Causal Inference in Neuroimaging

Leonardo Casarsa de Azevedo

17th August 2015

Summary

A central goal of neuroscience is to understand the causal mechanisms connecting brain and mind. In fact, every randomized experiment tries to answer some causal question, such as, "What would happen with some neural activity if specific stimuli were presented?", or "How would a cognitive function change if one manipulated some brain area?" Causal inference tries to use previous observations (or experiments) to provide insights into the result of experiments which have not yet happened. In this work, I explore the application of the causal Bayesian Networks framework in neuroimaging settings.

The work is divided into two parts. In the first, we introduce the necessary mathematical concepts, review a promising method for the causal interpretation of neuroimaging results, and propose an extension to the method in a different experimental paradigm. In the second, we empirically illustrate the importance of the causal methods introduced, while addressing the neuroscientific question, "Can we infer the causes of covert shifts of spatial attention using EEG?" We explain the results we obtained and finish with suggestions for the future directions of causal inference in neuroimaging.

Acknowledgements

I would like to thank my supervisor Dr.-Ing. Moritz Grosse-Wentrup for all his support and for pointing me to the right direction from the beginning. I wish to extend a heartfelt thank you to Tatiana Fomina for all the patience, insightful discussions and suggestions, both in the academic and personal levels, and Joachim Werner and Nina Flad, for all the indispensable help in the setting up of the experimental paradigm.

My gratitude is beyond words to my mother, Eliane Casarsa, and my brother, Frederico Augusto Casarsa de Azevedo, without whom I would not aspire to be a scientist, and to Silvio Martins, for his tireless encouragement during the past four years. Thank you, David Ryan Birkman, who helped me a lot during the critical period of writing the thesis.

Finally, I would like to acknowledge my colleagues in the Max Planck Institutes for Biological Cybernetics and Intelligent Systems, for creating such an inspiring environment, where curiosity is optimally mapped into learning and discovery.

Contents

Chapter 1

Introduction

A central goal of neuroscience is to understand the causal mechanisms connecting brain and mind.[1] Such causal knowledge would enable us, for instance, to predict how behavior would change after manipulating neural activity or to choose the right neural interventions for treatment of a mental disease. The most straightforward way of obtaining this knowledge is through experiments. As an example, one can randomly intervene in a brain area and observe the changes in behavior or in the progression of a disease. However it is seldom straightforward to intervene in the brain. Even if it were not, the space of possible interventions is often too large to be explored by trial and error.

Causal inference attempts to use the information obtained from previous experiments or observations to predict, in some sense, causal relations between the observed variables. It is motivated by the link between statistical dependence and cause, summarized in Reichenbach's Common Cause Principle [Reichenbach, 1956]. Specifically, the principle states that two statistically dependent variables are either the cause and the effect of one another or that there is a hidden common cause to them. However, it has been observed [Weichwald et al., 2015] that causal terminology has often been used in Neuroimaging, even when the empirical evidence does not imply causality. The authors then used the framework of causal Bayesian Networks (CBN) to derive a set of rules for causal inference appropriate to neuroimaging data.

The method from Weichwald et al. distinguishes encoding and decoding models for analysis of neural data, as well as experimental settings in which neural recordings occur before or after an experimental condition. Encoding

[1] We adopt the following definition for mind - The element or complex of elements in an individual that feels, perceives, thinks, wills, and especially reasons. A mind enables consciousness, perception, memory, and attention. "Mind." Merriam-Webster.com. Merriam-Webster, n.d. Web. 17 Aug. 2015. http://www.merriam-webster.com/dictionary/mind.

models try to predict neural features from an experimental condition, while decoding models infer the condition from the features. Thus, each model provides a different piece of information connecting features and condition. In classical neuroimaging design, a stimulus often precedes the neural recordings, which can be followed by a behavioral response. As the time sequence of recordings and experimental condition imposes causal constraints, the distinction between stimulus and response-based experiments helps eliminate causal hypotheses. The identification of brain features that are relevant for the encoding or decoding models in stimulus-based experiments may lead to unambiguous causal conclusions. The current limitation of this approach lies within response-based paradigms, where the presence of hidden confounding variables hinders the identification of genuine neural causes for behavior. We propose an experimental setting composed of both stimulus and response, in which neural causes for behavior can be disentangled from hidden common causes.

We illustrate the empirical relevance of our results on EEG data recorded during a covert visuospatial attention paradigm [Posner et al., 1980]. We find that this paradigm is particularly suited to our needs, since it is composed of both stimulus and response, it has been extensively employed in the neuroimaging and BCI literature [Posner et al., 1980, Cohen and Maunsell, 2010, Thut et al., 2006, Van Gerven et al., 2009, Kelly et al., 2005], and there have been results linking stimulus to neural activity to behavioral response [Cohen and Maunsell, 2010, Thut et al., 2006], but to our knowledge, none of them have investigated the presence of hidden common causes.

