Inferring High-Dimensional Causal Relations using Free Probability Theory

Diplomarbeit

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät II
Institut für Mathematik

Eingereicht von Jakob Zscheischler
geboren am 16.08.1985 in Ebersberg
Tübingen, August 2010

Prof. Dr. Markus Reiß

Institut für Mathematik
Humboldt-Universität zu Berlin

PD Dr. Dominik Janzing

Fakultät für Informatik
Universität Karlsruhe (TH)
MPI für biologische Kybernetik Tübingen

Acknowledgement:
I thank Dominik Janzing for the supervising at the MPI, all the fruitful discussions and that he had always time for my questions. I thank Prof. Dr. Markus Reiß for the mentoring in Berlin and Prof. Dr. Bernhard Schölkopf for giving me the chance to write my diploma thesis at the MPI in Tübingen. I thank Jonas Peters for many interesting conversations, his critical view and his friendship. I thank Kun Zhang for inspiring discussions and new ideas and Joris Mooij for sharing his extensive knowledge of computers and their software.
Many thanks to the Evangelisches Studienwerk Villigst for the financial and idealistic support during big parts of my studies. I thank my parents for their support.

List of Publications:

J. M. Maia, A. O. Manzi, F. Rasera, A. Krusche, H. Brandao, S. D. Miller, C. A. Querino, D. K. Adams, and J. Zscheischler (2010). Eddy covariance measurements over Amazonian rivers: the lower Negro river and the middle Solimoes. *Eos Trans. AGU, 91(26), Meet. Am Suppl.*

A. Radebach, J. Runge, J. Zscheischler, J. F. Donges, N. Marwan, and J. Kurths (2010). Evolving complex networks from global climatological fields on geodesic grids. *EGU General Assembly 2010, held 2-7 May, 2010 in Vienna, Austria, p.12723*

P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf (2010). Inferring deterministic causal relations. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI2010).* Best Student Paper Award

When we look about us towards external objects,
and consider the operation of causes, we are never able,
in a single instance, to discover any power or necessary connexion;
any quality, which binds the effect to the cause,
and renders the one an infallible consequence of the other. (David Hume, 1737)

Contents

1 Introduction

1.1 Problem

Imagine the following problem: We are given k pairs $(x_1, y_1), \ldots, (x_k, y_k)$ sampled from the joint distribution $P(X, Y)$, where X and Y are n- and m-dimensional random variables, respectively. We know there is a statistical independence between X and Y. With our given data, how can we infer whether

$$X \to Y \quad , \quad Y \to X$$

or none of these two is right?
Note that it can happen that both $X \to Y$ and $Y \to X$ are right. This problem is of special interest in causality research. Not considering a whole bunch of variables but only two is the most elementary problem in that branch of research. One can imagine many real world problems where this knowledge is of great interest. For example, is there a causal link between human action and global warming? Does a certain medical treatment actually help the patient or is it only because of outer circumstances that he/she got cured? How strong is the influence that our genes have on us?
In this work a method is presented which tries to differentiate between these cases. Our input will be samples of two high-dimensional variables. The main focus will lie on the situation where the number of dimensions exceeds the number of samples.

1.2 Causality

For centuries people had the dream to uncover causal relations. In the time of extensive scientific experimentation knowledge of causal dependencies would help a lot. Since causality is a very vague term with a wide range of definitions and interpretations, we will give a short definition here which should not be taken too close but rather as a working definition.

Definition 1.1 (Causality)
If we do an (hypothetical) intervention on X and observe a variation in the outcome of Y we say: X has a causal influence over Y and we write $X \to Y$.

Causal relationships enable predictions of the consequences of actions (Spirtes et al. [1993]). In most cases controlled randomized experiments constitute the primary tool for identifying causal relationships. However, in a lot of cases such experiments are either unethical, too expensive, or technically impossible. Think for example of medical treatment.

Therefore it is important to develop causal discovery methods to infer causal relationships from uncontrolled data. This already constitutes an important current research topic (here a small selection: Pearl [2000], Spirtes et al. [1993], Geiger and Heckerman [1994], Shimizu et al. [2006], Sun et al. [2008]).
Causal relations are usually described by a directed acyclic graph (DAG) where the observed variables build the nodes and arrows are drawn if there is a causal relationship from one node to another (Pearl [2000], see Figure 1.1).

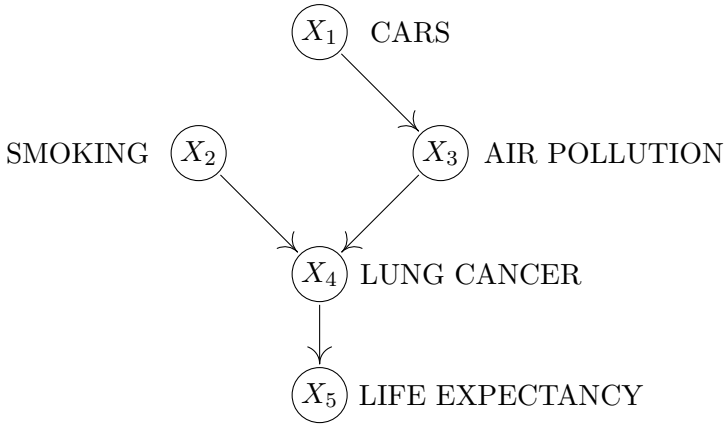
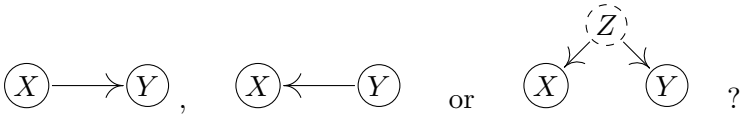


Figure 1.1: Example for a graph representing causal dependencies among five variables.

Inferring causal relations from a set of observed random variables is challenging if no controlled randomized studies can be made. Several approaches are known to solve this task, the most common one perhaps is the independence-based approach (Pearl [2000], Spirtes et al. [1993]) based on the causal Markov condition, that is, that each variable X_i is independent on all its non-descendants, given its parents PA_i (in a DAG G), and an assumption of faithfulness: We accept only those causal DAGs that explain all of the observed dependencies in the data. Furthermore, the DAG should not contain more, i.e. all inferred marginal and conditional *in*dependencies in the data should also derive from the structure of the DAG. Methods based on this assumption both have its advantages and limitations. One main problem is that approaches solely based on conditional independencies cannot distinguish between causally distinct models that impose the same set of independencies (so-called Markov-equivalent graphs). In particular, they cannot infer whether X causes Y or Y causes X for just two observed variables X and Y (Mooij and Janzing [2010]).
In the last few years several authors have proposed new ideas to do causal discovery, based on independent component analysis (ICA) or (more generally) additive-noise models. These methods assume that the effect is given by some (possibly nonlinear) function of the cause up to an additive noise that is statistically independent of the cause (Kano and Shimizu [2003], Shimizu et al. [2006], Hoyer et al. [2009], Peters et al. [2010]). A recent proposal generalizes this model class by further allowing nonlinear transformations of the effect (Zhang and Hyvärinen [2009]).

However, methods based on additive noise models fail for linear relationships with Gaussian noise. An unsolved problem for a long time has been how to deal with deterministic relationships between the observed variables. A first approach into this direction was done by Daniusis et al. [2010]. It is based on the same postulate we use in this work (Postulate 2.1).

Another difficulty, particularly for the case of two variables, is the so-called confounding problem. That is, assume we know that X and Y are correlated, we still do not know if this correlation is due to a common cause Z. For this reason it is essential to develop a method that can distinguish between



To solve this problem we need other prior assumptions than the causal Markov condition. Suppose we have the cause X and the effect Y. Our additional assumption will be that the cause X and the mechanism mapping X to Y are chosen independently from each other since both correspond to independent mechanisms of nature.

This is a rather vague idea. Therefore a more formal mathematical examination will be given below. At this point we will only give an intuition of the idea.

Imagine S to be a source creating x-values according to some distribution $P(X)$. M is a machine that gets these x as input and outputs y. Thus it represents the causal mechanism which reflects the mapping from x to y. It can be characterized by the conditional distribution of Y given X, i.e., $P(Y|X)$. Finally, the output y is characterized by the distribution $P(Y)$. We now claim that the machine M is "independent" from the input distribution $P(X)$ since these two entities represent independent mechanisms in nature. We will develop this idea a bit further at the beginning of Chapter 2.

In major parts this work will follow the line of Janzing et al. [2010], apart from the fact that it will focus on the *small sample case* with a deterministic relationship. That means, the model investigated will be

$$Y = AX$$

with high-dimensional Y and X and A being a matrix. The sample size will be smaller than the dimensionality of both X and Y. This case requires significant modifications of the approach of Janzing et al. [2010], since one has to find new ways to estimate A. New elements are the application of free probability theory, which is introduced in Section 2.2, a slightly different generation of the simulated data and new applications of the method to real world data.

1.3 Outline

We start with an introduction into free probability in Chapter 2. Free probability deals with noncommutative random variables and has strong connections to random matrices. At the end of this chapter we introduce two important lemmas, the Schur Lemma and Lévy's Lemma.
In Chapter 3 we establish identifiability results. That means, we prove how one can distinguish between cause and effect under the given assumptions. We need identifiability to derive an algorithm that gives us the right causal relations. Here, we constitute an asymmetry between the forward and the backward direction.
Chapter 4 proposes an inference method which results in a simple algorithm and uses the identifiability results from Chapter 3. We also present experiments there, both on simulated and real world data.
Chapter 5 gives a short extension to the nonlinear case.
Finally, Chapter 6 summarizes the work, gives an outlook and makes suggestions for future work.
Chapters 3-5 describe my own work, whereas in Chapter 1 and 2, I present the preliminaries.

2 Mathematical Tools

In this chapter we will first give some ideas concerning the basic assumption of this work. We will call this "philosophy". Afterwards we give an introduction into free probability theory. At its end we will present two important lemmas and provide a motivating example that gives an intuition about what kind of asymmetries between cause and effect can be expected in the next chapter.

2.1 Philosophy

One can argue that the most elementary problem in causal inference is to decide whether statistical dependencies between two random variables X and Y are due to a causal influence from X to Y ($X \to Y$), an influence from Y to X ($Y \to X$), or a possibly unobserved common cause Z influencing X and Y ($X \leftarrow Z \to X$).
Recent work (Sun et al. [2006], Sun et al. [2008], Janzing and Schölkopf [2010]) suggests that the shape of the joint distribution shows asymmetries between cause and effect, which often indicates the causal direction with some reliability. It is our aim in this work to establish such a kind of asymmetry between X and Y for the case that one is a linear transformation of the other.
One idea to do this is based on a postulate which was made by Lemeire and Dirkx [2006] and Janzing and Schölkopf [2010]. They postulate that if the causal model is $X \to Y$, the marginal and the conditional distribution, $P(X)$ and $P(Y|X)$, are *algorithmically* independent in that sense that the shortest "description" of $P(X, Y)$ is given by separate "descriptions" of the input distribution $P(X)$ and the conditional distribution $P(Y|X)$. This expresses the fact that both represent independent mechanisms of nature. Descriptions are here characterized through Kolmogorov complexity. Janzing and Schölkopf [2010] show some toy examples where such an independent choice of $P(X)$ and $P(Y|X)$ often leads to joint distributions, where $P(Y)$ and $P(X|Y)$ satisfy some suspicious relations indicating that $Y \to X$ is wrong. Since Kolmogorov complexity is known to be uncomputable we need to find other methods to exploit this independence.

Daniusis et al. [2010] formulated this assumption in a postulate for the deterministic relationship

$$Y = f(X)$$

which we want to introduce here:

Postulate 2.1
If $X \to Y$, the distribution of X and the function f mapping X to Y are independent since they correspond to independent mechanisms in nature.

Independence should not be understood as statistical independence here. We will define later how this is meant in our particular case. Janzing et al. [2010] have developed the same idea for multi-dimensional variables X and Y with a linear causal relation and noise, i.e.,

$$Y = AX + E$$

with multivariate X and Y, a matrix A and a noise term E, independent from X. For this specific case we can restate the postulate as

Postulate 2.2
If $X \to Y$, C_{XX} and the matrix $A^T A$ are independent since they correspond to independent mechanisms in nature.

Here C_{XX} denotes the covariance matrix of X. Our model assumption will be that C_{XX} is randomly drawn from a distribution that is invariant under transformations

$$C_{XX} \mapsto U C_{XX} U^T$$

with U randomly chosen according to the Haar measure out of the orthogonal group. This will be indicated with the term *independent*.

The eigenspaces of C_{XX} and $A^T A$ can lie distorted to each other, which we would expect in the generic case, or it can happen that eigenspaces of C_{XX} with big eigenvalues meet eigenspaces of $A^T A$ with big eigenvalues. In that case we would say that there are dependencies between these two matrices. To be able to investigate these relationships further we will introduce free probability theory in the next section which has a strong connection to random matrix theory.

2.2 Introduction into Free Probability Theory

Free probability theory is a theory dealing with noncommutative random variables. In contrast to classical probability theory, tensor products are replaced by free products, and independent random variables are replaced by free random variables. The theory arose from attempts to solve some longstanding problems about von Neumann algebras of free groups. Since its creation free probability formed connections to several other parts of mathematics: classical probability, operator algebras and the theory of random matrices. It also has connections with some mathematical models in theoretical physics (see Voiculescu [1997]).
We will first give two definitions, the definition of a unital algebra and the definition of a noncommutative probability space.

Definition 2.3 (Unital algebra, see f.e. Murphy [1990])
An *algebra* is a vector space $\mathscr{A}$ together with a bilinear map

$$\begin{array}{ccc} \mathscr{A}^2 & \to & \mathscr{A} \\ (a,b) & \mapsto & ab \end{array}$$

such that

$$a(bc) = (ab)c \quad (a, b, c \in \mathscr{A}) .$$

If $\mathscr{A}$ admits a unit 1 ($a1 = 1a = a$, for all $a \in \mathscr{A}$) and $\|1\| = 1$, where $\|\cdot\|$ is a norm, we say that $\mathscr{A}$ is a *unital algebra*.

Definition 2.4 (Noncommutative probability space, Voiculescu et al. [1992])
A *noncommutative probability space* is a unital algebra, $\mathscr{A}$ over $\mathbb{C}$ together with a linear functional, $\phi : \mathscr{A} \to \mathbb{C}$, such that $\phi(1) = 1$.

We now follow Speicher [2001] and start with a few necessary definitions regarding random matrices. Later we will present some results concerning free probability in relation with random matrices.
Let τ_n be the normalized trace on $n \times n$ matrices, i.e., for a $n \times n$ matrix A

$$\tau_n(A) := \frac{1}{n}\mathtt{tr}(A).$$

In the same way, we get the averaged trace $\tau_n \otimes \mathbb{E}$ for $n \times n$ random matrices. Therefore consider a sequence $(A_n)_{n\in\mathbb{N}}$ of $n \times n$ matrices, where the entries a_{ij} are random variables on some probability space Ω equipped with a probability measure P. We have

$$\tau_n \otimes \mathbb{E}(A_n) := \frac{1}{n}\sum_{i=1}^{n}\int_\Omega a_{ii}(\omega)dP(\omega).$$

Given these "states" $\tau_n \otimes \mathbb{E}$, we can talk about the s-th moment $\tau_n \otimes \mathbb{E}(A_n^s)$ of the random matrix A_n, and it is known that for "sufficiently nice" matrix ensembles these moments converge for $n \to \infty$ (Voiculescu [1991], for specific examples see Edelman and Rao [2005]). We give one example:

Example (The semicircle distribution) Wigner's law states that the spectral measure of a random symmetric $n \times n$-matrix with Gaussian-distributed elements with variance $1/\sqrt{n}$ tends to the (Wigner) semicircle distribution as n tends to infinity (Wigner [1955, 1958]). The semicircle distribution is given by

$$f(x) = \frac{2}{\pi R^2}\sqrt{R^2 - x^2}$$

for $-R < x < R$, and $f(x) = 0$ if $x > R$ or $x < -R$. A plot of the probability density function of this distribution for different values of R is shown in Figure 2.1.

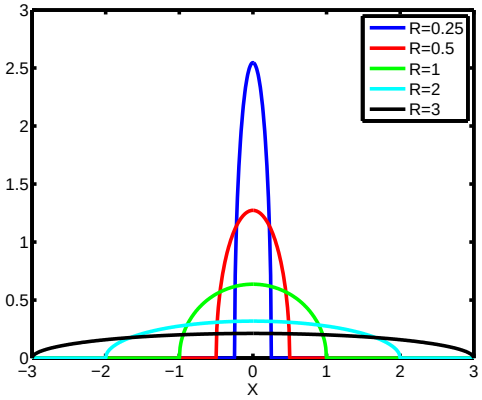


Figure 2.1: Probability density function of the Wigner semicircle distribution with different parameters R.

Remark 2.5

Above, we used the term *state* although it is reserved for positive linear functionals on C^*-algebras of norm one. We will see that the limit elements of random matrices are elements of some algebra. In free probability theory these algebras are often C^*-algebras, although in our case it is enough to consider ordinary algebras.

Assuming now that the limit of the s-th moment exist, let us denote it by α_s, i.e.,

$$\lim_{n\to\infty} \tau_n \otimes \mathbb{E}(A_n^s) =: \alpha_s. \tag{2.1}$$

Thus we can say that the limit $n = \infty$ consists exactly of the collection of all these moments α_k. We can identify these numbers as moments of some variable A. Going a bit further, we can view A as an element of some abstract algebra $\mathscr{A}$ which was generated by A and define a state ϕ on $\mathscr{A}$ through

$$\phi(A^s) := \alpha_s\,. \tag{2.2}$$

Now we can say that our random matrices A_n converge to the variable A in distribution which is defined by the next definition. We will denote this by $A_n \to A$.

Definition 2.6 (Convergence in distribution, Speicher [2001])

Consider $n \times n$-random matrices $A_n^{(1)}, \ldots, A_n^{(m)}$ and variables $A_1, \ldots, A_m \in \mathscr{A}$, where $\mathscr{A}$ is an algebra over $\mathbb{C}$ with a linear functional $\phi : \mathscr{A} \to \mathbb{C}$, such that $\phi(1) = 1$. We say that

$$(A_n^{(1)}, \ldots, A_n^{(m)}) \to (A_1, \ldots, A_m) \quad \text{in distribution,}$$

if

$$\lim_{n\to\infty} \tau_n \otimes \mathbb{E}(A_n^{(i_1)} \cdots A_n^{(i_k)}) = \phi(A_{i_1} \cdots A_{i_k})$$

for all choices of k with $i_1, \ldots, i_k \in \{1, \ldots, m\}$.

The limit elements $A_1, \ldots, A_m$ are now elements of a noncommutative algebra according to Definition 2.4. At this point we note that for a self-adjoint operator $A = A^*$ the collection of moments corresponds also to a probability measure μ_A on the real line, determined by

$$\phi(A^k) = \int_{\mathbb{R}} t^k d\mu_A(t)\,.$$

In particular, for a real-valued symmetric $n \times n$-matrix $A = A^T$ this measure is given by the eigenvalue distribution of A, i.e., it puts mass $1/n$ on each of the eigenvalues of A (counted with multiplicity):

$$\mu_A = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i} \tag{2.3}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. We will use this measure in Section 3.2.

