat{\beta} = 0.62$, which sounds sensible since the set of predictor variables is no longer sufficient. When we dropped one of then other X_j, we always obtained $\hat{\beta}$ zero or close to zero (in one case). Since the other variables influence the taste much weaker than X_{11}, the algorithm is not able to detect any significant confounding.

10. Data sets with shuffling the target variable

Here we describe a family of experiments where each single one cannot be assessed but one can discuss whether the collection of results seem sensible.

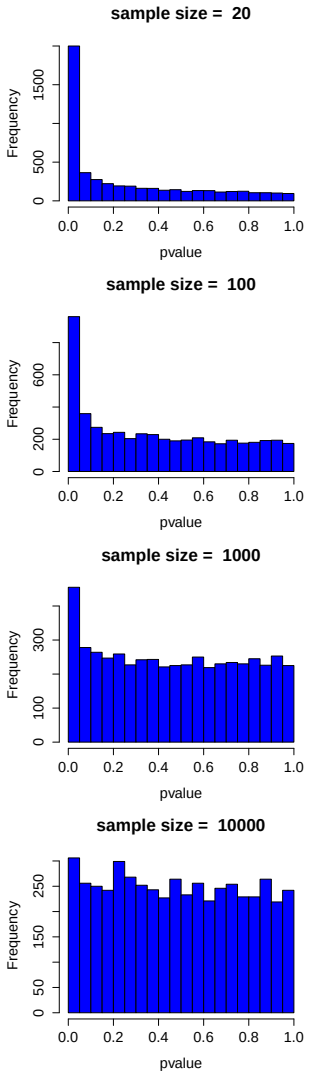

Figure 6. Distribution of p-values in the statistical test for unconfoundedness in a scenario without confounding.

If a data set contains $d+1$ correlated variables $X_1, \ldots, X_{d+1}$ we can take each X_j as hypothetical target variable Y and the remaining variables $\mathbf{X}^{(j)} := (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_{d+1})$ as hypothetical causes. Although we do not know whether some of these $d+1$ choices are purely causal in the sense that $X^{(j)}$ influences $Y^{(j)} := X_j$ without confounder, we know that not all of them are purely causal because not all the variables can be a sink node of the underlying causal DAG.

Since our model uses independent sources as in Independent Component Analysis (ICA) as basis it is natural to apply our method to data sets that have been used in the context of ICA, for instance data from Magnetoencephalographic Recordings (MEG)[4] used by Vigário et al. (1998). The

[4]The data set is available at `http://research.ics.aalto.fi/ica/eegmeg/MEG_data.html`

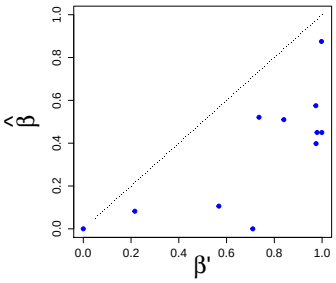

Figure 7. True and estimated confounding strength for the optical device used by Janzing & Schölkopf (2017).

data set contains a data matrix with $17,730$ samples of recordings from 122 channels in a whole-scalp Neuromag-122 neuromagnetometer. We have used the first 10 channels as $X_1, \ldots, X_{d+1}$ and took each of it as potential target and the remaining ones as potential causes. We the obtained for $j = 1, \ldots, 10$ the results $\hat{\beta} = 0.197, 0.257, 0.267, 0.292, 0.135, 0.228, 0.24, 0, 0.311, 0.040$. We do not know the ground truth, but it is reasonable that most of the cases are considered confounded by the algorithm, although we would have expected stronger values of confounding (assuming that all the channels are obtained by mixing the independent sources).

11. Discussion

We have shown that our idealized model assumptions make it possible to infer whether the observed correlations between the multi-dimensional predictor and the target variable are truly causal or an artifact of confounding or overfitting. For our assumptions, both cases of artifacts yield a 'dependence' between the covariance matrix of the potential cause and the regression vector for predicting the effect from the potential cause. Here, 'dependence' has the very simple meaning that principal components corresponding to small eigenvalues being over-represented in the decomposition of the regression vector while the meaning of 'dependence' for the scenario from Janzing & Schölkopf (2017) is more complex.

In our real data experiments, confounding seemed to be often underestimated, which suggests that real data generating process deviate from the model assumptions in a way that the effect of confounding is less visible by our method than the model predicts. Despite these limitations, our findings may inspire further search for hidden causal information in high-dimensional data and provide an intuition about the relevance of concentration of measure effects in causal inference.

References

Bloebaum, P., Janzing, D., Washio, T., Shmimizu, S., and Schölkopf, B. Cause-effect inference by comparing regression errors. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the 21th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, pp. 900–909. PMLR, 2018.

Cramér, H. (ed.). *Mathematical methods of statistics*. Princeton University Press, New Jersey, 1946.

Hoyer, P., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Proceedings of the conference Neural Information Processing Systems (NIPS) 2008*, Vancouver, Canada, 2009. MIT Press.

Hyvärinen, A., Karhunen, J., and Oja, E. (eds.). *Independent Component Analysis*. John Wiley & Sons., 1 edition, 2001.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Janzing, D. and Schölkopf, B. Detecting confounding in multivariate linear models. *Journal of Causal Inference*, 6(1), 2017. doi:10.1515/jci-2017-0013.

Kano, Y. and Shimizu, S. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pp. 261–270, Tokyo, Japan, 2003.

Lemeire, J. and Janzing, D. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, 23(2):227–249, 7 2012.

Lichman, M. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2013.

Marx, A. and Vreeken, J. Telling cause from effect using mdl-based local and global regression. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, pp. 307–316, 2017.

Mooij, J., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Pearl, J. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2000.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, 2017.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In J., Langford and Pineau, J. (eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 1255–1262. ACM, 2012.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search (Lecture notes in statistics)*. Springer-Verlag, New York, NY, 1993.

Sun, X., Janzing, D., and Schölkopf, B. Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pp. 1–11, Fort Lauderdale, FL, 2006.

Vigário, R., Jousmäki, V., Hämäläinen, M., Hari, R., and Oja, E. Independent component analysis for identification of in magnetoencephalographic recordings. In Jordan, M., Kearns, M., and Solla, S. (eds.), *Advances of Neural Information Processing 10, proceedings from the conference, Neural Information Processing Systems 1997*, pp. 229–235. MIT Press, 1998.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pp. 647–655, Arlington, Virginia, United States, 2009. AUAI Press.