Attention is not a unitary cognitive process, but rather, has various neurological underpinnings [Posner and Boies, 1971, Raz, 2004]. At least three functionally and anatomically different attentional systems have been identified through neuroimaging and patient population studies [Raz, 2004, Raz and Buhle, 2006, Posner and Boies, 1971], namely, alerting, orienting, and executive networks. The alerting network is responsible for the enhancement of processing of an expected stimulus and is composed mainly of frontal, thalamic, and parietal regions. The orienting network, related with the spatial shift of attentional focus, consists of the superior parietal cortex for covert shifts, and of the frontal eye fields and superior colliculus, for overt shifts of attention. The executive network is involved in voluntary control of attention, as well as in conflict management of processing in different neural areas, and has the anterior cingulate cortex as its main node. The degree of independence between the attentional networks remains unclear, but the three systems likely cooperate and work closely together [Raz and Buhle, 2006].

We chose the orientation change detection task, a paradigm for measuring

covert shifts of visuospatial attention based on Posner's central cue experiment [Posner et al., 1980]. Visuospatial attention is thought to improve perception for stimuli at an attended location, at the cost of stimuli elsewhere. Visuospatial attention shifts can be measured as an increase in the detection rate and decrease of detection latency of stimuli on the attended location [Posner et al., 1980]. Covert shifts of visuospatial attention, i.e., without head or eye movement, have been correlated with ipsilateral synchronization and contralateral desynchronization of α-Band (8-14 Hz) EEG activity over the occipital cortex [Klimesch et al., 1998, Thut et al., 2006, Sauseng et al., 2005,Rihs et al., 2007]. In particular, ipsilateral α-synchronization was found to dominate over contralateral desynchronization, to show a retinotopically organized patterns with stimulus location [Rihs et al., 2007] and this spatial pattern was able to predict detection latency of an upcoming visual target [Thut et al., 2006]. The prominent modulation of EEG activity with attention led to the development of Brain-Computer Interfaces, able to successfully decode the attended location with average performance greater than 70% [Kelly et al., 2005] and to serve as control signal for continuous control in a BCI setting [Bahramisharif et al., 2010].

Part I

Theory

Chapter 2

Causal Bayesian Networks

Causal Bayesian Networks is a framework for building graphical models around associational information and interpreting them causally. The associational information are conditional independence (CI) relations between the variables. New CI conditions can be generated from collections of CI relations through logical steps [Dawid, 2010]. However, these steps can become complicated, especially if the collection of CIs is large. Graphical models provide a way of compactly representing CIs and of inferring new ones geometrically [Pearl and Paz, 1985]. Due to the connection between causality and conditional independence, graphical models have become a useful language for causal inference [Reichenbach, 1956, Pearl, 2000, Spirtes et al., 2000]. In this chapter, I give an introduction to causal Bayesian Networks (CBNs). The chapter starts with the necessary definitions regarding graphs and probability, followed by the conditions under which one can interpret a graph probabilistically, and ends by relating causal inference and probabilistic graphs.

2.1 Graphs

A **graph** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an ordered pair consisting of a set of vertices, or nodes, $\mathcal{V}$ and a set of edges $\mathcal{E}$. An edge (A, B) in a graph is a pair of distinct vertices from $\mathcal{G}$, i.e., $(A, B) \in \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, with $A \neq B$. A directed edge is an edge composed by an ordered pair of vertices, in which $(A, B) \neq (B, A)$. The directed edge from A to B is named (A, B) and is represented as $A \to B$. There is an edge between A and B if $A \to B$ or $B \to A$.

A **path** between two nodes A and Z in $\mathcal{G}$ is a sequence of n vertices $\{V_1, \ldots, V_n\}$, with $V_1 = A$ and $V_n = Z$, such that there is an edge between any two adjacent nodes V_i and V_{i+1}, with $i = 1, \ldots, n-1$. A directed path

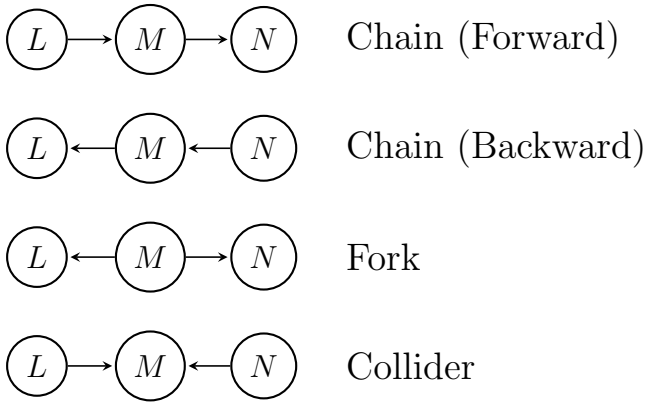

Figure 2.1: Edges between adjacent nodes L, M, N in an acyclic path can be oriented forward or backward. Three configurations emerge from the four possibilities: a chain (forward or backward), a fork, or a collider. If there is a collider in a path, the path is said to be blocked by the empty set.

from A to Z, $A \dashrightarrow Z$, is a path in which $V_i \to V_{i+1}$ for every i. Cyclic paths are paths in which any node appears more than once in the sequence. Empty paths, which contain only one node, will not be considered in this work, i.e., for us, every path has at least two distinct nodes.