Let A_n and B_n be two independent random matrices (all entries of A_n are independent of all entries of B_n) where all joint moments converge. Then we can say $(A_n, B_n) \to (A, B)$ What is the relation between A and B? Is there still something like independence? It turns out that the corresponding term here is *freeness* which is defined in the following sense:

Definition 2.7 (Freeness; Voiculescu [1985], Speicher [1997])
Let $\mathscr{A}$ be a unital algebra and $\phi : \mathscr{A} \to \mathbb{C}$ a linear functional on $\mathscr{A}$, which is unital, i.e., $\phi(1) = 1$. Then $a_1, \ldots, a_m \in \mathscr{A}$ are called *free* (with respect to ϕ) if

$$\phi[p_1(a_{i(1)}) \cdots p_k(a_{i(k)})] = 0$$

whenever

- $p_1, \ldots, p_k$ are polynomials in one variable
- $i(1) \neq i(2) \neq i(3) \neq \cdots \neq i(k)$ (only neighbouring elements are required to be distinct)
- $\phi[p_j(a_{i(j)})] = 0$ for all $j = 1, \ldots, k$.

Remark 2.8
For calculations with higher order terms there is a more convenient definition of freeness in terms of free cumulants. Free cumulants involve non-crossing partitions which will be introduced in the next section. There is an analogy to the independence in classical probability spaces: Any mixed cumulant involving independent variables is zero in classical probability theory. Freeness now is equivalent to the vanishing of mixed (free) cumulants in free probability theory. This is often easier to handle.

With this definition of freeness we can try to calculate mixed moments in terms of moments of the singular variables. Especially, if a and b are free, then the definition of freeness requires that

$$\phi[(a - \phi(a) \cdot 1)(b - \phi(b) \cdot 1)] = 0$$

which implies that

$$\phi(ab) = \phi(a) \cdot \phi(b). \tag{2.4}$$

Up to now, there is no difference to the results for classical independent random variables. But consider next for a and b free,

$$\phi[(a - \phi(a) \cdot 1)(b - \phi(b) \cdot 1)(a - \phi(a) \cdot 1)(b - \phi(b) \cdot 1)] = 0\,.$$

From that we can derive

$$\phi(abab) = \phi(aa) \cdot \phi(b) \cdot \phi(b) + \phi(a) \cdot \phi(a) \cdot \phi(bb) - \phi(a) \cdot \phi(b) \cdot \phi(a) \cdot \phi(b). \tag{2.5}$$

This shows that freeness is something different from classical independence; indeed it seems to be more complicated. Also, ϕ seems to play a similar role here as $\mathbb{E}$ does in classical probability theory. It is possible (at least in principle) to calculate all mixed

moments by reducing them to alternating products of centered variables as in the definition of freeness. Remembering that ϕ is the limit of the normalized trace of a sequence of random matrices, this fact holds approximately for independent matrices and sufficiently high dimension n. Later on (in Section 3.2) we want to calculate mixed moments of a certain structure, namely $\phi[(bp)^s]$ where p is a projection matrix. To do this we need the concept of non-crossing partitions which we will introduce now.

2.2.1 Non-Crossing Partitions

Freeness is defined in terms of mixed moments, but this definition is not easy to handle if we want to make concrete calculations. It is possible to describe freeness by another, more combinatorial approach which puts the main emphasis on so called "free cumulants". The nomenclature comes from classical probability theory, where corresponding elements are also called cumulants. There exists a combinatorial description of these classical cumulants which depends on partitions of sets. Now, free cumulants can be described combinatorially in a similar way, we only have to replace all partitions by so called "non-crossing partitions". Our presentation will follow Speicher [1997].

Definition 2.9

A *partition* of the set $S := \{1, \ldots, n\}$ is a decomposition

$$\pi = \{V_1, \ldots, V_r\}$$

of S into disjoint and non-empty sets V_i, i.e.

$$V_i \cap V_j = \emptyset \quad (i, j = 1, \ldots, r; i \neq j) \quad \text{and} \quad S = \bigcup_{i=1}^{r} V_i \,.$$

We call the V_i the *blocks* of π.
For $1 \leq p, q \leq n$ we write

$$p \sim_\pi q \quad \text{if } p \text{ and } q \text{ belong to the same block of } \pi \,.$$

A partition π is called *non-crossing* if the following does not occur: There exist $1 \leq p_1 < q_1 < p_2 < q_2 \leq n$ with

$$p_1 \sim_\pi p_2 \not\sim_\pi q_1 \sim_\pi q_2 \,.$$

The set of all non-crossing partitions of $1, \ldots, n$ is denoted by $NC(n)$.

Non-crossing partitions were introduced by Kreweras [1972] in a purely combinatorial context. An equivalent definition of non-crossing sets is the following: If we label the vertices of a regular n-gon with the numbers 1 through n, the convex hulls of the partition's different blocks are disjoint from each other, i.e., they also do not "cross" each other. A visualization of a non-crossing partition of a set with ten points is given in Figure 2.2.

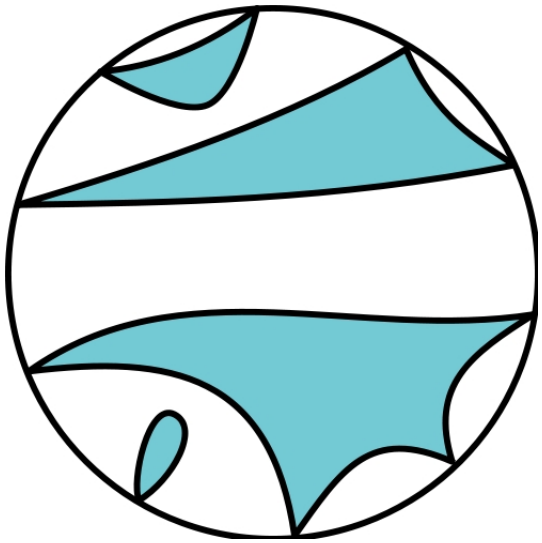


Figure 2.2: A non-crossing partition of ten points[1]

The number of non-crossing partitions of a set of size n is given by the so-called Catalan numbers (Speicher [1994], Corollary 2) which are given by

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

Definition 2.10

Let $\mathscr{A}$ be a unital C^*-algebra and $\phi : \mathscr{A} \to \mathbb{R}$ a linear functional. We define the *(free or non-crossing) cumulants*

$$k_n : \mathscr{A}^n \to \mathbb{R} \quad (n \in \mathbb{N})$$

(indirectly) by the following system of equations:

$$\phi(a_1 \cdots a_n) = \sum_{\pi \in NC(n)} k_\pi[a_1, \ldots, a_n] \quad (a_1, \ldots, a_n \in \mathscr{A}),$$

where k_π denotes a product of cumulants according to the block structure of π:

$$k_\pi[a_1, \ldots, a_n] := k_{V_1}[a_1, \ldots, a_n] \cdots k_{V_r}[a_1, \ldots, a_n] \quad \text{for} \quad \pi = \{V_1, \ldots, V_r\} \in NC(n)$$

and

$$k_V[a_1, \ldots, a_n] := k_{\#V}(a_{v_1}, \ldots, a_{v_l}) \quad \text{for} \quad V = (v_1, \cdots v_l)$$

Thus, one can calculate $k_{\#V}(a_{v_1}, \ldots, a_{v_l})$ and therefore $k_\pi[a_1, \ldots, a_n]$ with the help of $\phi(a_1 \cdots a_n)$. This definition will become clearer with the following examples.

[1] Figure taken from http://en.wikipedia.org/wiki/File:Noncrossing-partition.svg.

Remarks and Examples
1) The above equations have the form

$$\phi(a_1 \cdots a_n) = k_n(a_1, \ldots, a_n) + \text{smaller order terms}$$

and thus they can be resolved for the $k_n(a_1, \ldots, a_n)$ in a unique way.
2) Examples:

- $n = 1$

$$\phi(a_1) = k_1(a_1)\,.$$

- $n = 2$

$$\phi(a_1 a_2) = k_2(a_1, a_2) + k_1(a_1)k_1(a_2)\,,$$

thus

$$k_2(a_1, a_2) = \phi(a_1 a_2) - \phi(a_1)\phi(a_2)\,.$$

- $n = 3$

$$\begin{aligned}\phi(a_1 a_2 a_3) =& k_3(a_1, a_2 a_3) + k_1(a_1)k_2(a_2, a_3) + k_2(a_1, a_2)k_1(a_3) \\ &+ k_2(a_1, a_3)k_1(a_2) + k_1(a_1)k_1(a_2)k_1(a_3)\,,\end{aligned}$$

and thus

$$\begin{aligned}k_3(a_1, a_2, a_3) =& \phi(a_1 a_2 a_3) - \phi(a_1)\phi(a_2 a_3) - \phi(a_1 a_3)\phi(a_2) \\ &- \phi(a_1 a_2)\phi(a_3) + 2\phi(a_1)\phi(a_2)\phi(a_3)\,.\end{aligned}$$

3) The k_n are multi-linear in their n arguments.

For a random variable $a \in \mathscr{A}$ we put

$$k_n^a := k_n(a, \ldots, a)$$

and call $(k_n^a)_{n \geq 1}$ the *(free) cumulants of a*. We can now define freeness with terms of free cumulants:

Theorem 2.11 *(Speicher [1994], cf. Nica [1996])*
The following two statements are equivalent:

1. $a_1, \ldots, a_l$ *are free.*
2. $k_n(a_{i(1)}, \ldots, a_{i(n)}) = 0\,(n \in \mathbb{N})$ *whenever there are* $1 \leq p, q \leq n$ *with* $i(p) \neq i(q)$.

We want to apply this machinery onto the multiplication of random variables in $\mathscr{A}$. First we provide an important result from Voiculescu concerning the relation between random matrices and free probability theory.

Theorem 2.12 *(Asymptotic freeness, Speicher [1997], Section 9.2; cf. Voiculescu et al. [1992])*

1. *Let*
$$A^{(n)} = \left(a_{ij}^{(n)}\right)_{i,j=1}^{n} \quad \text{and} \quad B^{(n)} = \left(b_{ij}^{(n)}\right)_{i,j=1}^{n}$$
be symmetric $n \times n$-random matrices with
   a) *$a_{ij}^{(n)}$ $(1 \leq i \leq j \leq n)$ are independent and normally distributed (mean zero, variance $1/n$)*
   b) *$b_{ij}^{(n)}$ $(1 \leq i \leq j \leq n)$ are independent and normally distributed (mean zero, variance $1/n$)*
   c) *all $a_{ij}^{(n)}$ are independent from all $b_{kl}^{(n)}$.*

   Then $A^{(n)}$ and $B^{(n)}$ become free in the limit $n \to \infty$ with respect to ϕ (defined through (2.1) and (2.2)).
2. *Let A_n and C_n be real-valued symmetric deterministic $n \times n$-matrices whose eigenvalue distribution tend to some fixed probability measure μ and ν, respectively, in the limit $n \to \infty$. Consider now*
$$B_n = U_n C_n U_n^T,$$
where U_n is a random $n \times n$ orthogonal matrix from the ensemble $O(n)$ equipped with the Haar measure. Then A_n and B_n become free in the limit $n \to \infty$ with respect to ϕ.

This relates the independence of the matrix entries with freeness of the limit elements of these matrices. Note that part 2 of this theorem is much more general than part 1. In the first part, $A^{(n)}$ and $B^{(n)}$ are Gaussian random matrices and thus, by a result of Wigner [1955], their eigenvalue distributions tend towards the so-called semi-circle distribution for $n \to \infty$.
In part 2 of the theorem, however, we are not restricted to semi-circular distributions, but we can prescribe in the limit any distribution we want. Our focus in this work will lie on the second part, although in some situations we will assume distributions with compact support.
Imagine now we are given two sequences of $n \times n$-matrices C_n and P_n, where P_n are projection matrices ($P_n^2 = P_n$) with $\tau_n(P_n) \to c$ for some fixed c with $0 < c < 1$. The prescribed eigenvalue distribution of C_n is μ for $n \to \infty$ and we consider the randomly rotated version $B_n = U_n C_n U_n^T$ of this matrix. We are now looking for the moments of the limit distribution of $P_n B_n P_n$. Since all moments of B_n and P_n converge we can write
$$(B_n, P_n) \to (B, P),$$
where B has the prescribed distribution μ, and P is a projection with $\phi[P] = c$. By Voiculescu's theorem about asymptotic freeness, we know that B and P are free. In Section 3.2.2 we need to calculate $\phi[(PBP)^n]$. Due to the trace property of ϕ and the projection property $P^2 = P$, we have
$$\phi\left[(PBP)^n\right] = \phi\left[(BP)^n\right].$$

This situation was already discussed in Speicher [2001], Theorem 4.3. The result is the following

$$\phi[(BP)^n] = \sum_{\pi \in NC(n)} k_\pi[B, \ldots, B] c^{n+1-|\pi|}, \tag{2.6}$$

which is nothing else than a linear combination of the moments of B. We thus can describe $\phi[(BP)^n]$ in terms of $\phi[B^s]$ (s does not exceed n).

2.3 Two Helpful Lemmas

In this section we present two lemmas. The first will help us to make a statement about the average of a random rotated matrix. The second is a result from concentration of measure phenomenon and gives us an idea of how close a functional on a randomly chosen matrix is to its average. The former goes back to Schur [1905, 1906] and thus it is called the Schur Lemma. We will present here a formulation taken from Tung [1985]:

Lemma 2.13 *(Schur Lemma)*
Let $U(G)$ be an irreducible representation of a group G on the vector space V, and T be an arbitrary operator on V. If T commutes with all the operators $\{U(g), g \in G\}$, i.e. $TU(g) = U(g)T$, then T must be a multiple of the identity operator, i.e. $T = \lambda I$ where λ is a number and I the identity.

For our purposes we set V to be the vector space of all real-valued $n \times n$-matrices, i.e. $V = \mathbb{R}^{n \times n}$ and G is the orthogonal group $O(n)$. Let T be the average of the matrices UCU^T for some fixed matrix C, i.e.

$$T = \int_{U \in O(n)} UCU^T d\mu(U)\,.$$

where $d\mu$ denotes the Haar measure on $O(n)$. Since then obviously $T = UTU^T$ for any $U \in O(n)$, we have $TU = UT$. Thus T is commuting with every $U \in O(n)$. With the Schur Lemma it follows that $T = \lambda I$. Due to the cyclic property and the linearity of the trace we have

$$\mathtt{tr}(\lambda I) = \mathtt{tr}(T) = \mathtt{tr}\left(\int_{U \in O(n)} UCU^T d\mu(U)\right) = \mathtt{tr}(C)\,.$$

Therefore

$$\lambda = \frac{1}{n}\mathtt{tr}(C) = \tau_n(C). \tag{2.7}$$

Next, we introduce an important result which goes back to concentration of measure phenomenon (Ledoux [2001]) and is also used in Janzing et al. [2010], but we describe the derivation in more detail here. It shows that "nice" functions on high-dimensional spheres are concentrated around their median or mean.

Lemma 2.14 *(Lévy's Lemma, Ledoux [2001], cf. Popescu et al. [2005])*
Let $g : S_n \to \mathbb{R}$ be a Lipschitz continuous function on the n-dimensional sphere with

$$L := \max_{\gamma \neq \gamma'} \frac{|g(\gamma) - g(\gamma')|}{\|\gamma - \gamma'\|} .$$

If a point γ on S_n is chosen uniformly at random, it satisfies

$$|g(\gamma) - \bar{g}| \leq \epsilon$$

with probability at least $1 - \exp(-\kappa(n-1)\epsilon^2/L^2)$ for some constant κ, where $\bar{g}$ is either the median or the average of $g(\gamma)$.

We will later use Lévy's Lemma to make a statement about the trace of random matrices. Let us here discuss a special case for the function g. Set $g(\gamma) = \langle \gamma, B\gamma \rangle$ for some symmetric $n \times n$-matrix B and an n-dimensional unit vector γ. We will at this point shortly show that then $\bar{g} = \tau_n(B)$ and the Lipschitz constant of g is given by $2\|B\|$, with $\|\cdot\|$ denoting the operator norm

$$\|B\| = \max_{x \in \mathbb{R}^n} \frac{\|Bx\|}{\|x\|}.$$

First we calculate $\bar{g}$. Averaging over γ in S_n is the same as fixing γ and averaging over $U\gamma$ in $O(n)$. Using the Schur Lemma we can write

$$\begin{aligned}
\bar{g} &= \int_{\gamma \in S_n} \langle \gamma, B\gamma \rangle d\mu(\gamma) && (2.8)\\
&= \int_{U \in O(n)} \langle U\gamma, BU\gamma \rangle d\mu(U) && (2.9)\\
&= \langle \gamma, \int_{U \in O(n)} U^T BU d\mu(U)\gamma \rangle && (2.10)\\
&= \langle \gamma, \tau_n(B) I\gamma \rangle && (2.11)\\
&= \tau_n(B). && (2.12)
\end{aligned}$$

Here we applied the Schur Lemma in equation (2.11). Next we compute the Lipschitz constant. To this end let $\gamma \neq \gamma'$. Recall that due to the assumptions $\|\gamma\| = 1 = \|\gamma'\|$.

Then

$$\frac{|g(\gamma) - g(\gamma')|}{\|\gamma - \gamma'\|} = \frac{|\langle \gamma, B\gamma\rangle - \langle \gamma', B\gamma'\rangle|}{\|\gamma - \gamma'\|} \tag{2.13}$$

$$= \frac{|\langle \gamma, B\gamma\rangle - \langle \gamma, B\gamma'\rangle + \langle \gamma', B\gamma\rangle - \langle \gamma', B\gamma'\rangle|}{\|\gamma - \gamma'\|} \tag{2.14}$$

$$= \frac{|\langle \gamma, B(\gamma - \gamma')\rangle + \langle \gamma', B(\gamma - \gamma')\rangle|}{\|\gamma - \gamma'\|} \tag{2.15}$$

$$= \frac{|\langle B(\gamma + \gamma'), \gamma - \gamma'\rangle|}{\|\gamma - \gamma'\|} \tag{2.16}$$

$$\leq \frac{\|B(\gamma + \gamma')\|\|\gamma - \gamma'\|}{\|\gamma - \gamma'\|} \tag{2.17}$$

$$\leq \|B\|\|\gamma + \gamma'\| \tag{2.18}$$

$$\leq 2\|B\| \tag{2.19}$$

where we used the bilinearity of the inner product. In (2.14) and in (2.16) we made use of the symmetry of B and in (2.17) the Cauchy Schwarz inequality.

2.4 Motivating Example

Back to causality, we want to give an intuition about our main idea. Assume that X has a causal influence on Y. To see what kind of suspicious relations between $P(Y)$ and $P(X|Y)$ we can expect we will present a small toy example. The following introductory example is taken from Janzing et al. [2010].