A directed acyclic graph (**DAG**) is a graph with only directed edges and no directed cyclic paths. The relationship between variables in a DAG leads to useful properties. To explore them, the concepts of parent, ancestor, descendant, and non-descendant in a DAG must be defined.

A node A is an **ancestor** of node B, if $A \dashrightarrow B$, in which case B is called a **descendant** of A. Parents are the closest ancestors, i.e., A is a **parent** of B if $A \to B$. The set of all parents of B is called pa(B), and the set of all non-descendants is called nd(B).

I discuss the possible configurations of adjacent nodes on a path, which will lead to the concept of d-separation [Pearl, 1985].

2.1.1 d-Separation

Let L, M, N be three adjacent nodes on an (acyclic) path between A and Z. There are three possible variable configurations for nodes in a path, namely, a chain (forward or backward), a fork, and a collider (see Figure 2.1). Whenever there are only chains or forks in the path, the path is said to be **active**. Notice that L and N will only share ancestors in active paths. A path between A and Z is **blocked** by a set $\mathcal{S}$, with $A, Z \notin \mathcal{S}$, whenever it contains a middle node M, such that one of the two conditions is true:

- the path is active and M belongs to $\mathcal{S}$
- the path is not active, M is the center of a collider, and neither M or any of its descendants belong to $\mathcal{S}$.

It follows that every path containing a collider is blocked by the empty set. If all the paths between nodes A and Z in DAG $\mathcal{D}$ are blocked by some $\mathcal{S}$, A and Z are directed-separated, or ***d*-separated**, by $\mathcal{S}$, and represent it by $A \perp_d Z|\mathcal{S}$ $[\mathcal{D}]$ [Pearl, 1985]. Whenever it is clear from the context, $\mathcal{D}$ is omitted.

2.2 Probability

I provide the necessary notation and definition on random variables, probability, independence, and conditional independence. I refer to [Dudley, 2002] for a more detailed measure-theoretic take on Probability.

From here on out, random variables (r.v.) are written in uppercase, observations of r.v. in lowercase, and sets of random variables in calligraphic letters. All the random variables in this work will be multivariate real-valued, binary, or a mixture of both, i.e., assuming values in $\mathbb{R}^n \times \{0,1\}^k$, with $n, k = 0, 1, 2, \ldots$.[1]

$P(\mathcal{X})$ or $P(X_1, \ldots, X_m)$ is the joint distribution on the set of r.v. $\mathcal{X} = \{X_1, \ldots, X_m\}$, i.e., the countably additive probability measure on the cartesian product of random variables. $P(X_1)$, and $P(X_2|X_1)$ are, respectively, examples of marginal and conditional distributions. The corresponding joint, marginal and conditional probability mass or density functions are, respectively, $P(x_1, \ldots, x_m)$, $P(x_1)$, and $P(x_1|x_2)$. Although the notation of P is overloaded, its meaning should be clear in the context it is used.

Independence. X is independent of Y, or $X \perp\!\!\!\perp Y$,[2] whenever one can factorize the joint distribution,

$$P(X = x, Y = y) = P(X = x)P(Y = y), \tag{2.1}$$

for every x, y. $X \not\perp\!\!\!\perp Y$ denotes dependence between the variables.

Conditional Independence. X is independent of Y conditional on Z, or $X \perp\!\!\!\perp Y|Z$, whenever the conditional distributions factorize,

[1] If either n or k are 0, the variable assumes values only in $\{0,1\}^k$ and $\mathbb{R}^n$, respectively

[2] A more precise representation would be $X \perp\!\!\!\perp Y$ $[P]$ to denote that X and Y are independent in distribution P. However, this is usually clear from the context, and P is omitted for a more concise notation.

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z), \tag{2.2}$$

for every x, y, z. One can use a similar definition for the conditional independence of X and Y conditioned a set $\mathcal{Z} = \{Z_1, \ldots, Z_m\}$, and refer to it as $X \perp\!\!\!\perp Y|\mathcal{Z}$.

2.3 DAGs and Probability

Given a set of variables $\mathcal{V}$ and a collection $\mathscr{C}$ of conditional independence relations (CIR) among the variables, it is often possible to logically deduce new CIRs from $\mathscr{C}$ by applying universal conditional independence properties [Dawid, 2010]. However, these operation can become complicated, especially for large $\mathscr{C}$, and mathematical constructions to represent and manipulate CIRs become necessary. DAGs equipped with d-separation provide a useful graphical language for expressing dependencies among the variables.

A probability distribution P over a set of variables $\mathcal{V}$ can be represented by a DAG $\mathcal{D}$ with corresponding set of vertices $\mathcal{V}$ if,[3] for every $A, Z \in \mathcal{V}$ and $\mathcal{S} \subset \mathcal{V} \setminus \{A, Z\}$,

$$A \perp_d Z|\mathcal{S}\ [\mathcal{D}] \implies A \perp\!\!\!\perp Z|\mathcal{S}\ [P]. \tag{2.3}$$

An equivalent and common way of stating equation 2.3 is the Markov condition,

Markov Condition. A DAG $\mathcal{D}$ over $\mathcal{V}$ and a probability distribution P satisfy the Markov condition if, for each $V \in \mathcal{V}$, V is conditionally independent of its non-descendants, nd(V), given its parents, pa(V),

$$V \perp\!\!\!\perp \text{nd}(V)|\text{pa}(V). \tag{2.4}$$

$\mathcal{D}$ and P satisfy the Markov condition if, and only if,[4] the **Markov factorization** is true, i.e.,

$$P(\mathcal{V}) = \Pi_{i=1}^{n} P(V_i|\text{pa}(V_i)). \tag{2.5}$$

Although the Markov property allows to represent independences in the DAG, one cannot conversely infer dependencies from the DAG. If, however, one assumes that the converse of condition 2.3 also holds, one has a representation in which the presence and absence of d-separation imply conditional (in)dependency relations. This additional condition is called Faithfulness.

[3] Here, I purposefully make a notation overload to imply that the r.v. A and Z have corresponding nodes A and Z in the DAG $\mathcal{D}$.

[4] As long a a corresponding density function for P exists.

Faithfulness. A DAG $\mathcal{D}$ over $\mathcal{V}$ and a probability distribution P satisfy the faithfulness condition if,

$$A \perp_d Z|\mathcal{S}\ [\mathcal{D}] \iff A \perp\!\!\!\perp Z|\mathcal{S}\ [P]. \tag{2.6}$$

Though they provide powerful language, probabilistic DAG representations have limitations. To begin, DAG representations of CIRs are far from complete in the sense that there are collections of consistent CIRs that have no DAG representation [Dawid, 2010]. Additionally, many DAGs can be used to represent the same CIR. For example, it is easy to check that the Markov factorization of the three-node DAGs, consisting of a backward chain, a forward chain, and a fork (as in Figure 2.1), imply the same CIRs.

Distinct DAGs, which represent the same set of CIRs $\mathscr{C}$, are called **Markov equivalent**, and the set of all equivalent DAGs representing the same $\mathscr{C}$ is called a **Markov equivalence class**. Markov equivalence classes are fundamental for the causal analysis method to be discussed on the following.

DAGs in a Markov equivalence class are characterized by having the same skeleton and immoralities. A **skeleton** is the copy of a DAG with only undirected edges, while an **immorality** is a collider with disconnected parents, e.g., $L \to M \leftarrow N$, with no direct edge between L and N.

In a Markov equivalence class, some edges are the same for every DAG, providing an unambiguous representation, while others are ambiguously represented. One can represent the equivalence class with Complete Partially Directed Acyclic Graphs (CPDAG), in which the ambiguous edges are undirected, and the unambiguous are directed. I will now discuss an **independence-based** method for recovering a CPDAG from the data.

2.3.1 Independence-Based CPDAG Generation

Let $\mathcal{V}$ be a set of random variables observed under a stable regime over a series of repetitions, and let $\mathscr{C}$ be the collection of CIRs obtained from statistical tests from the data. Given the Markov condition and Faithfulness, one can generate a CPDAG from the data in two steps,

1. Estimation of skeleton

2. Orientation of edges

Different methods use different properties in each step, such as the PC [Pearl, 2000] or SGS [Spirtes et al., 2000] algorithms.

For the estimation of the skeleton, it is useful to note that two nodes $A, Z \in \mathcal{V}$ in a DAG are adjacent if and only if $A \not\perp\!\!\!\perp Z|\mathcal{S}$ for every $\mathcal{S} \subset \mathcal{V} \setminus \{A, B\}$. So starting with a fully connected skeleton, one can remove the undirected edges between each pair of variables whenever any conditional independence between the nodes is observed.

For the orientation of edges, one first checks for possible immoralities in the skeleton. One searches for structures of the form $L - M - N$, and tests if $L \not\perp\!\!\!\perp N|\mathcal{S}_M$, for every $\mathcal{S}_M$, with $M \in \mathcal{S} \subset \mathcal{V} \setminus \{L, M\}$. One can further orient edges to avoid the formation of cyclic DAGs and to respect problem-specific constraints.

Although independence-based methods [Pearl, 2000, Spirtes et al., 2000] are generally used for causal inference, I wanted to cover them before causality is mentioned, to emphasize that the properties in which they are based are already shared by probabilistic DAGs. The last step needed for the framework of causal Bayesian Networks is a definition of causality and its relationship to probability distributions.