Assume that X is a multivariate Gaussian variable with values in $\mathbb{R}^n$ and isotropic covariance matrix $C_{XX} = I$. Let Y be another $\mathbb{R}^n$-valued variable that is deterministically influenced by X via the linear relation $Y = AX$ for some $n \times n$ matrix A. This induces the covariance matrix

$$C_{YY} = AC_{XX}A^T = AA^T\,.$$

The converse causal hypothesis $Y \to X$ becomes unlikely because $P(Y)$ and $P(X|Y)$ are related in a suspicious way: $P(Y)$ is given by the covariance matrix AA^T whereas $P(X|Y)$ is given by A^{-1} with probability 1. A appears in both descriptions. Another point of view can be made if we look at symmetries: take $U \in O(n)$ where $O(n)$ denotes the orthogonal group and consider the set of covariance matrices $UC_{YY}U^T$. Among them, C_{YY} is very special since it is the only one that is transformed into the isotropic covariance matrix C_{XX}. More general speaking, in the light of the fact of how anisotropic the matrices

$$\tilde{C}_{XX} := A^{-1}UC_{YY}U^TA^{-T}$$

are for randomly chosen U, the hypothetical effect variable is surprisingly isotropic for $U = I$ (we use the short notation $A^{-T} := (A^{-1})^T$). We will show below that this remains true with high probability (in high dimensions) if we even start with an arbitrary covariance matrix C_{XX} and apply a random linear transformation A chosen independently of C_{XX}.

Remark 2.15

We assume Gaussianity here since this is the hardest case and contains a lot of symmetries. Other methods attacking the linear case at least assume non-Gaussian noise like for example Shimizu et al. [2006].

In the general setting

$$Y = AX + E, \quad E \perp\!\!\!\perp X$$

with C_{XX} being a randomly chosen covariance matrix. Since we want to check a certain kind of independence between $A^T A$ and C_{XX} we need to determine A only in terms of X and Y. One sees immediately that A is given by

$$A = C_{YX} C_{XX}^{-1}$$

since $C_{XE} = 0$. C_{YX} denotes here the cross-covariance matrix between X and Y. If we are given a finite amount of samples, all of these covariances have to be estimated. Fixing the dimensionality one can easily find consistent estimators if sample size tends to infinity. But what can be done in the case of ultra-highdimensional variables when only a few samples are given?
The challenge is nonetheless, to find good estimators. In this particular case, when dimensionality exceeds sample size, the empirical estimator of C_{XX} gets singular and thus A cannot be determined reliably. However, as it will turn out it is not necessary to estimate A entirely but only its trace. We will show in the next chapter, that this is still possible for the small sample case in the deterministic setting (without noise term E) but gets quite involved in the noisy setting.

3 Identifiability Results

In this chapter we will prove important identifiability results. They will give us the possibility to distinguish between the right and the wrong direction. Identifiability is very important in causality since it lays the ground for inference algorithms.

Given the true causal model $Y = AX + E$ and the hypothetical causal model for the backward direction $X = \tilde{A}Y + \tilde{E}$, we want to check whether the pair $(C_{YY}, \tilde{A})$ satisfies some suspicious relation that helps us to identify the wrong direction. To this end we compare the values

$$\tau_n(AC_{XX}A^T) \quad \text{and} \quad \tau_n(C_{XX})\tau_n(AA^T)\,. \tag{3.1}$$

First observe that the *expectation* of both values coincide if C_{XX} is randomly drawn from a distribution that is invariant under transformations

$$C_{XX} \mapsto UC_{XX}U^T.$$

To show this we use the linearity of the trace and the derivations we made in the last chapter, Section 2.3:

$$\begin{aligned}\int_{U\in O(n)} \tau_n(AUC_{XX}U^TA^T)d\mu(U) &= \tau_n(A\int_{U\in O(n)} UC_{XX}U^Td\mu(U)A^T)\\ &= \tau_n\big(A(\tau_n(C_{XX})I)A^T\big)\\ &= \tau_n(C_{XX})\tau_n(AA^T)\,.\end{aligned}$$

For our purpose it is decisive that in the typical case $UC_{XX}U^T$ is close to its average, i.e., the two expressions of (3.1) almost coincide. For the theoretical values this was already shown by Janzing et al. [2010]. The challenge is now to show a similar result for the sample-based quantities. If the number of dimensions exceeds the sample size, A cannot be estimated reliably any more on the whole space but only on a subspace. We will denote this estimated A by $\hat{A}$. However, for our purposes it suffices to find a good approximation of the trace of A^TA which is still possible. As it will turn out, this trace can be approximated by the trace of $\hat{A}^T\hat{A}$ times a scaling factor that is dependent only on the rank of the $\hat{A}$.

3.1 Notations

Let us first introduce some notations we will use throughout this work.

- Our causal model for $X \to Y$ is

$$Y = AX + E$$

with X being a random variable with values in $\mathbb{R}^n$, Y, E are random variables with values in $\mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$ and $X \perp\!\!\!\perp E$ (it follows that $C_{XE} = C_{EX} = 0$). E is the noise term. We refer to this model as the *forward* model. If not stated differently we will always assume without loss of generality that $X \to Y$ is the *ground truth*. The *backward* model then is

$$X = \tilde{A}Y + \tilde{E}$$

with the same X, Y and $\tilde{E}$ being a random variable in $\mathbb{R}^n$, $\tilde{A} \in \mathbb{R}^{n \times m}$.

- For a matrix C we denote the empirical estimator by $\hat{C} = (\hat{c}_{ij})_{i,j \in \{1,\dots,n\}}$. To estimate the covariance matrix we use the standard estimator which is given by

$$\hat{c}_{ij} = \frac{1}{k-1} \sum_{l=1}^{k} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)$$

where

$$\bar{x}_i = \frac{1}{k} \sum_{l=1}^{k} x_{il}$$

denotes the sample mean, n is the dimensionality of C and k is the number of samples.

- We say the two matrices B and C are *independent* if for two fixed matrices B and D, C is a random rotation of D, i.e., $C = UDU$ with U drawn from $O(n)$ at random according to the Harr measure. C_{XX} will be this randomly rotated matrix, whereas we assume $A^T A$ to be fixed. If $X \to Y$ we will always assume that $A^T A$ and C_{XX} are independent.

- In the next section we will consider sequences of matrices with finite moments of all orders. Given a $n \times n$-matrix B_n we can talk about a real-valued random variable Z_n with probability measure μ_{B_n} as in (2.3). Z_n then reflects the distribution of the eigenvalues of B_n.

- We will denote the sample size with k and the rank of C_{XX} with r. We assume $k \leq n$. Thus if $P(X)$ is a density, almost surely $r = k - 1$, since for k given i.i.d. samples of X, the rank of its covariance matrix is $k - 1$.

- When we write $\lim_{n \to \infty}$ in the following part we mean convergence in probability. We say, a sequence of random variables X_n *converges in probability* towards X if for all $\epsilon > 0$

$$\lim_{n \to \infty} \Pr(|X_n - X| \geq \epsilon) = 0.$$

3.2 The Deterministic (Noiseless) Case

We will now discuss the deterministic case, i.e., our model does not contain a noise term E. Let $A_{(m,n)}$ be a sequence of (deterministic) $m \times n$-matrices such that the limit distribution of $A_{(m,n)}^T A_{(m,n)}$ exists. This formulation is from now on always meant in the spirit of free probability theory, i.e., we claim that

$$\lim_{m,n\to\infty} \tau_n\big((A_{(m,n)}^T A_{(m,n)})^s\big)$$

exist for all s. One can ignore the expectation here since the entries of $A_{(m,n)}$ are fixed. Referring to this we will also say that *all moments of* $A_{(m,n)}^T A_{(m,n)}$ *converge*.
Let $X_{(n)}$ be a sequence of n-dimensional multivariate real-valued random variables such that the normalized trace $\tau_n(C_{XX})$ converges for $n \to \infty$.
Let $Y_{(m)}$ be a sequence of $\mathbb{R}^m$-valued variables that is deterministically influenced by $X_{(n)}$ via the linear relation

$$Y_{(m)} = A_{(m,n)} X_{(n)} \,.$$

Finally, let $\hat{C}_{XX}^{(n)}$ be the sequence of empirical covariances for k randomly observed samples (x_i, y_i) $(i = 1, \ldots, k)$. As it will be shown in Lemma 3.9, the normalized trace of $\hat{C}_{XX}^{(n)}$ then converges independently from n, if $k \to \infty$ (the same is true for $\hat{C}_{YY}^{(m)}$ and m).

Since the case where $k \gg m, n$ was already discussed in Janzing et al. [2010] we want to focus here on the more difficult case where $k \leq \min(m, n)$. Note that in this case, if C_{XX} has full rank, almost surely

$$\texttt{rank}(\hat{C}_{XX}) = k - 1 =: r \,.$$

Thus we can assume without loss of generality that $\texttt{rank}(\hat{C}_{XX}) = \texttt{rank}(\hat{C}_{YY}) = r$. In this section we want to investigate what happens if $m, n \to \infty$ and $k \leq \min(m, n)$. To deal with this we assume that m, n and k go to infinity with the same rate. Thus we have the relations $m = \kappa n$ for some constant κ, and $r/\min(m, n)$ is fixed, i.e.

$$\frac{r}{\min(m, n)} \to c \quad \text{with} \quad 0 < c < 1 \quad \text{as} \quad m, n \to \infty \,.$$

We will discuss at the end of this section what will happen if $r/\min(m, n) \to 0$. Since m, n and r are now dependent only on one variable we set $\rho := \min(m, n)$ and denote the matrices $A_{(m,n)}$ with A_ρ. We will also index all other matrices dependent on both m and n with the sub- or superscript ρ, like for example C_{XY}^ρ. On the other hand, if a matrix is dependent only on either m or n we keep the index $^{(m)}$ and $^{(n)}$, respectively.

3.2.1 Establishing an Equality in the Forward Direction

Let

$$\hat{C}_{XX}^{(n)} = U_{(n)} \Sigma_{(n)} U_{(n)}^T$$

denote the reduced singular value decomposition of the empirical covariance matrix $\hat{C}^{(n)}_{XX}$. $U_{(n)}$ is given by $(u_1, \ldots, u_r)_{(n)}$ with orthonormal vectors $u_i^{(n)}$ and $\Sigma_{(n)}$ is a diagonal $r \times r$-matrix with the non-zero eigenvalues $\sigma_i^{(n)}$ of $\hat{C}^{(n)}_{XX}$ on its diagonal. One then can also write

$$\hat{C}^{(n)}_{XX} = \sum_{i=1}^{k} \sigma_i u_i u_i^T$$

(here we omitted the indices $^{(n)}$ for better readability, but remember that also u_i and σ_i depend on n). We can now estimate A_ρ by

$$\hat{A}_\rho = \hat{C}^{\rho}_{YX} \hat{C}^{+(n)}_{XX} = A_\rho \hat{C}^{(n)}_{XX} \hat{C}^{+(n)}_{XX} = A_\rho U_{(n)} U^T_{(n)} \tag{3.2}$$

where $\hat{C}^{+(n)}_{XX}$ denotes the Moore-Penrose pseudoinverse of $\hat{C}^{(n)}_{XX}$, given by

$$\hat{C}^{+(n)}_{XX} = U_{(n)} \Sigma^{-1}_{(n)} U^T_{(n)}.$$

We can establish our first result for the forward direction.

Theorem 3.1 *(Equality in the forward direction)*
Assume that either

a) the growing rate of the operator norm of $A_\rho^T A_\rho$ is bounded by n, i.e., $\|A_\rho^T A_\rho\|/n < \infty$ for $n, \rho \to \infty$ or

b) the eigenvalue distribution of both $C^{(n)}_{XX}$ and $A_\rho^T A_\rho$ converge to some probability density μ and ν, respectively.

Then, if $A_\rho^T A_\rho$ and $C^{(n)}_{XX}$ are independent and $0 < r/\rho \to c < 1$ for $\rho \to \infty$,

$$\lim_{\rho \to \infty} \left[\tau_n(\hat{A}_\rho^T \hat{A}_\rho \hat{C}^{(n)}_{XX}) - \frac{n}{r} \tau_n(\hat{A}_\rho^T \hat{A}_\rho) \tau_n(\hat{C}^{(n)}_{XX}) \right] = 0\,. \tag{3.3}$$

We will provide two different proofs for this lemma. One "classical" one under the assumption a) which does not use free probability but Lévy's Lemma instead. The other one, assuming b), will show how helpful free probability theory can be in this kind of setting.

Proof without free probability theory (a): Due to the following property of the Moore-Penrose pseudoinverse

$$DD^+D = D$$

for some matrix D, we get

$$\hat{A}_\rho \hat{C}^{(n)}_{XX} \hat{A}_\rho^T = A_\rho \hat{C}^{(n)}_{XX} \hat{C}^{+(n)}_{XX} \hat{C}^{(n)}_{XX} \hat{C}^{+(n)}_{XX} \hat{C}^{(n)}_{XX} A_\rho^T = A_\rho \hat{C}^{(n)}_{XX} A_\rho^T\,.$$

Thus

$$\tau_n(\hat{A}_\rho^T \hat{A}_\rho \hat{C}^{(n)}_{XX}) = \tau_n(A_\rho^T A_\rho \hat{C}^{(n)}_{XX}) \tag{3.4}$$

by the cyclic property of the trace.
We set $B_\rho := A_\rho^T A_\rho$ and derive

$$\lim_{\rho\to\infty}\left[\tau_n(A_\rho^T A_\rho \hat{C}_{XX}^{(n)}) - \frac{n}{r}\tau_n(\hat{A}_\rho^T \hat{A}_\rho)\tau_n(\hat{C}_{XX}^{(n)})\right] \tag{3.5}$$

$$= \lim_{\rho\to\infty}\left[\tau_n(B_\rho \hat{C}_{XX}^{(n)}) - \frac{n}{r}\tau_n(U_{(n)}U_{(n)}^T A_\rho^T A_\rho U_{(n)}U_{(n)}^T)\tau_n(\hat{C}_{XX}^{(n)})\right] \tag{3.6}$$

$$= \lim_{\rho\to\infty}\left[\tau_n(B_\rho \hat{C}_{XX}^{(n)}) - \frac{n}{r}\tau_n(B_\rho U_{(n)}U_{(n)}^T)\tau_n(\hat{C}_{XX}^{(n)})\right] \tag{3.7}$$

$$= \lim_{\rho\to\infty}\left[\tau_n(B_\rho \sum_{i=1}^r \sigma_i u_i u_i^T) - \frac{n}{r}\tau_n(B_\rho \sum_{i=1}^r u_i u_i^T)\tau_n(\hat{C}_{XX}^{(n)})\right] \tag{3.8}$$

$$= \lim_{\rho\to\infty}\left[\tau_n(B_\rho \sum_{i=1}^r \sigma_i u_i u_i^T) - \tau_n(B_\rho \sum_{i=1}^r \frac{n}{r}\tau_n(\hat{C}_{XX}^{(n)}) u_i u_i^T)\right] \tag{3.9}$$

$$= \lim_{\rho\to\infty}\left[\tau_n\big(B_\rho \sum_{i=1}^r \sigma_i u_i u_i^T - B_\rho \sum_{i=1}^r \frac{n}{r}\tau_n(\hat{C}_{XX}^{(n)}) u_i u_i^T\big)\right] \tag{3.10}$$

$$= \lim_{\rho\to\infty}\left[\tau_n\big(B_\rho \sum_{i=1}^r (\sigma_i - \frac{n}{r}\tau_n(\hat{C}_{XX}^{(n)})) u_i u_i^T\big)\right] \tag{3.11}$$

$$= \lim_{\rho\to\infty}\left[\sum_{i=1}^r (\sigma_i - \tau_r(\hat{C}_{XX}^{(n)}))\tau_n(B_\rho u_i u_i^T)\right] \tag{3.12}$$

$$= \lim_{\rho\to\infty} \sum_{i=1}^r (\sigma_i - \tau_r(\hat{C}_{XX}^{(n)})) \lim_{\rho\to\infty} \tau_n(B_\rho u_i u_i^T)\,. \tag{3.13}$$

In (3.7) we used the cyclic property of the trace and that $P := U_{(n)}U_{(n)}^T$ is a projection and thus $P^2 = P$. In (3.8) we substituted $U_{(n)}U_{(n)}^T$ by $\sum_{i=1}^k u_i u_i^T$. In (3.9) and (3.10) we used the linearity of the trace. The change of multiplication and limes in (3.13) is possible since all relevant quantities converge. Also we used the notation $\tau_r(\cdot) = \mathtt{tr}(\cdot)/r$.

Now we apply Lévy's Lemma (Lemma 2.14): define the function $f(u) := \langle u, B_\rho u\rangle = \mathtt{tr}(B_\rho u u^T)$. Then $\bar{f} = \tau_n(B_\rho)$ if u has norm 1 and is randomly chosen according to a rotation invariant prior (see (2.7)). We can assume this for the u_i's since $\hat{C}_{XX}^{(n)}$ and B_ρ are independent. It follows

$$|\tau_n(B_\rho u_i u_i^T) - \frac{1}{n}\tau_n(B_\rho)| \le \frac{\epsilon}{n}\|B_\rho\| \tag{3.14}$$

with probability at least $1 - \exp(-\kappa(n-1)\epsilon^2)$ (replace ϵ with $\epsilon\|B_\rho\|$ in Lemma 2.14). Now, due to our assumption *a)*, $\|B_\rho\|/n < \infty$ for all ρ, we get for every $\epsilon > 0$

$$\lim_{n\to\infty} \Pr\left(\left|\tau_n(B_\rho u_i u_i^T) - \frac{1}{n}\tau_n(B_\rho)\right| \ge \epsilon\right) = 0\,.$$

Because of

$$\sum_{i=1}^r (\sigma_i - \tau_r(\hat{C}_{XX}^{(n)})) = \sum_{i=1}^r \sigma_i - r\frac{1}{r}\mathtt{tr}(\hat{C}_{XX}^{(n)})) = 0$$

it is clear that (3.13) is zero. This proves the lemma. □

Proof with free probability theory (b): For a sequence of real valued deterministic $n \times n$-matrices D_n it holds

$$\lim_{n\to\infty} \tau_n(D_n) = \lim_{n\to\infty} \tau_n \otimes \mathbb{E}(B_n)\,,$$

i.e., if $D_n \to D$,

$$\phi(D) = \lim_{n\to\infty} \tau_n(D_n)\,.$$

We use (3.4) and set again $B_\rho = A_\rho^T A_\rho$. We want to apply the second part of Voiculescu's theorem (Theorem 2.12). B_ρ and $\hat{C}_{XX}^{(n)}$ fulfill its assumptions: Both their eigenvalue distributions tend to some fixed probability measure μ and ν, respectively, and $\hat{C}_{XX}^{(n)}$ can be viewed as some randomly rotated matrix. Ergo B_ρ and $\hat{C}_{XX}^{(n)}$ become free in the limit $\rho, n \to \infty$. We set

$$B = \lim_{\rho\to\infty} B_\rho$$

and

$$C = \lim_{\rho\to\infty} \hat{C}_{XX}^{(n)}\,.$$

Let us define $P_\rho = \hat{C}_{XX}^{(n)} \hat{C}_{XX}^{+(n)}$. If $\hat{C}_{XX}^{(n)}$ is randomly rotated against B_ρ, then also P_ρ. Due to the assumption that $r/\rho \to c$ we get $\tau_n(P_\rho) \to c$ for $n \to \infty$ and thus, since $c < 1$, all moments of P_ρ converge. Again by Theorem 2.12, P_ρ and B_ρ are free in the limit. We can set

$$P = \lim_{\rho\to\infty} P_\rho\,.$$

Let us conclude for a moment what we got so far: B and C as well as B and P are free with respect to ϕ. By a simple property of free variables (cf. Equation (2.4)) we have

$$\phi(BC) = \phi(B)\phi(C) \quad \text{and} \quad \phi(BP) = \phi(B)\phi(P)$$

and thus

$$\begin{aligned}\lim_{\rho\to\infty} \tau_n(A_\rho \hat{C}_{XX}^{(n)} A_\rho^T) = \phi(BC) = \phi(B)\phi(C) = \frac{\phi(BP)\phi(C)}{\phi(P)} = \frac{n}{r}\phi(PBP)\phi(C) \\ = \lim_{\rho\to\infty} \frac{n}{r} \tau_n(\hat{A}_\rho^T \hat{A}_\rho)\tau_n(\hat{C}_{XX}^{(n)})\end{aligned}$$

since $\phi(P) = r/n$. Note that we used the trace property of ϕ and $P^2 = P$. □

As one can see, the proof using free probability theory is a bit shorter. However, we used different assumptions for both proofs and it seems that assumption a) is a bit weaker. To see that, consider a sequence of $n \times n$-matrices A_n where every entry is drawn i.i.d. from a standard normal distribution. Then $\tau_n(A_n^T A_n) = n\hat{\sigma}(x_{i,j}) \approx n$ and $\|A_n^T A_n\|/n < \infty$ for all n. To obtain converging moments this sequence would need some normalization on the entries and thus it has no limit distribution for $n \to \infty$. Hence it violates assumption b).