2.4 Causality and Probability

Previously, DAGs were used to express and infer CIRs. In the framework of causal Bayesian Networks, DAGs express causal relations, and each direct edge corresponds to a causal statement,

$$A \to B \iff A \text{ directly causes } B. \tag{2.7}$$

Causality in this framework is understood in terms of interventions. An intervention on a variable A in a DAG $\mathcal{D}$ representing $\mathcal{V}$ is an external manipulation of the values of A alone, i.e., independent on the causes (ancestors) of A. A classical example of interventions are randomized experiments in which the values of A are set to an appropriate distribution, such as uniform or gaussian.

A is said to be a cause of B if and only if an intervention on A changes the probability distribution of B. The resulting distribution over variables $\mathcal{V}$ after an intervention in $A \in \mathcal{V}$ is denoted as $P(\mathcal{V}|do(A))$. Therefore,

$$A \text{ causes } B \iff P(B|do(A)) \neq P(B) \text{ for some values of } A \text{ and } B.$$

Intervening is not the same as conditioning. For example, if A is the cause of B, then $P(B|A) \neq P(B)$, and $P(A|B) \neq P(A)$ but $P(A|do(B)) = P(A)$. Nevertheless, inspired by Reichenbach's Common Cause Principle, it is desirable to have a correspondence between causal relations and CIRs. For

identifying graphical relations with CIRs, the Markov condition in probabilistic DAGs can be used. It can be shown that a causal DAG in which the Markov condition holds will also follow the Common Cause Principle [Reichenbach, 1956].

If additionally to the Markov condition, Faithfulness holds, one can use the independence-based methods described previously to generate a causal CPDAG. Causal inference in CBNs consists of generating a causal CPDAG and interpreting the directed edges causally. In the following chapter, I show how one can use problem-specific constraints to obtain a causal CPDAG with more oriented edges.

As a side note, there is no a priori reason to believe that causal DAGs should follow Markov or faithfulness conditions. In fact, there are special quantum systems which do not obey the causal Markov condition [Elby, 1992], as there are sequences of faithful distributions which converge to unfaithful ones [Robins et al., 2003]. Nevertheless, if one is to infer causality, one must accept the necessary assumptions. In the CBN framework, it is assumed that the Markov and Faithfulness conditions are true.

Chapter 3

Causality in Neuroimaging

A promising application of CBNs was recently presented by Weichwald et. al [Weichwald et al., 2015]. The authors observed that, in the neuroimaging literature, the results of encoding and decoding methods are often interpreted causally, and investigated when the causal conclusions are supported by the empirical data. The underlying procedure was similar to an independence-based CPDAG generation, namely, first the CIRs resulting from the analysis methods were collected, then the skeleton was estimated and the edges were oriented based on the presence of immoralities, and experimental constraints specific to the neuroimaging setup. I will not derive the set of causal interpretation rules from [Weichwald et al., 2015], but rather I will explain the context in which they can be properly understood, and provide new insights to the rules. I conclude with the discussion of further causal interpretation rules in a neuroimaging setup not explored by the authors.

3.1 Encoding and Decoding models

The most popular and sensitive techniques for analyzing neurimaging information distinguish between encoding and decoding models [Naselaris et al., 2011]. The two models work in mirrored perspectives. While encoding infers the conditional probability distribution $P(\mathcal{X}|C)$ over a set of brain recordings in $\mathcal{X}$, given experimental conditions C, decoding uses $\mathcal{X}$ to predict C, i.e., it models $P(C|\mathcal{X})$. Although it has been argued [Naselaris et al., 2011] that encoding models are superior to decoding ones in the sense that only encoding can provide a complete functional description of a region of interest, it has been shown [Weichwald et al., 2015] that a combination of the two models applied to the same data might produce novel causal insights.

The first step in the causal inference method is the collection of condi-

tional independence relations (CIR). For that, one needs to distinguish features $X \in \mathcal{X}$, which are relevant for the encoding or decoding of a condition C. It can be shown that relevance to encoding or decoding is equivalent to the following conditional dependence relations [Strobl et al., 2008, Weichwald et al., 2015],

$$X \in \mathcal{X}^{+\mathrm{enc(C)}} \iff X \not\perp\!\!\!\perp C \tag{3.1}$$

$$X \in \mathcal{X}_{+\mathrm{dec(C)}} \iff X \not\perp\!\!\!\perp C | \mathcal{X} \setminus X \tag{3.2}$$

Therefore, a neural feature X is directly connected to the condition C in the causal CPDAG only if it is relevant for both encoding and decoding. This is only a necessary condition, though, since adjacent nodes in a CPDAG are dependent on each other conditional on each other possible subset $\mathcal{S} \subset \mathcal{X} \setminus X$. From this observation, it becomes clear that the causal interpretation rules derived from decoding and encoding models could never identify direct from indirect causal relations.

3.2 Experimental Constraints

I investigate the particular constraints that neuroimaging experimental design imposes for the orientation of edges in the CPDAG. The setting consists of recording of brain activity $\mathcal{X}$ either after a stimulus S or before a behavioral response R. Stimuli are often randomized and therefore cannot have ancestors in the causal DAG. Due to the inherent time structure, behavioral responses R cannot be a cause of previously recorded activity. CPDAGs from stimulus or response-based experiments will, therefore, have different constraints. Additional constraints can be derived from paradigms which contain both stimulus and response phases, as long as the time sequencing of stimulus, neural recordings and response are preserved. I further discuss one causal inference method in the section 3.4.

3.3 Causal Interpretation Rules

I now give two examples from [Weichwald et al., 2015] on causal inference for features relevant for encoding in stimulus and response-based settings.

$\mathbf{X} \in \mathcal{X}^{+\mathbf{enc(S)}}$: X and S are dependent. From the Markov condition, there is an active path between X and S. Due to randomization, S can have no ancestor in the causal DAG. Therefore, $S \dashrightarrow X$.

$\mathbf{X} \in \mathcal{X}^{+\mathbf{enc(R)}}$: X and R are dependent. From the Markov condition, there is an active path between X and R. Due to time sequencing, $R \not\dashrightarrow X$. It is still possible that either $X \dashrightarrow R$ or that there is a fork (common cause) X and R.

Through a sequence of similar reasoning steps, the authors derived the causal interpretation rules in table 3.1.

	relevance in encoding	relevance in decoding	causal interpretation	rule
stimulus-based	√		X effect of S	S1
	×		X no effect of S	S2
		√	⚠ inconclusive	S3
		×	⚠ inconclusive	S4
	√	√	X effect of S	S5
	√	×	X indirect effect of S	S6
	×	√	provides brain state context	S7
	×	×	neither effect nor provides brain state context	S8
response-based	√		⚠ inconclusive	R1
	×		X no cause of R	R2
		√	⚠ inconclusive	R3
		×	⚠ inconclusive	R4
	√	√	⚠ inconclusive	R5
	√	×	X no direct cause of R	R6
	×	√	provides brain state context	R7
	×	×	neither cause nor provides brain state context	R8

Table 3.1: Causal interpretation rules for relevant (√) and/or irrelevant (×) features X_i in encoding and decoding models for stimulus ($C = S$) and response-based ($C = R$) paradigms. Extracted from [Weichwald et al., 2015], with permission from the authors.

I would like to detail some insights which can be obtained from the causal interpretation rules.

1. All (observed) direct effects of S are in $\mathcal{X}^{+enc(S)}_{+dec(S)}$.

2. All (observed) direct causes of R are in $\mathcal{X}^{+enc(R)}_{+dec(R)}$.

3. Every cause of R in $\mathcal{X}^{+enc(R)}_{-dec(R)}$ is indirect.

In the discussion of rules S5 and R5 from table 3.1, the authors [Weichwald et al., 2015] declared that the intuition that $X_i \in \mathcal{X}^{+enc(C)}_{+dec(C)}$, with $C \in \{S, R\}$ is in some sense closer to C than $X_i \in \mathcal{X}^{+enc(C)}$ was wrong. I argue, however, that since all (observed) directly related features are in $\mathcal{X}^{+enc(C)}_{+dec(C)}$, the features in the set are closer to C in a non-deterministic way.

If the objective of an experimenter is to find features which are direct effects of a stimulus, one should prefer features in $\mathcal{X}^{+enc(S)}_{+dec(S)}$ than in $\mathcal{X}^{+enc(S)}_{-dec(S)}$ or $\mathcal{X}^{+enc(S)}$ for two reasons. First, if there is a indirect effect, there must be a direct one, though the converse is not true. Second, by randomly picking a feature X in $\mathcal{X}^{+enc(S)}_{+dec(S)}$, there is a probability $\alpha > 0$ of this feature being a indirect effect, and $1-\alpha$ of it being a direct effect. However, if $X \in \mathcal{X}^{+enc(S)}_{-dec(S)}$, X will be an indirect effect with probability 1. In other words, although $\mathcal{X}^{+enc(S)}_{+dec(S)}$ may contain direct or indirect effects, the proportion of direct effects in $\mathcal{X}^{+enc(S)}$ and $\mathcal{X}^{+enc(S)}_{-dec(S)}$ is less than in $\mathcal{X}^{+enc(S)}_{+dec(S)}$.

Likewise, in response-based settings, although there might be no cause among the observed neural features, $\mathcal{X}^{+enc(R)}_{+dec(R)}$ are still closer to R. If there were a test for identifying genuine causes of R, there could be a cause present in $\mathcal{X}^{+enc(R)}_{+dec(R)}$ and none in $\mathcal{X}^{+enc(R)}_{-dec(R)}$, while the converse is not true. Therefore, given that there is a cause for response in the neural features, the proportion of direct causes is greater in $\mathcal{X}^{+enc(R)}_{+dec(R)}$.

In fact, there are tests for detecting the absence of hidden common causes in an experimental setting composed of both stimulus and response.