Theorem 3.1 shows (except from the quotient r/n which comes from the projection) that for two independent matrices A and B the renormalized trace is approximately multiplicative, i.e.,

$$\tau_n(AB) \approx \tau_n(A)\tau_n(B)\,.$$

We will refer to this as the *trace multiplicativity*.

3.2.2 Establishing an Inequality in the Backward Direction

To get an idea what happens in the backward direction, we first want to give an intuition why we expect a violation of the trace multiplicativity there, more precisely, why the values of (3.1) won't coincide. Since this is very important, we will formulate our permanent assumption once again: The sequence $A_\rho^T A_\rho$ is fixed and the matrices $C_{XX}^{(n)}$ are chosen at random from a distribution that is invariant under rotations. We will say $A_\rho^T A_\rho$ and $C_{XX}^{(n)}$ are independent. Another assumption we want to make in this part is the assumption *b*) of Theorem 3.1, namely that the eigenvalue distributions of both $C_{XX}^{(n)}$ and $A_\rho^T A_\rho$ converge to some probability density.
We first describe how Janzing et al. [2010] show this for the large sample case and then explain why this cannot be applied in our case. Let for a moment $m = n$ and A be invertible. Given the forward model with $A^T A$ and C_{XX} independent and a nontrivial eigenvalue distribution on $A^T A$, we have (using the shorthand $A^{-T} = (A^{-1})^T$)

$$\begin{aligned}
\tau_n(A^{-T}A^{-1}C_{YY}) - \tau_n(A^{-T}A^{-1})\tau_n(C_{YY}) &= \tau_n(C_{XX}) - \tau_n(A^{-T}A^{-1})\tau_n(A^T A C_{XX}) \\
&\approx \tau_n(C_{XX}) - \tau_n(A^{-T}A^{-1})\tau_n(A^T A)\tau_n(C_{XX}) \\
&= \tau_n(C_{XX})(1 - \tau_n(A^{-T}A^{-1})\tau_n(A^T A)) \\
&= \tau_n(C_{XX})\mathtt{Cov}(1/Z, Z) \\
&< 0,
\end{aligned}$$

where Z is a real-valued random variable whose probability measure is given by

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_1}\,,$$

with λ_i denoting the eigenvalues of $A^T A$. Thus, Z reflects the empirical distribution of the eigenvalues of $A^T A$ and $A^{-T}A^{-1}$, respectively (cf. equation (2.3), see also Janzing et al. [2010], Theorem 2).
Why could we get in trouble showing this for the sample-based quantities? First of all, the estimators of A and $\tilde{A}$ do not have full rank and therefore are not inverse to each other. But we will see that there is a similar nice relationship between these two estimators which is shown in Lemma 3.3. The other restraint could be that we don't know what happens with the distribution of Z if we are given only a small number of samples compared to the number of dimensions. But this distribution is crucial to establish the asymmetry since we need to use the fact that $\mathtt{Cov}(Z, 1/Z) < 0$. With this problem we deal in Lemma 3.4. We start with a relationship between covariance and cross-covariance matrices.

Lemma 3.2

Let $k = r + 1$ be the number of samples and let $\mathtt{rank}(\hat{C}_{XX}) = \mathtt{rank}(\hat{C}_{YY}) = r < \rho = \min\{\dim(\hat{C}_{XX}), \dim(\hat{C}_{YY})\}$. Then

$$\hat{C}_{XY}\hat{C}_{YY}^{+}\hat{C}_{YX} = \hat{C}_{XX}$$

and

$$\hat{C}_{YX}\hat{C}_{XX}^{+}\hat{C}_{XY} = \hat{C}_{YY}.$$

Proof: Let X and Y be the $n \times k$- and $m \times k$-matrix of observations, respectively. We define the centralized observation matrices $\bar{X} = X - \mu \mathtt{1}_k^T$ and $\bar{Y} = Y - \nu \mathtt{1}_k^T$ where μ and ν are the row mean of X and Y, respectively, and $\mathtt{1}_k$ is the k-dimensional column vector containing only 1's. Then

$$\hat{C}_{XY} = \frac{1}{r}\bar{X}\bar{Y}^T, \quad \hat{C}_{YY}^{+} = \frac{1}{r}(\bar{Y}\bar{Y}^T)^{+} \quad \text{and} \quad \hat{C}_{YX} = \frac{1}{r}\bar{Y}\bar{X}^T.$$

We use a property of the pseudoinverse

$$\bar{Y}^T(\bar{Y}\bar{Y}^T)^{+} = \bar{Y}^{+}$$

and get

$$\hat{C}_{XY}\hat{C}_{YY}^{+}\hat{C}_{YX} = \frac{1}{r}\bar{X}\bar{Y}^T(\bar{Y}\bar{Y}^T)^{+}\bar{Y}\bar{X}^T = \frac{1}{r}\bar{X}\bar{Y}^{+}\bar{Y}\bar{X}^T.$$

$\bar{Y}^{+}\bar{Y}$ is an orthogonal projector onto the range of $\bar{Y}^T$ which agrees with the orthogonal complement of the kernel of $\bar{Y}$. We already know that the kernel of $\bar{Y}$ is one-dimensional (since $\bar{Y}\bar{Y}^T$ has rank r which is one smaller then the number of columns of $\bar{Y}$). Let $Y = (y_{ij})_{m\times k}$ and $\bar{y}_i = \frac{1}{k}\sum_{j=1}^{k} y_{ij}$. Now for any row $\bar{Y}_i$ of $\bar{Y}$

$$\bar{Y}_i \mathtt{1}_k = \sum_{j=1}^{k}(y_{ij} - \bar{y}_i) = \sum_{j=1}^{k} y_{ij} - \sum_{l=1}^{k}\frac{1}{k}\sum_{j=1}^{k} y_{ij} = 0$$

which implies that $\mathtt{ker}(\bar{Y})$ is spanned by $\mathtt{1}_k$. The same is true for $\mathtt{ker}(\bar{X})$, i.e., these two kernels coincide. $\mathtt{range}(\bar{X}^T)$ and $\mathtt{range}(\bar{Y}^T)$ are mappings from $\mathbb{R}^n \to \mathbb{R}^k$ and $\mathbb{R}^m \to \mathbb{R}^k$, respectively. We now have $\mathtt{range}(\bar{Y}^T) = \mathtt{1}_k = \mathtt{range}(\bar{X}^T)$ and so $\bar{Y}^{+}\bar{Y}$ is the identity on $\mathtt{range}(\bar{X}^T)$. Thus $\bar{Y}^{+}\bar{Y}\bar{X}^T = \bar{X}^T$. Changing the roles of X and Y proves the second inequality. The statement follows. □

Now we look at the estimators of the transfer matrices more carefully. Similar to the forward direction we estimate $\tilde{A}$ in the backward model by

$$\hat{\tilde{A}}_\rho = \hat{C}_{XY}^{\rho}\hat{C}_{YY}^{+(n)}. \tag{3.15}$$

First observe that $\hat{\tilde{A}}\hat{\tilde{A}}^T$ is the pseudoinverse of $\hat{A}^T\hat{A}$ (index $_\rho$ omitted). This is shown by the next lemma.

Lemma 3.3
Let $\hat{A}$ and $\tilde{A}$ be as defined in (3.2) and (3.15), respectively. Then it holds

$$(\hat{\tilde{A}}\hat{\tilde{A}}^T)^+ = \hat{A}^T\hat{A}\,. \tag{3.16}$$

Proof: We have to show the four properties characterizing the Moore-Penrose pseudoinverse:
(1) and (2): Obviously $\hat{\tilde{A}}\hat{\tilde{A}}^T\hat{A}^T\hat{A}$ and $\hat{A}^T\hat{A}\hat{\tilde{A}}\hat{\tilde{A}}^T$ are both symmetric.
(3) With the help of Lemma 3.2 we derive

$$\begin{aligned}
\hat{\tilde{A}}\hat{\tilde{A}}^T\hat{A}^T\hat{A}\hat{\tilde{A}}\hat{\tilde{A}}^T &= \hat{C}_{XY}\hat{C}_{YY}^+\hat{C}_{YY}^+\hat{C}_{YX}\hat{C}_{XX}^+\hat{C}_{XY}\hat{C}_{YX}\hat{C}_{XX}^+\hat{C}_{XY}\hat{C}_{YY}^+\hat{C}_{YY}^+\hat{C}_{YX}\\
&= \hat{C}_{XY}\hat{C}_{YY}^+\hat{C}_{YY}^+\hat{C}_{YY}\hat{C}_{YY}\hat{C}_{YY}^+\hat{C}_{YY}^+\hat{C}_{YX}\\
&= \hat{C}_{XY}\hat{C}_{YY}^+\hat{C}_{YY}^+\hat{C}_{YX}\\
&= \hat{\tilde{A}}\hat{\tilde{A}}^T
\end{aligned}$$

(4)

$$\begin{aligned}
\hat{A}^T\hat{A}\hat{\tilde{A}}\hat{\tilde{A}}^T\hat{A}^T\hat{A} &= \hat{C}_{XX}^+\hat{C}_{XY}\hat{C}_{YX}\hat{C}_{XX}^+\hat{C}_{XY}\hat{C}_{YY}^+\hat{C}_{YY}^+\hat{C}_{YX}\hat{C}_{XX}^+\hat{C}_{XY}\hat{C}_{YX}\hat{C}_{XX}^+ && (3.17)\\
&= \hat{C}_{XX}^+\hat{C}_{XY}\hat{C}_{YY}\hat{C}_{YY}^+\hat{C}_{YY}^+\hat{C}_{YY}\hat{C}_{YX}\hat{C}_{XX}^+ && (3.18)\\
&= \hat{C}_{XX}^+\hat{C}_{XY}\hat{C}_{YY}\hat{C}_{YY}^+\hat{C}_{YX}\hat{C}_{XX}^+ && (3.19)\\
&= \hat{C}_{XX}^+\hat{C}_{XY}\hat{C}_{YX}\hat{C}_{XX}^+ && (3.20)\\
&= \hat{A}^T\hat{A} && (3.21)
\end{aligned}$$

where (3.20) follows from the fact that $\hat{C}_{YY}\hat{C}_{YY}^+$ is an orthogonal projector onto the range of $\hat{C}_{YY}$ which coincides with the range of $\hat{C}_{YX}$ and thus

$$\hat{C}_{YY}\hat{C}_{YY}^+\hat{C}_{YX} = \hat{C}_{YX}$$

(cf. also the proof of Lemma 3.2). This proves the Lemma. □

Next we introduce a sequence of random variables Z_ρ whose distribution is the empirical distribution of the *non-zero* eigenvalues of $\hat{A}_\rho^T\hat{A}_\rho$, which are all positive. Their probability measures are given by

$$\mu_\rho = \frac{1}{r}\sum_{i=1}^{r}\delta_{\lambda_i(\rho)}$$

where the $\lambda_i(\rho)$ denote the r non-zero eigenvalues of $\hat{A}_\rho^T\hat{A}_\rho$. Then, because of Lemma 3.3, $1/Z_\rho$ reflects the empirical distribution of the non-zero eigenvalues of $\hat{\tilde{A}}_\rho^T\hat{\tilde{A}}_\rho$. Since we assumed independence of $C_{XX}^{(n)}$ and $A_\rho^TA_\rho$, as in the second proof of Theorem 3.1 we get that $\hat{C}_{XX}^{(n)}$ and $\hat{A}_\rho^T\hat{A}_\rho$ are asymptotically free and furthermore that $\hat{C}_{XX}^{(n)}\hat{C}_{XX}^{+(n)}$ and $\hat{A}_\rho^T\hat{A}_\rho$ are asymptotically free. Thus, again by setting $P := \lim_{\rho\to\infty}\hat{C}_{XX}^{(n)}\hat{C}_{XX}^{+(n)}$ and $B := \lim_{\rho\to\infty}A_\rho^TA_\rho$, P and B are free.

Set now

$$Z := \lim_{\rho\to\infty} Z_\rho \,. \tag{3.22}$$

Then

$$\mathbb{E}(Z) = \lim_{\rho\to\infty} \tau_r(\hat{A}_\rho^T \hat{A}_\rho) = \frac{n}{r}\phi[PBP] = \frac{n}{r}\phi[BP] = \frac{n}{r}\phi[B]\phi[P] = \phi[B] \,. \tag{3.23}$$

This follows from $\tau_n(P_\rho) = r/n$ for all n ($r/n \to c$ in the limit $\rho \to \infty$). Also

$$\mathbb{E}(Z^2) = \lim_{\rho\to\infty} \tau_r\left((\hat{A}_\rho^T \hat{A}_\rho)^2\right) = \frac{n}{r} \lim_{n\to\infty} \tau_n\left((P_\rho B_\rho P_\rho)^2\right) = \frac{n}{r}\phi\left[(BP)^2\right] \,. \tag{3.24}$$

Next we prove two statements about the variance of Z that are important for the proof of the theorem for the backward direction.

Lemma 3.4

Let the limit distribution of the eigenvalues of $A_\rho^T A_\rho$ have non-zero variance. Let $0 < r/\rho \to c < 1$ for $\rho \to \infty$. Then, if Z is defined like above (see Equation (3.22) and the paragraph before)

$$\texttt{Var}(Z) > 0.$$

Remark 3.5

By *limit distribution of the eigenvalues of* $A_\rho^T A_\rho$ the following is meant: We can assign a random variable to each matrix $A_\rho^T A_\rho$ whose probability measure is determined through their eigenvalues by (2.3). Due to the assumptions made on the sequence A_ρ, these random variables converge to a limit random variable whose distribution we call the *limit distribution of the eigenvalues of* $A_\rho^T A_\rho$.

Proof: To investigate $\texttt{Var}(Z)$ we use free probability theory. We use Equations (3.23) and (3.24) and get

$$\texttt{Var}(Z) = \mathbb{E}(Z^2) - \mathbb{E}(Z)^2 = \frac{1}{c}\phi[(PB)^2] - \phi[B]^2 \,.$$

With (2.5) and $c = \phi[P]$ we calculate

$$\begin{aligned}
\texttt{Var}(Z) &= \frac{1}{c}\phi[(PB)^2] - \phi[B]^2 = \frac{1}{c}\phi[PBPB] - \phi[B]^2 && (3.25)\\
&= \frac{1}{c}\big(\phi[PP]\phi[B]\phi[B] + \phi[P]\phi[P]\phi[BB] - \phi[P]\phi[B]\phi[P]\phi[B]\big) - \phi[B]^2 && (3.26)\\
&= \phi[B]^2 + c\phi[B^2] - c\phi[B]^2 - \phi[B]^2 && (3.27)\\
&= c(\phi[B^2] - \phi[B]^2) && (3.28)\\
&= c\texttt{Var}(\tilde{Z}) && (3.29)\\
&\neq 0 \,. && (3.30)
\end{aligned}$$

$\tilde{Z}$ denotes here a random variable whose distribution is the limit distribution of the eigenvalues of $A_\rho^T A_\rho$. □

Lemma 3.6

Let Z be a random variable that takes only positive values. Then

$$\texttt{Var}(Z) \neq 0 \quad \Longrightarrow \quad \texttt{Cov}(Z, 1/Z) < 0.$$

Proof: If $\texttt{Var}(Z) \neq 0$ then Z is not constant. Therefore with Cauchy-Schwarz

$$1 = \mathbb{E}(\sqrt{Z}\sqrt{1/Z}) \leq \mathbb{E}(Z)\mathbb{E}(1/Z)$$

with equality if and only if Z and $1/Z$ are linearly dependent (f.e. Bickel and Doksum [1991]) which is not the case here. Thus

$$0 > 1 - \mathbb{E}(Z)\mathbb{E}(1/Z) = \texttt{Cov}(Z, 1/Z)$$

□

Finally we derive a systematic violation of the multiplicativity of the traces in the backward direction and describe how to compute the exact values of the emerging difference in the limit.