3.4 Genuine Cause Identification

In an experimental setting composed of both stimulus and response, with $S \not\perp\!\!\!\perp R$, it is possible to identify effects of stimulus which are genuine causes of response. Indeed, if a brain feature X d-separates S and R, there can be no hidden common cause between X and R. To see that, first remember from 2.1.1 that if $S \perp_d R|X$, every active path between S and R must contain X. Now, because $S \not\perp\!\!\!\perp R$, there is at least one active path between S and R. Given that a randomized S can have no ancestors in the causal DAG, $S \not\perp\!\!\!\perp R \implies S \dashrightarrow R$ follows from the Markov condition. Finally, one needs a statistical test for identifying d-separation. Such a test follows from Faithfulness, expressly, $S \perp\!\!\!\perp R|X \implies S \perp_d R|X$. The procedure for identification of genuine cause is summarized in the following.

Test for Genuine Cause. In a neuroimaging experiment based on both stimulus S and response R, the following conditions are sufficient for concluding, in the CBN framework, that a brain feature X causes R:

1. $S \not\perp\!\!\!\perp R$
2. $S \perp\!\!\!\perp R|X$.

It should be emphasized that the previous test is sufficient, but not necessary, for X being a cause of R. Additionally, whether X is a direct or indirect cause for R is not identifiable by these rules.

A straightforward consequence of the previous conditions is that for every genuine cause X, $X \in \mathcal{X}^{+enc(S,R)}$.[1] In particular, whenever $X \in \mathcal{X}^{+enc(S,R)}_{-dec(S,R)}$ is a genuine cause of R, there will be a feature $Y \in \mathcal{X}^{+enc(S,R)} \setminus \mathcal{X}^{+enc(S,R)}_{-dec(S,R)}$, which is also a genuine cause for R. The converse, however, is in general not true. One can therefore start the search for genuine causes with $\mathcal{X}^{+enc(S,R)} \setminus \mathcal{X}^{+enc(S,R)}_{-dec(S,R)}$.

[1] For compact notation, I consider $\mathcal{X}^{+enc(S,R)} := \mathcal{X}^{+enc(S)} \cap \mathcal{X}^{+enc(R)}$.

Part II

Experiments

Chapter 4

Methods

In this chapter, we describe the experimental setup and explain the methods used for behavioral and neuroimaging analysis. We picked a spatially cued orientation change detection paradigm and modeled the experimental outcomes as a Bayesian Network to derive conditions under which there is a dependence between stimulus and response. We explain which behavioral features we used to characterize a response and the constraints that led to this decision. Furthermore, we describe the recording setup and the pre-processing steps necessary for extracting the relevant brain features. Finally, we explain the classification methods as well as the statistical tests used for the causal interpretation rules for encoding and decoding models in each of the stimulus and response-based settings, and present an additional conditional independence test used for investigating the presence of hidden confounders in the data.

4.1 Experimental Setup

A spatially cued orientation change detection task was used. This paradigm was designed as a variant of Posner's central cue experiment [Posner et al., 1980]. The sequence of events depicting the task are shown in Figure 4.1.

Participants were seated in front of a CRT monitor and were asked to place their heads on a chin rest located 50 cm from the monitor, centered in relation to the monitor, and to move as little as possible during the recording session. Eye movements were controlled using an iView X RED pt eye tracking system (SensoMotoric Instruments GmbH (SMI), Teltow, Germany), with sampling at 50 Hz. Stimuli were presented on a 22 inches CRT screen with a refresh rate of 85 Hz and maximum resolution of 1920 $\times$ 1440. The eye tracker was mounted to the left of the monitor and the