Theorem 3.7 *(Inequality in the backward direction)*

Assume that the eigenvalue distribution of both $C_{XX}^{(n)}$ and $A_\rho^T A_\rho$ tend to some probability measure as $\rho \to \infty$. Let $A_\rho^T A_\rho$ and $C_{XX}^{(n)}$ be independent and $0 < r/\rho \to c < 1$ for $\rho \to \infty$. Let the limit distribution of eigenvalues of $A_\rho^T A_\rho$ have non-zero variance. Then

$$\lim_{\rho\to\infty} \left[\tau_m(\hat{\tilde{A}}_\rho^T \hat{\tilde{A}}_\rho \hat{C}_{YY}^{(m)}) - \frac{m}{r}\tau_m(\hat{\tilde{A}}_\rho^T \hat{\tilde{A}}_\rho)\tau_m(\hat{C}_{YY}^{(m)})\right] < 0. \tag{3.31}$$

Proof: Due to Lemma 3.2 it holds

$$\hat{\tilde{A}}_\rho \hat{C}_{YY}^{(m)} \hat{\tilde{A}}_\rho^T = \hat{C}_{XY}^{\rho} \hat{C}_{YY}^{+(m)} \hat{C}_{YY}^{(m)} \hat{C}_{YY}^{+(m)} \hat{C}_{YX}^{\rho} = \hat{C}_{XY}^{\rho} \hat{C}_{YY}^{+(m)} \hat{C}_{YX}^{\rho} = \hat{C}_{XX}^{(n)}\,. \tag{3.32}$$

Therefore, with the cyclic property of the trace,

$$\lim_{\rho\to\infty} \left(\tau_m(\hat{\tilde{A}}_\rho^T \hat{\tilde{A}}_\rho \hat{C}_{YY}^{(m)}) - \tau_r(\hat{\tilde{A}}_\rho^T \hat{\tilde{A}}_\rho)\tau_m(\hat{C}_{YY}^{(m)})\right) \tag{3.33}$$

$$= \lim_{\rho\to\infty} \tau_m(\hat{C}_{XX}^{(n)}) - \lim_{\rho\to\infty} \tau_r(\hat{\tilde{A}}_\rho^T \hat{\tilde{A}}_\rho) \lim_{\rho\to\infty} \tau_m(A_\rho \hat{C}_{XX}^{(n)} A_\rho^T) \tag{3.34}$$

$$= \lim_{\rho\to\infty} \tau_m(\hat{C}_{XX}^{(n)}) - \lim_{\rho\to\infty} \tau_r(\hat{\tilde{A}}_\rho^T \hat{\tilde{A}}_\rho)\tau_r(\hat{A}_\rho^T \hat{A}_\rho)\tau_m(\hat{C}_{XX}^{(n)}) \tag{3.35}$$

$$= \lim_{\rho\to\infty} \left[\left(1 - \tau_r(\hat{\tilde{A}}_\rho \hat{\tilde{A}}_\rho^T)\tau_r(\hat{A}_\rho^T \hat{A}_\rho)\right)\tau_m(\hat{C}_{XX}^{(n)})\right] \tag{3.36}$$

$$= (1 - \mathbb{E}(Z)\mathbb{E}(1/Z)) \lim_{\rho\to\infty} \frac{n}{m}\tau_n(\hat{C}_{XX}^{(n)}) \tag{3.37}$$

$$= \texttt{Cov}(Z, 1/Z) \lim_{\rho\to\infty} \frac{n}{m}\tau_n(\hat{C}_{XX}^{(n)}) \tag{3.38}$$

$$< 0\,. \tag{3.39}$$

In (3.34) on the left side we used (3.32) and in (3.35) Theorem 3.1. (3.39) follows from Lemma 3.4, Lemma 3.6 and the fact that $\tau_n(\hat{C}_{XX})$ is always positive. Remember that n/m is a constant. □

Remark 3.8

Theorem 3.7 already shows that for a given model $Y = AX$ the strength of the violation of the trace multiplicity in the backward direction depends on the eigenvalue distribution of $A^T A$, in particular on its variance, and on the quotient r/ρ.

We have just shown that in contrast to the forward direction we found a significant violation of the trace multiplicity in the backward direction. It is determined by the covariance of the eigenvalue distribution of A estimated on an r-dimensional subspace and the inverse of this distribution. The result is similar to that shown in Janzing et al. [2010] with the difference that due to the small sample setting it is restricted to the subspace which is spanned by the samples. Thereby one expects the violation to be weaker here than in the large sample case.

A natural question which arises now is whether we can determine the term $\mathbb{E}(1/Z)$ any further with the objective of getting an idea about the strength of the violation of the trace multiplicity. Such a result also would help us to construct a statistical test. It turns out that this is in principle possible but we end up with an infinite sum that only depends on the moments of $A_\rho^T A_\rho$. In the following part we assume that the eigenvalue distribution of $A_\rho^T A_\rho$ tends to some fixed probability measure with compact support.

The steps are the following: Since we want to compute $\mathbb{E}(1/Z)$ we first describe $1/Z$ in terms of Z by means of the well known geometric series because we can calculate the moments of Z. Then we apply free probability theory, in particular Equation (2.6).
The geometric series is given by

$$\frac{1}{1-q} = \sum_{l=0}^{\infty} q^l \quad \text{for} \quad |q| < 1\,.$$

It follows for $x = 1 - q$

$$\frac{1}{x} = \sum_{l=0}^{\infty} (1-x)^l \quad \text{for} \quad |1-x| < 1\,.$$

Z is a positive real-valued random variable. Recall our assumption that it has compact support, i.e., we can find a constant d such that $|1 - Z/d| < 1$. Thus we can assume without loss of generality $|1 - Z| < 1$. We get

$$\mathbb{E}(1/Z) = \mathbb{E}\left(\sum_{l=0}^{\infty}(1-Z)^l\right) = \mathbb{E}\left(\sum_{l=0}^{\infty}\sum_{s=0}^{l}\binom{l}{s}(-Z)^s\right) = \sum_{l=0}^{\infty}\sum_{s=0}^{l}\binom{l}{s}(-1)^s\mathbb{E}(Z^s)\,. \tag{3.40}$$

The terms $\mathbb{E}(Z^s)$ can be determined further with the help of free probability theory. Since Z was defined by the non-zero eigenvalues of the limit distribution of $P_\rho B_\rho P_\rho$ its moments are given by Equation (2.6) through

$$\mathbb{E}(Z^l) = \frac{1}{c}\phi\left[(BP)^l\right] = \sum_{\pi \in NC(l)} k_\pi[B,\ldots,B]c^{l-|\pi|}\,, \tag{3.41}$$

with $c = \phi[P] = r/n$. With the help of the examples after Definition 2.10 we exemplary determine the first three summands of (3.41):

$$\mathbb{E}(Z) = \phi[B],$$

$$\mathbb{E}(Z^2) = c\phi[B^2] - \phi[B]^2,$$

$$\mathbb{E}(Z^3) = c^2\phi(B^3] - 3c\phi[B]\phi[B^2] + 2\phi[B]^3,$$

and so forth.

3.2.3 The Case $k/n \to 0$

We shortly want to discuss the case $k/\rho \to 0$ here. Then also $r/n \to 0$. Let for a moment $m = n = \rho$ and $k = (\log \rho + 1)$. Can we still expect a significant violation of the trace multiplicativity in this situation? Since r is given by $\log \rho$ and

$$\tau(P_\rho) = \frac{\log \rho}{\rho}$$

tends to zero if $\rho \to \infty$, the limit distribution of the eigenvalues of P is just zero. This also renders all moments of Z and $1/Z$ to be zero. As a consequence the asymmetry we derived between the forward and the backward direction collapses since all relevant quantities tend to zero.

3.3 The Averaged Trace of the Covariance Estimator

Up to now we often used the terms $\tau_n(\hat{C}_{XX})$ and $\tau_n(\hat{C}_{YY})$ but it is not clear how close these quantities are to their exact values. Without loss of generality we only speak about $\hat{C}_{XX}$ here. One can imagine that if there are no dependencies among the dimensions, i.e. $x_i \perp\!\!\!\perp x_j$ for $i \neq j$ (the covariance matrix is diagonal), the normalized trace of the sample covariance matrix should converge quite rapidly for k and n going to infinity with equal rate. The following lemma, mainly based on observations made by Hoeffding [1963] for independent variables shows that even if there are dependencies among the different dimensions of X, the normalized trace of their covariance matrix converges independently from the number of dimensions, i.e., the only relevant factor in this convergence is the sample size $k = r + 1$.

Lemma 3.9 *(Sum of Dependent Variables)*
Let $X_1, \ldots, X_n$ be (dependent) random variables with $\mu_i = \mathbb{E}(X_i)$. Assume that $a \leq X_i \leq b$ for every $i = 1, \ldots, n$. Then

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{(b-a)^2}\right). \tag{3.42}$$

Proof: Note that this lemma is a slightly easier version of Theorem 2.1 in Janson [2004]. We follow the proof given there which is based on Hoeffding [1963].
Let $B = \sum_{i=1}^{n} X_i$. Subtracting the mean, we assume $\mathbb{E}(X_i) = 0$ for all $1 \leq i \leq n$. And thus $\mathbb{E}(B) = 0$. Then by Hoeffding [1963] (4.16), for every real h,

$$\mathbb{E}\exp(hA_i) \leq \exp\left(\frac{1}{8}h^2(b-a)^2\right).$$

By Jensen's inequality, for real u,

$$\exp(uB) = \exp\left(\sum_{i=1}^{n}\frac{1}{n}nuX_j\right) \leq \sum_{i=1}^{n}\frac{1}{n}\exp\left(nuX_i\right).$$

Taking the expectation leads to

$$\mathbb{E}\exp(uB) \leq \sum_{i=1}^{n}\frac{1}{n}\mathbb{E}\exp\left(nuX_i\right) \leq \sum_{i=1}^{n}\frac{1}{n}\exp\left(\frac{n^2u^2}{8}(b-a)^2\right).$$

We set $T := n(b-a)$, and find

$$\mathbb{E}\exp(uB) \leq \exp\left(\frac{1}{8}T^2u^2\right), u \in \mathbb{R}.$$

Hence, for $u \geq 0$, using Markov's inequality,

$$\Pr[B \geq t] = P(e^{uB} \geq e^{ut}) \leq e^{-ut}\mathbb{E}e^{uB} \leq \exp\left(\frac{1}{8}T^2u^2 - ut\right),$$

and the optimal choice $u := 4t/T^2$ yields

$$\Pr[B \geq t] \leq \exp(-2t^2/T^2).$$

Thus we have

$$\Pr[B \geq nt] \leq \exp\left(\frac{-2(nt)^2}{(n(b-a))^2}\right) = \exp\left(\frac{-2t^2}{(b-a)^2}\right)$$

By considering $-B$ we get the same result for $\Pr[B - \mathbb{E}B \leq -t]$ and hence the statement follows. □

Remark 3.10

We can view the sample variances $\hat{\sigma}_i$ on the main diagonal of $\hat{C}_{XX}$ as random variables X_i with expectation σ_i. Since we assumed that X has finite second moment these variances are finite, too. With growing sample size k the interval $[a, b]$ gets smaller and smaller, independent from n. Thus, the larger m and n (and therefore also k), the closer are $\tau_n(\hat{C}_{XX})$ and $\tau_m(\hat{C}_{YY})$ to its theoretical values $\tau_n(C_{XX})$ and $\tau_m(C_{YY})$, respectively. Note that if the dimensions are independent from each other, a larger dimension would speed up the convergence rate. This can be derived from a classical result made by Hoeffding [1963].

3.4 The Noisy Case

The noisy case is much more involved and we haven't found a completely satisfying solution yet. Still we want to say a few words about what the problems are and how one possibly can tackle them. Recall the model

$$Y = AX + E$$

with $A \in \mathbb{R}^{m\times n}$ and $E \perp\!\!\!\perp X$. We assume $\tau_m(C_{EE}) \nrightarrow 0$ as m tends to infinity. We further assume that the norm of $A_\rho^T A_\rho$ is growing slower than n, i.e., $\|A_\rho^T A_\rho\|/n \to 0$ as ρ tends to infinity. This is fulfilled if all moments of $A_\rho^T A_\rho$ converge. We first show that in the deterministic case we had consistent estimators for $\tau_n(C_{XX})$, $\tau_n(A^T A C_{XX})$ and $\tau_n(A^T A)$. Under consistency we want to understand here that if we have an estimator T_ρ for some parameter θ_ρ (also dependent on ρ) we claim that for every $\epsilon > 0$

$$\lim_{\rho\to\infty} \Pr\left[|T_\rho - \theta_\rho| \geq \epsilon\right] = 0\,.$$

Let us recall our estimation of A,

$$\hat{A} = \hat{C}_{YX}\hat{C}_{XX}^{+} = A\hat{C}_{XX}\hat{C}_{XX}^{+}\,. \tag{3.43}$$

This definition lead to the estimator

$$T_\rho^A = \frac{r}{n}\tau_n(\hat{A_\rho}^T\hat{A}_\rho) \quad \text{for} \quad \tau_n(A_\rho^T A_\rho)$$

and

$$T_\rho^{AC} = \tau_n(\hat{A}_\rho^T\hat{A}_\rho\hat{C}_{XX}^{(n)}) \quad \text{for} \quad \tau_n(A_\rho^T A_\rho C_{XX}^{(n)})$$

in the case that $r < \rho$. The consistency of $\tau_n(\hat{C}_{XX})$ for $\tau_n(C_{XX})$ is shown by Lemma 3.9. Due to the equalities

$$\tau_n(\hat{A}^T\hat{A}\hat{C}_{XX}) = \tau_n(\hat{C}_{XX}^{+}\hat{C}_{XX}A^T A\hat{C}_{XX}\hat{C}_{XX}^{+}\hat{C}_{XX}) = \tau_n(A^T A\hat{C}_{XX}) = \tau_n(\hat{C}_{YY})$$

and

$$\tau_n(A^T A C_{XX}) = \tau_n(AC_{XX}A^T) = \tau_n(C_{YY})$$

and again with Lemma 3.9 we conclude that also T_ρ^{AC} is a consistent estimator. It remains to show the consistency of T_ρ^A. This is done with the help of Lévy's Lemma (Lemma 2.14) and applying the first theorem from Janzing et al. [2010] which states:

Theorem 3.11 *(Traces are typically multiplicative)*
Let C be a symmetric, positive definite $n\times n$-matrix and A an arbitrary $m\times n$-matrix (with entries in $\mathbb{R}$). Let U be randomly chosen from $O(n)$ according to the unique $O(n)$-invariant distribution (i.e. the Haar measure). Introducing the operator norm

$$\|B\| := \max_{\|x\|=1}\|Bx\|,$$

we have

$$|\tau_n(A^T AUCU^T) - \tau_n(C)\tau_n(A^T A)| \leq 2\epsilon\|C\|\|A^T A\|$$

with probability at least $q := 1-\exp(-\kappa(n-1)\epsilon^2)$ for some constant κ (independent of C, A, n, m, ϵ).

A proof can be found in Janzing et al. [2010].
As done before we can write $\tau_n(\hat{A}_\rho^T \hat{A}_\rho) = \tau_n(\hat{C}_{XX}^{(n)+} \hat{C}_{XX}^{(n)} A_\rho^T A_\rho \hat{C}_{XX}^{(n)} \hat{C}_{XX}^{(n)+}) = \tau_n(P_\rho A_\rho^T A_\rho P_\rho)$ with $P_\rho = \hat{C}_{XX}^{(n)} \hat{C}_{XX}^{(n)+}$. We can prove

Corollary 3.12

With the above model and $r+1$ given samples under the condition $r < n$, it holds

$$\frac{n}{r}\tau_n(\hat{A}_\rho^T \hat{A}_\rho) \quad \textit{is a consistent estimator for} \quad \tau_n(A_\rho^T A_\rho)\,.$$

Proof: For an arbitrary $\epsilon > 0$ we have with Theorem 3.11

$$\left|\tau_n(\hat{A}_\rho^T \hat{A}_\rho) - \tau_n(A_\rho^T A_\rho)\tau_n(P_\rho)\right| \leq 2\epsilon \tag{3.44}$$

with probability at least $1 - \exp(-\kappa(n-1)\epsilon^2/\|A_\rho^T A_\rho\|)$ since $\|P_\rho\| = 1$. Due to the assumption $\|A_\rho^T A_\rho\|/n \to 0$ and $\tau_n(P_\rho) \to r/n$ it follows

$$\Pr\left[\left|\frac{n}{r}\tau_n(\hat{A_\rho}^T \hat{A}_\rho) - \tau_n(A_\rho^T A_\rho)\right| > \epsilon\right] \to 0$$

as ρ tends to infinity. Thus we have shown consistency. □

Remark 3.13

One could also argue with free probability theory: since $B_\rho = A_\rho^T A_\rho$ and P_ρ are independent there limit elements B and P are free and thus $\phi[BP] = \phi[B]\phi[P]$.

In the noisy case though, estimating A by (3.43) does not lead to a consistent estimator of $\tau_n(A^T A C_{XX})$ since $\tau_n(\hat{A}^T \hat{A} \hat{C}_{XX})$ contains some error terms that do not vanish if not $k \gg n$. More precisely

$$\begin{aligned}\tau_n(\hat{A}^T \hat{A} \hat{C}_{XX}) &= \tau_n(\hat{C}_{XX}^+ \hat{C}_{XY} \hat{C}_{XY} \hat{C}_{XX}^+ \hat{C}_{XX}) \\ &= \tau_n(A^T A \hat{C}_{XX}) + 2\tau_n(A \hat{C}_{XX} \hat{C}_{XX}^+ \hat{C}_{XE}) + \tau_n(\hat{C}_{EX} \hat{C}_{XX}^+ \hat{C}_{XE})\,.\end{aligned}$$

Under the condition that $r < n$, the last term

$$\tau_n(\hat{C}_{EX} \hat{C}_{XX}^+ \hat{C}_{XE}) = \tau_n(\hat{C}_{EE})$$

by a slight variation of Lemma 3.2 and therefore does not vanish for $n \to \infty$. We will encounter a similar term in $\tau_n(\hat{A}^T \hat{A})$. This brings us to the challenge of either finding consistent estimators for these two quantities or reducing dimensionality, for example with the help of projections.

3.4.1 Dimensionality Reduction

In the following we will sketch an idea which aims to reduce dimensionality. This idea came up in a discussion with Kun Zhang[2].

[2]Kun Zhang. *Private communication*, 2010

Assume $m = n$. First we choose an n-dimensional unit vector θ at random according to the Haar measure on S_n. We multiply our model $Y = AX + E$ from the left with θ^T and get

$$\theta^T Y = \theta^T AX + \theta^T E,$$

which we rewrite as

$$\check{Y} = \check{A}X + \check{E},$$

with $\check{Y}$ and $\check{E}$ now being one-dimensional and $\tilde{A}$ is a $1 \times n$-vector. With Lévy's Lemma one can show that then

$$\check{A}\check{A}^T \approx \tau_n(A^T A)\,.$$

Thus we are thrown back to an ordinary regression problem with n-dimensional X and one-dimensional $\check{Y}$. But since we still have the limitation of $k < n$ with k denoting the number of samples, we cannot apply traditional tools designed for solving this sort of regression problem. On the other hand this is a common problem in modern statistics and machine learning since growing computing power and other technology made it possible to collect data of unprecedented size and complexity. One has to find the relevant dimensions in X, those who carry the majority of the information and ignore the others. This problem is called "variable selection" in the statistics literature. For a recent review on this subject, see Fan and Lv [2010].
To solve this task one has to find additional assumptions that usually hold in practice and that may help to estimate the transformation matrix $\check{A}$. Although X and Y are high-dimensional, often only a subset of all X_i's is significant for $\check{Y}$. In other words, the large matrix A only has a small number of non-zero entries, making it possible to find a consistent estimator even if $k < n$. Therefore, we assume $\check{A}$ to be sparse and would like to find its estimate by making use of this assumption.
Fan and Lv [2008] propose a method to deal with this kind of problem. To find out which X_i are significant for $\check{Y}$ they use sure independence screening (SIS), a method explained more detailed in Fan et al. [2009]. With SIS and the right sparsity assumption it is possible to estimate $\check{A}$ reliably and then test the quantities on the trace multiplicativity. For the backward direction just exchange the variables and repeat the whole procedure.

4 Inference Algorithm and Experiments

In this chapter we want to use our insights obtained from Chapter 3 to develop an algorithm for deciding whether

$$X \to Y \quad \text{or} \quad Y \to X$$

for a deterministic relationship between X and Y. The asymmetry we developed is based on the trace multiplicativity of $A^T A C_{XX}$ in the forward direction and the fact that this is violated in the backward direction. We therefore introduce a scale-invariant measure for the strength of this violation similar to that of Janzing et al. [2010].

Definition 4.1 (A scale-invariant measure)
Given the estimators $\hat{C}_{XX}$, $\hat{A} = \hat{C}_{YX}\hat{C}_{XX}^{+}$ and $r = \texttt{rank}(\hat{A}^T\hat{A})$ we define

$$\Delta(C_{XX}, A) := \log \tau_n(\hat{A}^T\hat{A}\hat{C}_{XX}) - \log \tau_n(\hat{C}_{XX}) - \log\left(\frac{n}{r}\tau_n(\hat{A}^T\hat{A})\right) .$$

Then $\Delta(C_{YY}, \tilde{A})$ is given by

$$\Delta(C_{YY}, \tilde{A}) := \log \tau_m(\hat{\tilde{A}}^T\hat{\tilde{A}}\hat{C}_{YY}) - \log \tau_m(\hat{C}_{YY}) - \log\left(\frac{m}{r}\tau_m(\hat{\tilde{A}}^T\hat{\tilde{A}})\right) .$$

We can state a "theorem" that should not be taken too strict (cf. Janzing et al. [2010]):

Theorem 4.2
If the dimensionality of X and Y is sufficiently high and the sample size is a significant fraction of the dimensionality it holds

$$\Delta(C_{XX}, A) + \Delta(C_{YY}, \tilde{A}) < 0. \tag{4.1}$$

The prove is immediate from Theorem 3.1 and Theorem 3.7.
The idea of the inference method now works indirectly: We assume that for the right direction the effect Y is given by a linear transformation of X, i.e., $Y = AX$ and that $A^T A$ and C_{XX} are independent. We calculate both $\Delta(C_{XX}, A)$ and $\Delta(C_{YY}, \tilde{A})$. These numbers indicate how strong the trace multiplicativity is violated. Since it should hold $\Delta(C_{XX}, A) \approx 0$ for the forward direction we propose the following inference rule:

Inference method: *Given $\Delta(C_{XX}, A)$ and $\Delta(C_{YY}, \tilde{A})$, infer that $X \to Y$ if $\Delta(C_{XX}, A)$ is closer to zero and $Y \to X$ if $\Delta(C_{YY}, \tilde{A})$ is closer to zero (see Algorithm 1).*

With the help of experiments we would like to clarify: Does the result obtained for dimension to infinity already hold for moderate dimensions? Is the multiplicativity of trace sufficiently violated with a rather small sample size? How large must we choose ϵ in Algorithm 1 to obtain reliable results? In real data sets, is the causal structure sufficiently close to our model with independent choices of $A^T A$ and C_{XX}?

Algorithm 1 Identifying linear causal relations via traces

Input: $(x_1, y_1), \ldots, (x_k, y_k)$, $(k \leq m, n$; with $n = \dim(x_i)$, $m = \dim(y_i))$
Compute $\hat{C}_{XX}$ and $\hat{A} = \hat{C}_{YX}\hat{C}^{+}_{XX}$
Compute $\hat{C}_{YY}$ and $\hat{\tilde{A}} = \hat{C}_{XY}\hat{C}^{+}_{YY}$
Compute $r_1 = \texttt{rank}(\hat{A})$ and $r_2 = \texttt{rank}(\hat{\tilde{A}})$
if $\left|\log \tau_n(\hat{A}^T\hat{A}\hat{C}_{XX}) - \log \tau_n(\hat{C}_{XX}) - \log\left(\frac{n}{r_1}\tau_n(\hat{A}^T\hat{A})\right)\right| > \epsilon +$ $\left|\log \tau_m(\hat{\tilde{A}}^T\hat{\tilde{A}}\hat{C}_{YY}) - \log \tau_m(\hat{C}_{YY}) - \log\left(\frac{m}{r_2}\tau_m(\hat{\tilde{A}}^T\hat{\tilde{A}})\right)\right|$ **then**
  write "Y is the cause"
else
  if $\left|\log \tau_m(\hat{\tilde{A}}^T\hat{\tilde{A}}\hat{C}_{YY}) - \log \tau_m(\hat{C}_{YY}) - \log\left(\frac{m}{r_2}\tau_m(\hat{\tilde{A}}^T\hat{\tilde{A}})\right)\right| > \epsilon +$ $\left|\log \tau_n(\hat{A}^T\hat{A}\hat{C}_{XX}) - \log \tau_n(\hat{C}_{XX}) - \log\left(\frac{n}{r_1}\tau_n(\hat{A}^T\hat{A})\right)\right|$ **then**
    write "X is the cause"
  else
    write "cause cannot be identified"
  end if
end if

4.1 Experiments with Simulated Data

We present experiments with simulated data. First we will explain how the data were created. Focusing on the deterministic case, we want to generate random models

$$Y = AX.$$

To this end we constructed the $m \times n$ matrix A as follows: We set $l := min(m, n)$ and chose a random l-dimensional diagonal matrix D^A with i.i.d. entries from some distribution with finite moments. We filled it up with zeros and obtained an $m \times n$ matrix. Then we multiplied it on one side with a random m-dimensional orthogonal matrix, on the other side with an n-dimensional one, i.e.,

$$A = U^A_m D^A U^A_n \qquad \left(U^A_m \in O(m), U^A_n \in O(n)\right).$$

This ensures that the eigenvalues of A^TA coincide with the square of the eigenvalues of D^A and therefore all moments of A^TA converge.
Next we generated a random covariance matrix C_{XX} in the same fashion: A diagonal matrix D^C with i.i.d. entries taken at random from a distribution which only takes positive values was multiplied with a random orthogonal matrix from both sides:

$$C_{XX} = U^C_n D^C U^C_n \qquad \left(U^C_n \in O(n)\right).$$

Let now B_X be the unique square root of C_{XX}, i.e., $B_X = \sqrt{C_{XX}}$. We sampled an $n \times k$-matrix $\tilde{X}$ by drawing each entry i.i.d. from some distribution with finite moments and multiplied it from the left with B_X, i.e.

$$X = B_X\tilde{X}.$$

Then X is a data matrix whose columns consist of k different samples x_i chosen independent at random from a certain distribution with covariance matrix C_{XX}. Finally we set $Y = AX$.
The random orthogonal matrices are created following the algorithm of Diaconis and Shahshahani [1987], Method B, explained in the next section. This is the fastest known algorithm that produces random orthogonal matrices ensuring rotation invariance.

4.2 Uniform Distributed Random Orthogonal Matrices

In this section we will explain an algorithm which outputs uniform distributed random orthogonal matrices of dimension n. We follow a paper from Diaconis and Shahshahani [1987]. We first give a short theory part where we introduce the idea of the algorithm. It needs some elements from group theory.

4.2.1 The Subgroup Algorithm

Let G be a finite group. Let $G_0 = G \supset G_1 \supset \cdots \supset G_r$ be a nested chain of subgroups (not necessarily normal). Let C_i be coset representatives for G_{i+1} in G_i, $0 \leq i < r$. A coset representative is a representative in a equivalent class sense. What this will be will become clearer in the next section. G can be represented as

$$G \cong C_0 \times C_1 \times \cdots \times C_{r-1} \times C_r \tag{4.2}$$

in the sense that each $g \in G$ has a unique representation $g = g_0 g_1 \cdots g_r$ with $g_0 \in C_i$ and $g_r \in G_r$. It follows that if the g_i are chosen uniformly at random in their respective domains and multiplied together, the resulting product element g will be uniformly distributed on G.
The subgroup algorithm works, in essentially the same way, for any compact topological group G. For an illustration of this take a look at Diaconis and Shahshahani [1987], section 4. The algorithm was first developed by G. W. Stewart. He gives a clear discussion in Stewart [1980]. The idea is to find a closed subgroup $H \subset G$, chose an element of H at random, chose a coset representative at random, and multiply.

4.2.2 Generation of Random Orthogonal Transformations

We now want to apply the above theory to a concrete example, namely the orthogonal group. Let $O(n)$ be the group of $n \times n$ orthogonal matrices. $O(n)$ has a natural uniform distribution called Haar measure. In probabilistic notation, the random matrix X is uniform distributed if

$$\Pr\{X \in A\} = \Pr\{X \in \Gamma A\}$$

for every $A \subset O(n)$ and $\Gamma \in O(n)$.
In two dimensions, a random X can be specified as

$$\begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -b\sin(\theta) & b\cos(\theta) \end{pmatrix},$$

with θ uniform on $[0, 2\pi]$ and $b = \pm 1$ with probability $\frac{1}{2}$.
The algorithm is based on the tower

$$O(n) \supset O(n-1) \supset O(n-2) \supset \cdots \supset O(2),$$

with $O(n-1)$ the subgroup of $O(n)$ fixing the vector e_1. If we knew how to choose a random element of $O(n-1)$ and coset representatives for $O(n-1)$ in $O(n)$ at random, then the subgroup algorithm and induction finish the job.
Coset representatives for $O(n-1)$ in $O(n)$ can be specified by saying where e_1 goes. Thus the coset space $O(n)/O(n-1)$ can be identified with S^{n-1} – the $(n-1)$-dimensional sphere in $\mathbb{R}^n$. For $x \in S^{n-1}$ (as a column vector) define the Householder reflection

$$I - 2xx^T.$$

For every $v \in S^{n-1}$ this reflection with x chosen as

$$x = (e_1 - v)/c \quad \text{and} \quad c = \sqrt{(e_1 - v)^T(e_1 - v)} \tag{4.3}$$

takes e_1 into v. Choosing v at random results in a randomly chosen coset representative. This results in the following:

Lemma 4.3 *(Uniform distribution on $O(n)$)*
Let v be chosen at random on the $(n-1)$-sphere. Let Γ_1 be chosen at random on $O(n-1)$. Then, with x defined as in (4.3)

$$(I - 2xx^T) \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \Gamma_1 & \\ 0 & & & \end{pmatrix} \tag{4.4}$$

is uniformly distributed on $O(n)$.

This follows from the observations made around (4.2). The standard way of choosing v at random on S^{n-1} is to take $v = (z_1, \ldots, z_n)/\sqrt{(z_1^2, \ldots, z_n^2)}$ with z_i independent standard normal. From here, induction gives a simple algorithm for choosing a uniformly distributed element of $O(n)$. It is an $O(n^3)$ algorithm, but with a smaller constant than other classical algorithms with the same complexity, making a substantial difference in computing time.

4.3 Results of Experiments with Simulated Data

We present results of experiments with simulated data for the deterministic model. We sampled the singular values of A, i.e., the diagonal entries of D^A independently from a standard normal distribution. The eigenvalues of C_{XX}, i.e., the entries of the diagonal matrix D^C were drawn independently from absolute values of the standard normal distribution.

In all experiments shown in the following part we used $\rho = m = n$. Nevertheless, we also did experiments with $m \neq n$ and got similar results. Figure 4.1 shows both the performance and the values of delta for both directions. The sample size is half of the dimensionality, i.e., $k = n/2$. As can be seen, already at around 20 dimensions, the performance of the method exceeds 90%. Figure 4.2 shows the performance and the values of delta as a function of sample size while the dimensionality is fixed to fifty. A samples size of around 20 seems to be enough to reliably distinguish between cause and effect in that case. In Algorithm 1 we need to specify an ϵ as an indicator when the difference of the values of the two deltas is big enough to decide in favor for one direction. For the two Figures 4.1 and 4.2 we chose $\epsilon = 0.3$.
To show the fact that it is crucial that the quotient k/n does not tend to zero, Figure 4.4 shows similar plots as Figure 4.1, except that now the sample size equals two times the logarithm of the number of dimensions and hence k/ρ tends to zero as ρ tends to infinity. As one can see in the right plot of Figure 4.4, $\Delta_{Y\to X}$ tends to zero with growing dimensionality and while doing so the performance drops. In this setting we used $\epsilon = 0.2$.

We also tested the method with small noise, still using Algorithm 1. C_{EE} was generated in the same way as C_{XX}, although with an adjustable parameter σ governing the scaling of the noise with respect to the signal: $\sigma = 0$ yields the deterministic setting, $\sigma = 1$ equates the power of the noise to that of the signal. Figure 4.3 shows the results obtained with $\sigma = 0.3$. As already remarked in the last chapter, if $X \to Y$, $\Delta_{X\to Y}$ does not converge to zero as dimensionality rises. However, it is still much closer to zero than $\Delta_{Y\to X}$. Of course, this could be due to the manner of sampling. We will further discuss this point in Section 4.4.1.

4.4 Experiments with Real World Data

We will present experiments with real world data from the field of climate research. We took several variables from Reanalysis data.[3] Reanalysis data is a technique to produce multiple climate variables. Previously observed climate data for temperature, wind speed, and pressure is recorded, observations are analyzed and interpolated onto a system of grids. Then a 3-D forecasting model is initialized with this observational data. The output is a simulated data set at 6-hourly, daily, and monthly time steps of many unobservable climate variables. For our purposes we used monthly mean data on a T62 Gaussian grid (192×94 grid points, latitude $\times$ longitude, see Figure 4.6).
Out of the big pool of possible variables, we chose five, where we are relatively sure about the ground truth. We took *precipitation rate*, *volumetric soil moisture*, *specific humidity at two meters*, *upward longwave radiation flux* and *near infrared beam downward solar flux*. Data is given in monthly means from 1/1948 until 6/2010. This gives us in total a dimensionality of 18048 ($= 192 \times 94$ points) and a sample size of 750 ($= 12 \times 62.5$ months).

[3] We thank the NCEP/NCAR 40-year reanalysis project (Kalnay et al. [1996]). NCEP Reanalysis Derived data is provided by the National Oceanic and Atmospheric Administration (NOAA/OAR/ESRL/PSD), Boulder, Colorado, USA. We took the data from their Website at http://www.esrl.noaa.gov/psd/.

We first examine the two pairs

$$X = \textit{precipitation rate (prate)}, \quad Y_1 = \textit{volumetric soil moisture (soilm)} \tag{4.5}$$

and

$$X = \textit{precipitation rate (prate)}, \quad Y_2 = \textit{specific humidity at two meters (sphum)}.$$

More precisely, X and Y_1 are multidimensional random variables

$$X = \begin{pmatrix} \textit{prate at grid point } x_1 \\ \textit{prate at grid point } x_2 \\ \vdots \\ \textit{prate at grid point } x_n \end{pmatrix}, \quad Y_1 = \begin{pmatrix} \textit{soilm at grid point } x_1 \\ \textit{soilm at grid point } x_2 \\ \vdots \\ \textit{soilm at grid point } x_n \end{pmatrix}. \tag{4.6}$$

Different samples of this random variables are now given by the values for the different months. For the whole chapter it will hold that the definition of random variables as in (4.5) is meant in the spirit of (4.6).
The meaning of precipitation rate is clear. Volumetric soil moisture is a measure of how much water is contained in the soil and is given in fraction of a volume unit. Specific humidity is the ratio of water vapor to air. Since we believe that precipitation strongly effects soil moisture and humidity whereas the effect in the other direction is nearly negligible we get the ground truths $X \to Y_1$ and $X \to Y_2$, respectively (of course one can argue that in regions with high soil moisture or humidity the formation of clouds rises which can result in rain but the effect in the other direction should be much stronger. Additionally, if we take points lying far away from each other we this backward relation should get even more random).
For the variable Y_2 we have two options, namely the level from $0 - 10cm$ or the level from $10 - 200cm$. Since the results did not differ much we will report here only the results of the former. Volumetric soil moisture can only be measured over land, such that the 18048 grid points boil down to 5914 points ($n = 5914$), which are exactly the grid points lying over land. Now we take the different locations as different dimensions and the points in time as samples. It is clear the there are a lot of inner-structural dependencies. For this reason we do not take all points into account but choose a widely spread fraction out of them.

Table 4.1 to Table 4.4 show the values for $\Delta_{X \to Y_1}$, $\Delta_{Y_1 \to X}$, $\Delta_{X \to Y_2}$ and $\Delta_{Y_2 \to X}$, respectively, for different subsets out of the whole dataset. The subsets are the following:

$$[1 : i : 5914, 1 : j : 750]$$

where i and j are the step size for the grid and time points, respectively. i takes values from 10 to 100 with step size 10, j takes values from 2 to 12 with step size 2. Taking for example every sixth month and every 10th grid point we obtain dimensionality 591 and sample size 125.

The calculated delta values were clipped to two digits behind the point. As one can see, except in one position ($i = 100$, $j = 10$), every value of Table 4.1 is closer to zero than its corresponding value for the backward direction in Table 4.2. The results for the second pair of variables are similar: Only two values in the first column of Table 4.4 are closer to zero than the corresponding values in Table 4.3. Choosing $\epsilon = 0.3$ would give us a wrong direction only in one case.

There is one assumption concerning the transfer matrix A contained in the method. That is, that the entries of A are independent and identically distributed. Due to the structure of our real world data sets, one can object that this assumption is clearly violated here¿ We will discuss this in the next subsection. Nonetheless, the method gives good results. Moreover, in most of the inference methods in causality, sampled data are assumed to be independent and identically distributed, too. Taking consecutive months in our examples with high probability would violate this assumption since weather is dependent on time. Surprisingly, the method seems to be fairly robust against these two assumptions since even by taking every second month into account it outputs the right direction in nearly all cases. This is true for the next pair of variables, too. Note that for some combinations of i and j the sample size exceeds the number of dimensions.

We present the results of another pair of variables

$$X = \textit{upward longwave radiation flux (uwlrf)}$$

and

$$Y = \textit{near infrared beam downward solar flux (nbdsf)}$$

Radiation fluxes are given in terms of the quantity of radiant energy flowing through unit area of a surface in unit time. Near infrared beam downward solar flux measures part of the energy coming from the sun. This energy, measured in radiation, is reflected at the surface of the earth and one part of the reflected energy happens to be upward longwave radiation flux. Thus the ground truth here is $Y \to X$. Reflected radiation from the surface can be reflected again at the clouds and come back to the earth but also here, this relation is much weaker the that in the other direction. Additionally, since we take widely spread points in space, it is unlikely that we meet exactly those points, where the reflected radiation comes back to earth again.
The corresponding tables for $\Delta_{X\to Y}$ and $\Delta_{Y\to X}$ are Table 4.5 and Table 4.6. Observe that every value of the Table 4.6 for $\Delta_{Y\to X}$, standing for the right direction is closer to zero than its corresponding value in Table 4.5, standing for the wrong direction.
In this second real data set we can use the information around the whole globe, i.e. the dimensionality gets much (approximately three times) larger than in the first case. Still we observe at no point that the method outputs the wrong direction.

4.4.1 Discussion of Results

The proposed method for the small sample case and Algorithm 1 are designed for the deterministic case. However, one can expect real world data being never free from noise.

Still the algorithm outputs the right direction in nearly all cases. What might be a plausible explanation? As we already discussed before it could be possible that $\Delta_{X\to Y}$ (X being the cause) still converges to a value closer to zero than $\Delta_{Y\to X}$ in the noisy case. Another possibility is the following, namely that the difference of both deltas always have the same sign, i.e., it usually holds

$$\Delta_{X\to Y} - \Delta_{Y\to X} > 0\,.$$

We made this observation both with simulated data and with experiments on real world data. Nonetheless, Janzing et al. [2010] have shown that this cannot hold in the generality of all cases. The authors constructed a counter example by assuming A to be orthogonal. The value of delta for the backward direction then gets positive. It still might be that this case rarely appears in nature and therefore is a rather academic exception. This cannot be clarified satisfyingly in this work and must be postponed to future work.

Let us discuss another point here: The real world data sets we investigated have a specific structure. For example rain is falling on a certain place, making the soil moist at this same place which should render A to be diagonal. We tested this by just taking very few values out of the grid and taking the whole time period into account. Doing this we should get a quite reliable estimate of this specific A. We found that the values on the diagonal are quite big compared to the other values. However, the are some matrices where this relation is not very strong. Figure 4.5 shows an image plot for two 20-dimensional example matrices which nicely show the expected diagonal structure. On the left side we used the dimensions $[1001 : 100 : 2901]$ whereas on the right side we used $[2050 : 100 : 3950]$. Both times we took all 750 sample to estimate the transfer matrices. Anyway, as it is shown in a small example in Janzing et al. [2010], the method works even if both A and C_{XX} are diagonal matrices, as long as the diagonal elements have significant variance.

Moreover, we made an additional observation. That is, the estimated A for the above data sets seems to have a specific vertical structure. Values in the same column are likely to be close to each other. Assuming such a structure on A can lead to a simplified variable selection method, which is similar to the idea proposed in Section 3.4.1. We present a short, rather heuristic algorithm and its results on one of the above data sets in the following section.

We also tested quadratic regions of 400 grid points. In that setting the inner dependencies should be very high. Nevertheless, we obtained similarly good results as in the other cases. That suggests that the proposed method is possibly robust against these dependencies.

4.4.2 A Heuristic Approach towards the Noisy Case

Here we report some results we obtained from a simplified algorithm we derived out of the ideas of Section 3.4.1. It reduces only the dimensionality of the cause, i.e., if we want to calculate $\Delta_{X\to Y}$ we throw out some dimensions of X and hence get a rectangular A. This algorithm assumes that A has a certain structure, namely that some columns

are negligible. A parameter α has to be chosen. This parameter will give the final dimensionality of X and should be significantly smaller than k.
As $X \to Y$ we transform Y to an orthonormal basis. Then we measure the correlation between these basis vectors and the different dimensions of X. More precisely, if X and Y are the $n \times k$- and $m \times k$-observation matrices, respectively, let then

$$\tilde{Y} = TY$$

where T is a transformation matrix such that $C_{\tilde{Y}\tilde{Y}} = I$. Now calculate the correlations between the rows X_i and $\tilde{Y}$

$$c_i = \frac{1}{m} \sum_{j=1}^{m} X_i \tilde{Y}_j^T$$

and sort it. Starting from the dimension with the lowest correlation, delete rows X_i until only α dimensions are left. Now the new A can be estimated by

$$\hat{A} = \hat{C}_{YX} \hat{C}_{XX}^{-1}$$

since $\hat{C}_{XX}$ is invertible now. If α is significantly smaller than k, this new A can be estimated quite reliably. Additionally, $\hat{A}^T \hat{A}$ has full rank. Finally, one can test the trace multiplicativity by calculating $\Delta_{X \to Y}$. For the other direction just exchange X and Y.
For this method the parameter α has to be chosen in the beginning. We made good experiences with $\alpha = 2k/3$ and $\alpha = k/2$. Table 4.7 and Table 4.8 show the results for

$$X = \textit{precipitation rate} \quad \text{and} \quad Y = \textit{volumetric soil moisture}$$

with $\alpha = 2k/3$. Note that we only calculated the values of delta if the dimensionality exceeds the sample size. As one can see we already got promising results with this simple method since only on one position $(i = 12, j = 30)$ we would infer the wrong direction by taking an ϵ smaller than 0.2. We are also working on a more advanced method to attack the variable selection problem involving sure independence screening (cf. Section 3.4.1).

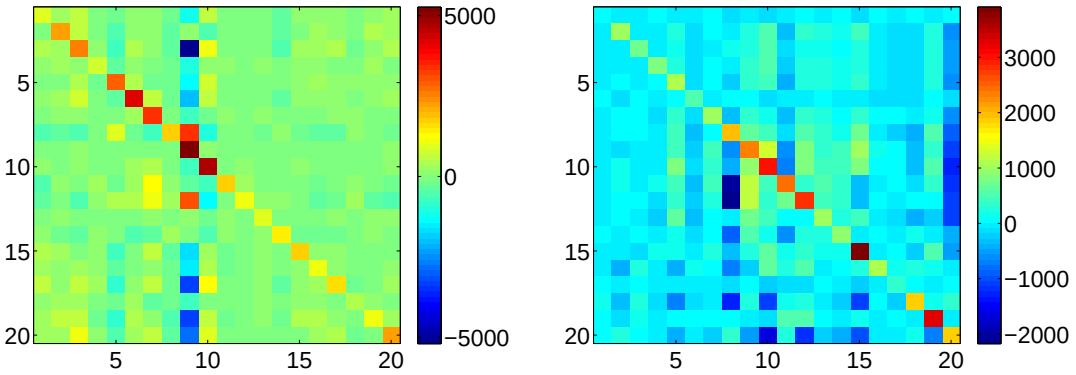


Figure 4.5: Visualization of the estimation of two representative transfer matrices for $X = \textit{prate}$ and $Y = \textit{soilm}$ and the model $Y = AX$. Out of the 5914 available dimensions (grid points) only 20 were taken into account, namely [1001:100:2901] and [2050:100:3950] for the left and the right plot, respectively. A was estimated with all 750 samples by $\hat{A} = \hat{C}_{YX} \hat{C}_{XX}^{-1}$.

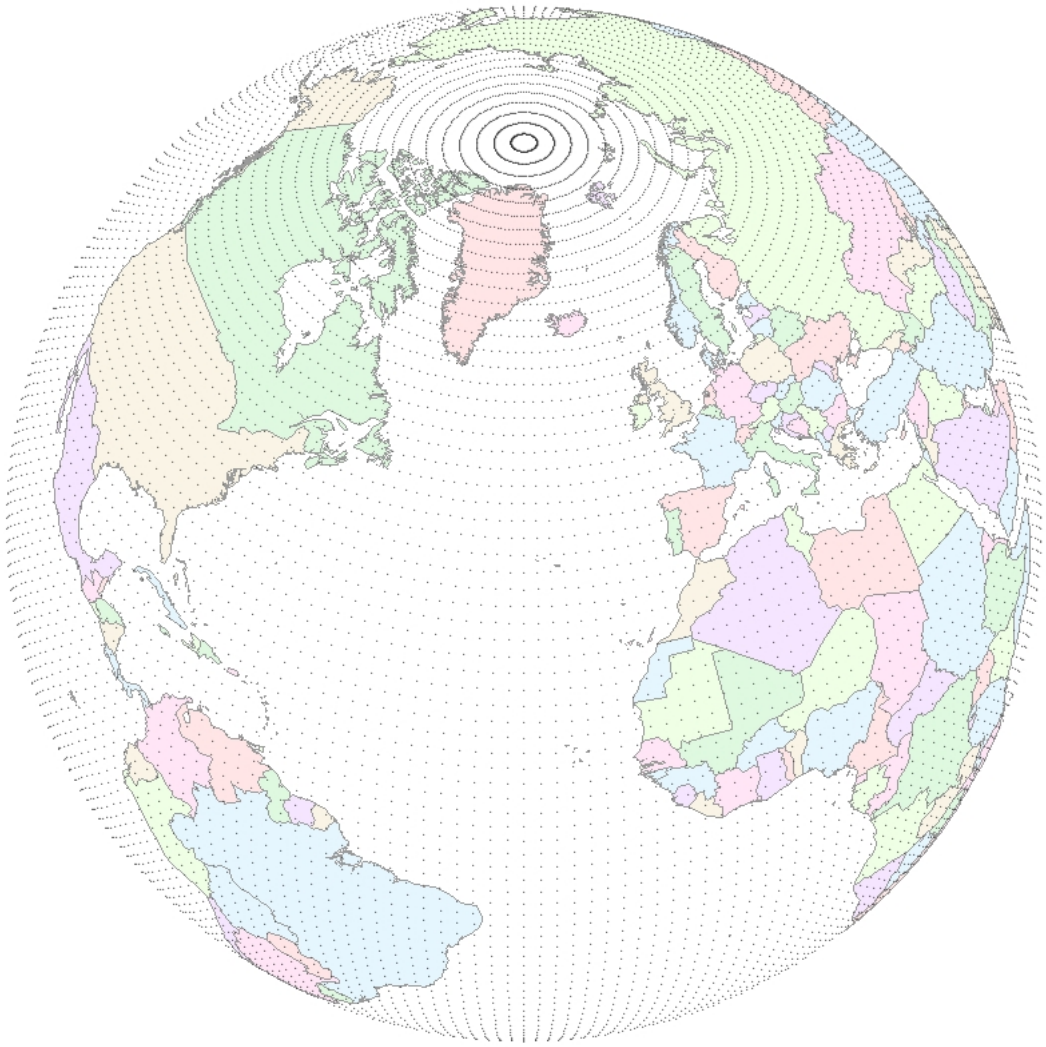


Figure 4.6: Visualization of a Gaussian grid[4]

[4]Figure taken from http://en.wikipedia.org/wiki/File:NCEP_T62_gaussian_grid.png

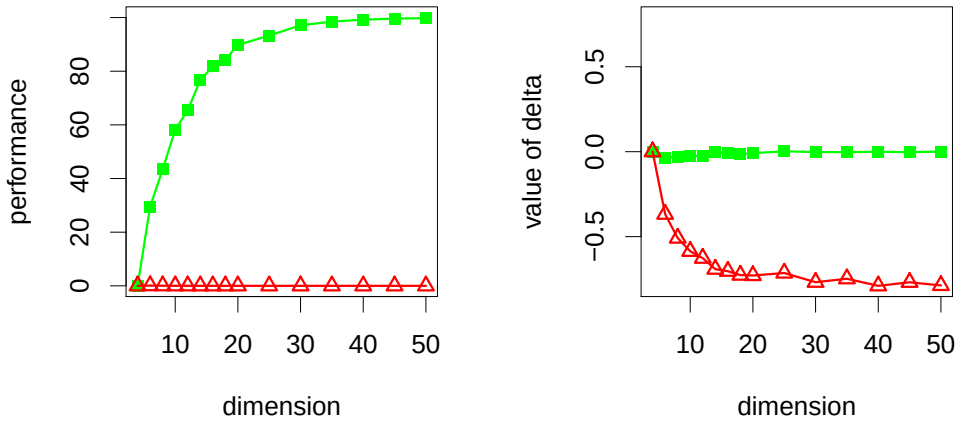


Figure 4.1: Deterministic setting. (Left) Performance of the method as a function of the input dimensionality n, when the output dimensionality $m = n$ and sample size is $k = n/2$. The green curve (rectangles) denotes the percentage of simulations in which the true causal direction was selected, while the red curve (triangles) gives the percentage of wrong answers. We used $\epsilon = 0.3$ and did 1000 simulations. (Right) Mean values of $\Delta_{X\to Y}$ (green curve, rectangles) and $\Delta_{Y\to X}$ (red curve, triangles).

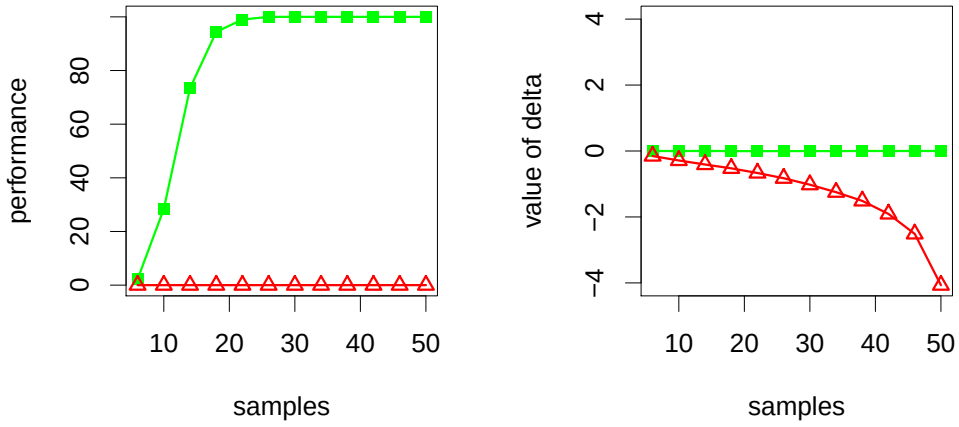


Figure 4.2: Deterministic setting. (Left) Performance of the method as a function of the sample size while fixing the input dimensionality $n = 50$, when the output dimensionality $m = n$. The green curve (rectangles) denotes the percentage of simulations in which the true causal direction was selected, while the red curve (triangles) gives the percentage of wrong answers. We used $\epsilon = 0.3$ and did 500 simulations. (Right) Mean values of $\Delta_{X\to Y}$ (green curve, rectangles) and $\Delta_{Y\to X}$ (red curve, triangles).

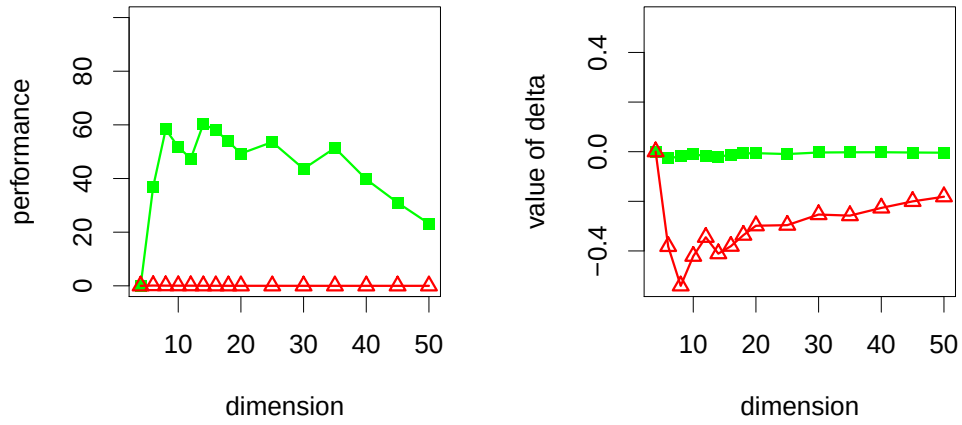


Figure 4.3: Deterministic setting. (Left) Performance of the method as a function of the input dimensionality n, when the output dimensionality $m = n$ and sample size is $k = \lceil 2 \log n \rceil$. The green curve (rectangles) denotes the percentage of simulations in which the true causal direction was selected, while the red curve (triangles) gives the percentage of wrong answers. We used $\epsilon = 0.3$ and did 1000 simulations. (Right) Mean values of $\Delta_{X \to Y}$ (green curve, rectangles) and $\Delta_{Y \to X}$ (red curve, triangles).

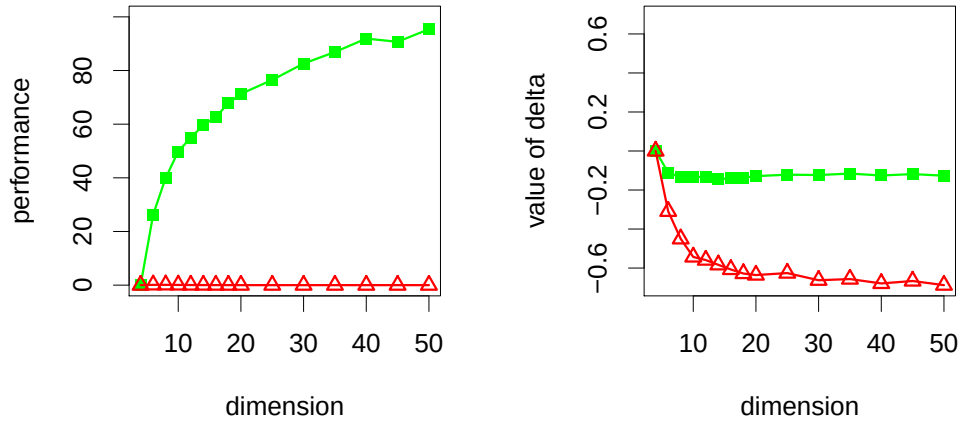


Figure 4.4: Non-deterministic setting, $\sigma = 0.3$. (Left) Performance of the method as a function of the input dimensionality n, when the output dimensionality $m = n$ and sample size is $k = n/2$. The green curve (rectangles) denotes the percentage of simulations in which the true causal direction was selected, while the red curve (triangles) gives the percentage of wrong answers. We used $\epsilon = 0.3$ and did 1000 simulations. (Right) Mean values of $\Delta_{X \to Y}$ (green curve, rectangles) and $\Delta_{Y \to X}$ (red curve, triangles).

j \ i	10	20	30	40	50	60	70	80	90	100
2	-0.69	-1.07	-1.36	-1.07	-1.23	-1.02	-2.12	-1.77	-2.14	-1.63
4	0.00	-1.08	-1.86	-2.08	-2.13	-1.57	-1.68	-2.09	-2.22	-2.09
6	0.18	-0.30	-1.22	-1.39	-1.71	-1.82	-1.89	-2.31	-2.35	-2.02
8	0.44	0.11	-0.45	-1.42	-2.18	-2.48	-2.57	-2.95	-2.69	-2.88
10	0.26	-0.01	-0.26	-1.08	-1.44	-2.32	-3.44	-3.19	-3.29	-3.77
12	-0.42	-0.74	-1.07	-1.69	-2.14	-2.47	-3.94	-4.57	-4.92	-5.73

Table 4.1: Values of $\Delta_{X \to Y_1}$ for different settings of i and j ($X = prate$, $Y_1 = soilm$).

j \ i	10	20	30	40	50	60	70	80	90	100
2	-1.54	-1.86	-1.85	-1.82	-2.06	-2.73	-2.24	-2.79	-3.70	-2.95
4	-1.88	-2.31	-2.61	-2.38	-2.42	-2.81	-2.74	-2.84	-3.46	-2.95
6	-1.39	-2.13	-2.07	-2.58	-2.51	-2.11	-2.64	-2.78	-3.27	-2.64
8	-1.46	-2.05	-2.49	-3.14	-3.19	-3.24	-3.57	-3.34	-4.00	-3.36
10	-1.22	-1.62	-1.98	-2.87	-3.07	-3.82	-3.65	-3.67	-4.31	-3.69
12	-0.61	-1.33	-1.97	-3.30	-5.23	-5.21	-5.80	-5.90	-6.22	-6.65

Table 4.2: Values of $\Delta_{Y_1 \to X}$ for different settings of i and j ($X = prate$, $Y_1 = soilm$).

j \ i	10	20	30	40	50	60	70	80	90	100
2	1.12	0.71	0.16	-0.26	-0.30	-0.09	0.06	-0.25	0.19	-0.01
4	1.28	1.15	0.86	0.62	0.20	-0.17	-0.64	-1.25	-1.37	-1.39
6	1.78	1.69	1.52	1.42	1.18	1.01	0.84	0.56	0.34	-0.07
8	1.35	1.30	1.17	1.09	0.92	0.85	0.67	0.52	0.38	0.17
10	1.32	1.28	1.19	1.12	1.00	0.92	0.82	0.70	0.62	0.44
12	0.02	-0.04	-0.07	-0.21	-0.29	-0.38	-0.42	-0.45	-0.61	-0.81

Table 4.3: Values of $\Delta_{X \to Y_2}$ for different settings of i and j ($X = prate$, $Y_2 = sphum$).

j \ i	10	20	30	40	50	60	70	80	90	100
2	-0.73	-1.34	-1.82	-1.80	-2.16	-2.11	-2.02	-1.98	-1.97	-2.23
4	-1.47	-2.05	-2.40	-2.42	-2.73	-2.73	-2.70	-2.74	-3.10	-3.08
6	-1.74	-2.32	-2.72	-2.73	-3.00	-3.01	-3.05	-3.01	-3.50	-3.51
8	-2.11	-2.24	-2.22	-2.41	-2.45	-2.52	-2.70	-2.81	-2.96	-3.31
10	-2.12	-2.16	-2.16	-2.30	-2.32	-2.42	-2.47	-2.60	-2.78	-2.95
12	-0.65	-0.67	-0.75	-0.80	-0.81	-0.92	-1.03	-1.17	-1.12	-1.26

Table 4.4: Values of $\Delta_{Y_2 \to X}$ for different settings of i and j ($X = prate$, $Y_2 = sphum$).

j \ i	10	20	30	40	50	60	70	80	90	100
2	1.35	0.64	-0.06	0.09	0.16	0.14	-0.24	-0.25	-0.19	-0.15
4	0.58	-0.15	-0.23	-0.76	-0.67	-0.75	-1.12	-1.15	-1.15	-1.06
6	0.54	-0.23	-1.01	-0.89	-0.80	-1.55	-1.33	-1.4243	-1.38	-1.36
8	-0.47	-1.05	-1.11	-1.46	-1.39	-1.53	-1.76	-1.85	-1.85	-1.89
10	-0.91	-1.18	-1.27	-1.31	-1.40	-1.62	-1.52	-1.64	-1.90	-2.20
12	-1.05	-1.12	-1.13	-1.25	-1.33	-1.39	-1.50	-1.53	-1.71	-1.86

Table 4.5: Values of $\Delta_{X\to Y}$ for different settings of i and j ($X = uwlrf$, $Y = nbdsf$).

j \ i	10	20	30	40	50	60	70	80	90	100
2	0.97	0.30	0.39	-0.04	0.05	0.16	-0.27	-0.19	-0.10	-0.14
4	0.20	-0.21	-0.16	-0.52	-0.48	-0.41	-0.82	-0.84	-0.84	-0.74
6	0.60	0.13	0.11	-0.03	-0.17	-0.09	-0.47	-0.50	-0.47	-0.39
8	-0.05	-0.09	-0.04	-0.19	-0.25	-0.21	-0.47	-0.47	-0.63	-0.72
10	-0.31	-0.32	-0.31	-0.40	-0.53	-0.42	-0.62	-0.66	-0.75	-0.80
12	-0.24	-0.26	-0.27	-0.34	-0.44	-0.45	-0.65	-0.64	-0.78	-0.91

Table 4.6: Values of $\Delta_{Y\to X}$ for different settings of i and j ($X = uwlrf$, $Y = nbdsf$).

j \ i	10	20	30	40	50	60	70	80	90
2	-0.45	-	-	-	-	-	-	-	-
4	-0.56	-0.10	-0.34	-	-	-	-	-	-
6	-1.05	-0.66	-0.04	-0.21	-	-	-	-	-
8	-0.90	-0.45	-0.09	-0.25	-0.13	-0.39	-	-	-
10	-0.93	-0.71	-0.17	-0.42	-0.32	-0.22	-0.39	-	-
12	-1.12	-1.17	-1.13	-1.09	-1.04	-1.34	-1.28	-1.67	-1.39

Table 4.7: Heuristic method. Values of $\Delta_{Y\to X}$ for different settings of i and j ($X = prate$, $Y = soilm$), $\alpha = 2k/3$.

j \ i	10	20	30	40	50	60	70	80	90
2	-1.45	-	-	-	-	-	-	-	-
4	-2.43	-2.40	-2.24	-	-	-	-	-	-
6	-2.71	-2.63	-2.14	-2.11	-	-	-	-	-
8	-3.12	-2.71	-2.37	-2.68	-2.74	-2.23	-	-	-
10	-2.93	-2.80	-1.91	-2.03	-2.56	-2.06	-2.19	-	-
12	-1.52	-1.25	-0.92	-1.88	-1.26	-2.12	-3.34	-3.95	-4.57

Table 4.8: Heuristic method. Values of $\Delta_{Y\to X}$ for different settings of i and j ($X = prate$, $Y = soilm$), $\alpha = 2k/3$.

5 Extension to the Nonlinear Case

In high dimensions one often restricts the attention to linear models since more complex classes of functions tend to overfit the data unless the sample size is significantly higher than the dimension. Nevertheless, we want to say a few words about the nonlinear case. In Daniusis et al. [2010] we attacked the problem of inferring deterministic causal relations, i.e., when the underlying model is

$$Y = f(X)$$

for a cause X, an effect Y and a nonlinear function f. We assume f to be a diffeomorphism between the two domains of X and Y. The established theory in the paper is based on the already cited Postulate 2.1, namely that the probability density of the cause X is "independent" of the function f. With this formulation is meant That means that the structure of both is somehow independent in the sense that the peaks of $P(X)$ and positions where f has steep slop are uncorrelated. Under this assumption we obtain dependencies between the output and the inverted function. More precisely, no correlation between p_X and the slope of f implies a *positive* correlation between p_Y and the slope of $g := f^{-1}$.
Figure 5.1 illustrates what kind of dependencies between the output distribution and the function we can expect in the one-dimensional case. Note that the peaks of p_Y are exactly a the same position as the points of steep slope of $g = f^{-1}$. Such a nice illustration is not possible for the multidimensional case but one can imagine that we will encounter the same sort of dependencies there. The probability density of the output is given by

$$p_Y(y) = \left|\nabla f\left(f^{-1}(y)\right)\right| p_X\left(f^{-1}(y)\right).$$

As one can see p_Y depends on the shape of ∇f.
Starting with Postulate 2.1 we derived a second postulate, which in loose words states that the irregularities of the output are given by the irregularities of the input plus the irregularities of the function. We measure *irregularities* by estimating the relative entropy between the quantity to be measured and a smooth reference measure, usually an exponential family. We can view this as a certain kind of "distance" measure between the complexity of the probability densities and the function on one side and some smooth "simple" family of functions on the other side. Due to its foundation in information geometry the method got called *Information Geometric Causal Inference (IGCI).*

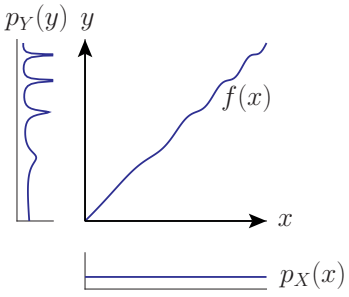


Figure 5.1: Illustration of the dependencies that occur if the density of X (input density) is uniform. Note that the output density $p_Y(y)$ is strongly peaked where the derivative of f' is small.

5.1 Information Geometric Causal Inference

ICGI in principle works for arbitrary domains. In our case where X and Y are high-dimensional random variables with probability densities p_X and p_Y we choose as reference measures the closest isotropic Gaussians to p_X and p_Y, respectively, as Subsection 3.3 of Daniusis et al. [2010] suggests. These Gaussians are given by the same mean as X and Y and the covariances $\tau_n(C_{XX})I$ and $\tau_m(C_{YY})I$, respectively, where I denotes the identity matrix. Let us denote them by u and v. Equation (9) in Section 3 of Daniusis et al. [2010] then tells us how to calculate the crucial quantity, namely

$$C_{X\to Y} = S(u) - S(v) + \int \log|\det J_f| p(x)dx \tag{5.1}$$

where J_f denotes the Jacobian of f and $S(\cdot)$ denotes the differential entropy, given by

$$S(p) = -\int_X p(x)\log(p(x))dx$$

for some probability distribution p on a probability space X. We found the following causal inference rule:

Given $C_{X\to Y}$, infer that X causes Y if $C_{X\to Y} < 0$, or that Y causes X if $C_{X\to Y} > 0$.

By whitening the data, i.e., the data are linearly transformed such that they have the identity as covariance matrix, the reference measures become the same and therefore $S(u) = S(v)$. Thus, (5.1) boils down to $C_{X\to Y} = \int \log|\det J_f| p(x)dx$ and it is decisive here to have a good estimator for the determinant of the Jacobian of f. We propose the following:
Denoting the indices of the n-nearest neighbours of x_i with $n_1(x_i), \ldots, n_n(x_i)$, then estimate $C_{X\to Y}$ by

$$\hat{C}_{X\to Y} = \frac{1}{k}\sum_{i=1}^{k} \log \left| \begin{array}{c} y^T_{n_1(x_i)} - y^T_i \\ \vdots \\ y^T_{n_n(x_i)} - y^T_i \end{array} \right| \Big/ \left| \begin{array}{c} x^T_{n_1(x_i)} - x^T_i \\ \vdots \\ x^T_{n_n(x_i)} - x^T_i \end{array} \right|, \tag{5.2}$$

where $|\cdot|$ is shorthand for the absolute value of the determinant. This estimator requires more samples than the number of dimensions since otherwise the determinant is not defined. There are several good k-nearest neigbour (kNN) algorithms available, we used for example one from Arya and Mount [1993][5].
Like it is shown in Daniusis et al. [2010] this inference method is robust in the low noise regime, i.e., in the case where

$$Y = f(X) + E$$

with a small noise term E.
Thus, we have a method to infer causal relations also in the nonlinear case, although in that case it is required that sample size exceeds dimensions.

[5] Code packages can be downloaded from Mount and Arya [2006] and Bagon [2009].

6 Discussion and Outlook

We have presented a method that is able to infer linear causal relations between two high-dimensional variables. We developed the theory based on concentration of measure phenomenon and free probability theory and proved identifiability results for the deterministic case. The main results are given by two Theorems in Chapter 3, namely Theorem 3.1 and Theorem 3.7. The proposed algorithm, presented in Chapter 4 is based on these findings. Although we couldn't show a similar result for the noisy case, experiments both on simulated and on real world data suggest that at least in the small noise regime we can apply the same method as in the deterministic case. A promising idea to attack the noisy case is an approach towards variable selection. We gave an outline and presented first experiments in this direction. Future work has to show if this approach can lead to a satisfactory algorithm. We also gave an idea of how to attack the nonlinear case.

We tested our algorithm both on simulated and real world data. The results are promising but it should be tested more extensively in particular with real world data sets, also to question the importance of the i.i.d. assumption for the entries of A. In the real world data experiments one could expect that if A was chosen randomly its eigenvalues should follow a semi-circle distribution. But in our case we always considered $A^T A$ which is a positive definite matrix and thus all eigenvalues are positive such that they cannot form a semi-circle distribution. The same is true for C_{XX}.

6.1 Towards a Statistical Test

We did not discuss the problem of statistical significance in the proposed inference method yet. In other words: How should we choose the ϵ in Algorithm 1? One possible approach here, already suggested in Janzing et al. [2010], may be to generate a large number of orthogonal matrices U, to apply them on X and then calculate the distribution of the values $\Delta_{UX \to Y}$. The observed value then defines a p-value and we can infer the right direction by comparing the p-values for both directions. For high dimensions though, this is computationally very expensive. Even with the fast algorithm for computing random orthogonal matrices which we presented in Section 4.2, this becomes intractable for dimensionality approaching 1000. Nevertheless, it can be used if input and output dimension is reasonably small.

Another approach could be derived from our insights about the violation of the trace multiplicativity in the backward direction. With the equations (3.40) and (3.41) we can determine the values of $\Delta_{Y \to X}$ for the wrong direction, depending on the sample size. It is possible then to estimate a distribution of these deltas and choose ϵ following a simple statistical test criterion.

6.2 Outlook

Voiculescu's theorem from Chapter 2 (Theorem 2.12) already suggests that there are different types of matrix ensembles resulting in free variables in the infinity limit. This means that the results shown in this work, in particular Theorem 3.1 and Theorem 3.7 also hold for slight variations in the assumptions made on the transfer matrix A.

We want to discuss another point: The method presented here can be extended in a natural way: Throughout the whole work, we only discussed the second moment of the respective distributions, but it is in principle possible to test the trace multiplicity on higher moments or powers of the covariance matrix, too. For example, one could test whether

$$\tau\left((A^T A C_{XX})^s\right) \approx \tau\left((A^T A)^s\right)\tau\left(C_{XX}^s\right) .$$

One can even imagine to use kernel methods for independence testing. Kernel methods are quite common in machine learning, Schölkopf and Smola [2002] give a good introduction into the field. There exists a widely used criterion for independence testing based on kernel methods, which is the so-called Hilbert-Schmidt independence criterion (HSIC). This criterion is able to detect all kinds of dependencies contained in the data (for a short description of HSIC see Gretton et al. [2008]). One could think of an adjustment of HSIC for our case.

The main assumption we made in the model is that the covariance of X is chosen independent of the transfer matrix A under a rotation invariant prior. This is a relatively strong assumption. One may find a way to obtain the same results under weaker conditions, in particular since it is not clear if this assumption holds in nature. It still remains to be fully clarified whether in the experiments of Section 4.4 the method inferred the right causal directions because the data sets fulfilled our assumptions, or because some other inherent structure of the data makes the method to give always these results. In other words, these few examples still do not verify the method completely. To give an example why to be precautious: In the examined data sets it is very likely that they contain a strong confounder, namely the season. That means the season both influences precipitation and soil moisture (and the other variables, too). Although by taking only every 12th month ($j = 12$) we circumvent this problem but still, a better way to do the experiments would be to use anomaly data. Anomaly data one can obtain in our case by subtracting from every (monthly) value the mean of that month averaged over all years. For further verification it is inevitable to extensively test the method, particularly on real world data sets of different domains.

Bibliography

S. Arya and D. M. Mount. Approximate nearest neighbor queries in fixed dimensions. In *Proceedings of the 4th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 271–280, 1993.

S. Bagon. Matlab class for ANN, 2009. URL `http://www.wisdom.weizmann.ac.il/~bagon/matlab.html`. Website, 10.4.2010.

P. J. Bickel and K. A. Doksum. *Mathematical Statistics*. Prentice-Hall, 1991.

P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI2010)*, 2010.

P. Diaconis and M. Shahshahani. The subgroup algorithm for generating uniform random variables. *Probability in the Engineering and Informational Sciences*, 1:15–32, 1987.

A. Edelman and N. Rao. Random matrix theory. *Acta Numerica*, 14:233–297, 2005.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of Royal Statistical Society*, 70:849–911, 2008.

J. Fan and J. Lv. A selective review of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010.

J. Fan, R. Samworth, and Y. Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10:1829–1853, 2009.

D. Geiger and D. Heckerman. Learning Gaussian networks. In *Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence (UAI1994)*, pages 235–243, 1994.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20:585–592, 2008.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with addidive noise models. In *Proceedings of the conference Neural Information Processing Systems (NIPS2008)*, 2009.

S. Janson. Large deviations for sums of partly dependent random variables. *Random Structures and Algorithms*, 24(3):234–248, 2004.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. To appear in *IEEE Transactions on Information Theory*, 2010.

D. Janzing, P. O. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

E. Kalnay et al. The ncep/ncar 40-year reanalysis project. *Bulletin of the American Meteorological Society*, (77):437–470, 1996.

Y. Kano and S. Shimizu. Causal inference using nonnormality. In *Proceedings of the International Symposium of Science of Modeling-The 30th anniversary of the Information Criterion (AIC)*, volume 1, pages 261–270, 2003.

G. Kreweras. Sur les partitions non croisées d'un cycle. *Disecrete Mathematics*, 1(4): 333–350, 1972.

M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Society, 2001.

J. Lemeire and E. Dirkx. Causal models as minimal descriptions of multivariate systems. Unpublished, accessible at http://parallel.vub.ac.be/, 2006.

J. Mooij and D. Janzing. Distinguishing between cause and effect. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 6:147–156, 2010.

D. M. Mount and S. Arya. ANN: A library for approximate nearest neighbor searching, 2006. URL `http://www.cs.umd.edu/~mount/ANN/`. Website, 10.4.2010.

G. J. Murphy. *C*-Algebras and Operator Theory*. Academic Press, 1990.

A. Nica. R-transforms of free joint distributions, and non-crossing partitions. *Journal of Functional Analysis*, 135:271–296, 1996.

J. Pearl. *Causality*. Cambridge University Press, 2000.

J. Peters, D. Janzing, and B. Schölkopf. Identifying cause and effect on discrete data using additive noise models. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9, pages 597–604, 2010.

S. Popescu, A. J. Short, and A. Winter. The foundations of statistical mechanics from entanglement: Individual states vs. averages. *Preprint on arXiv:quant-ph/0511225v3*, 2005.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

I. Schur. Neue Begründung der Theorie der Gruppencharaktere. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, pages 406–432, 1905.

I. Schur. Arithmetische Untersuchungen über endliche Gruppen linearer Substitutionen. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, pages 164–184, 1906.

S. Shimizu, P. O. Hoyer, and A. Hyvärinen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

R. Speicher. Multiplicative functions on the lattice of non-crossing partitions and free convolution. *Mathematische Annalen*, 298:611–628, 1994.

R. Speicher. Free probability theory and non-crossing partitions. *Séminaire Lotharingien de Combinatoire*, 39:38 pages, 1997.

R. Speicher. Free probability theory and random matrices. *Lectures at the European Summer School "Asymptotic Combinatorics with Applications to Mathematical Physics", St. Petersburg (Russia)*, 2001.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. Springer-Verlag, 1993.

G. W. Stewart. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, 17(3):403–409, 1980.

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

X. Sun, D. Janzing, and B. Schölkopf. Distinguishing between cause and effect via kernel-based complexity measures for conditional probability densities. *Neurocomputing*, pages 1248–1256, 2008.

W.-K. Tung. *Group Theory in Physics*. World Scientific Publishing Company, 1985.

D. Voiculescu. *Operator algebras and their connections with topology*, volume 1132 of *Lecture Notes in Mathematics*, chapter Symmetries of some reduced free product C*-algebras, pages 556–588. Springer-Verlag, 1985.

D. Voiculescu. Limit laws for random matrices and free products. *Inventiones Mathematicae*, 104:201–220, 1991.

D. Voiculescu, editor. *Free Probability Theory*. American Mathematical Society, 1997.

D. Voiculescu, K. J. Dykema, and A. Nica. *Free random variables*. American Mathematical Society, 1992.

E. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62:548–564, 1955.

E. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67:325–328, 1958.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI2009)*, 2009.

Selbstständigkeitserklärung

Selbstständigkeitserklärung:

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Tübingen, den

Einverständniserklärung:

Hiermit erkläre ich mich einverstanden, dass ein Exemplar meiner Diplomarbeit in der Bibliothek des Institutes für Mathematik verbleibt.

Tübingen, den