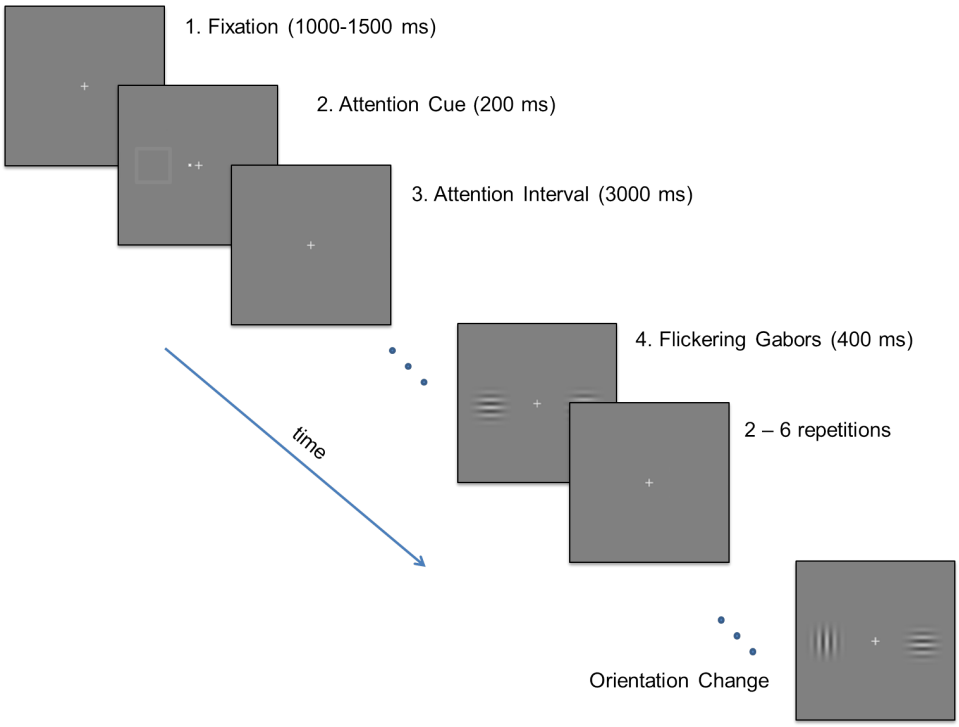

Figure 4.1: Schematics of the orientation change detection task. A trial contained the following sequence of events: Fixation, Attention Cue, Attention Interval, Flickering Gabors, and Orientation Change. The brain state features used in our data analysis were extracted from the signal recorded during the 3 s of the attention interval. During the Flickering Gabors phase, two symmetric parallel Gabors flashed on the screen for 200 ms and off for 200 ms, for a total of 2 to 6 times. Upon an unsignaled on-flash, one of the Gabors had its orientation changed by an angle θ, adapted to each participant. 80% of the trials were valid cue trials, in which the change of orientation and attention cue appeared in the same hemifield of vision. The participant had 800 ms after change of orientation to report the side in which the stimulus changed. Participants received positive auditory feedback for correctly detecting the change, in valid or invalid cue trials.

infrared emitter was mounted underneath it. No stimulus was presented vertically, therefore, we chose to track eye movements only horizontally. The eye tracker was calibrated before every experimental session, at least once every 10 min during a session, as well as whenever tracking was lost. We wrote our experiments in MATLAB (The MathWorks, Inc.), using the Psychophysics Toolbox extensions [Brainard, 1997, Kleiner et al., 2007].

A trial started with the presentation of a white fixation cross (0.5° long, symmetric) in the center of a uniform gray background. During a trial, participants were asked to fix their eyes on the fixation cross for the duration of the trial.

After the 1000 ms fixation period, a white dot (0.2° diameter) and a 90% gray square frame (10° long) were presented for 200 ms, located respectively at 1° and 15° from the center, either to the left or to the right of the screen. The dot and frame served as spatial attention cues, instructing participants to covertly shift their attention to the corresponding hemifield of vision. The square frame's position and size matched the position and size of the Gabor patches presented afterwards.

A 3000 ms attention interval followed, during which only the fixation cross was exhibited. The brain state features used in our analysis were extracted from the EEG signal recorded during this interval.

In the flickering Gabors phase, two achromatic, 10% ± 5 contrast, odd-symmetric Gabor patches located at 15° from the center in each visual hemifield flashed synchronously on for 200 ms and off for 200 ms for a maximum of six repetitions. During an unsignaled on-flash, the orientation of one of the Gabor patches changed, and the participant was asked to detect the change as quickly as possible by pressing a button with the right or left index fingers. Response was considered correct if button press and orientation change happened on the same side. A beep of higher (800 Hz, 0.25 s) or lower (200 Hz, 0.25 s) frequencies signaled correct and incorrect responses, respectively. Correct responses after 800 ms of orientation change were considered incorrect in the behavioral analysis, but were signalized as correct to the participant.

The trial ended with a blank screen and a 2000 ms intertrial interval, during which the participants could blink and freely move their eyes. Whenever eye fixation was broken outside the intertrial interval, the trial was aborted and restarted. After every 60 trials (around 15 min), the participants had a longer break (2-5 min long), during which they could freely move. Participants were allowed to take additional breaks during a session whenever they felt tired.

Stimulus change never occurred on the first on-flash in order to prevent precipitated guessing. 20% of the trials were invalid cue trials, in which the

attention cue and orientation change appeared in different sides of the screen, while the remaining 80% were valid cue trials. The participant received positive feedback for correctly detecting a stimulus change regardless of the attention cue.

Each participant performed one experimental session with three runs of 120 trials. Each experimental session was preceded by two training runs, for the purpose of ensuring that the task was well understood, reducing the effects of behavioral learning, and allowing a proper choice of angle for orientation change. The only difference between training and trial runs was that the attention cues did not disappear during the attention interval in the training runs, to give the participants the opportunity to learn how to covertly shift attention. A proper angle for orientation change was the one in which the mean detection rate and latency were significantly larger and shorter, respectively, in valid than in invalid cue trials. If no angle fulfilled these conditions, a third training run with new angles was introduced.

4.2 Behavioral Analysis

Covert shifts of attention affect the efficiency of detecting events at different spatial locations [Posner et al., 1980]. Detection rate and detection latency are expected to be higher and lower, respectively, in attended versus non-attended locations. Although it is not possible, in our framework, to assess the attentional state of a participant on a single-trial level, we assume for our analysis that the mean behavioral outcome of valid cue trials, in which the attention cue and orientation change occur to the same side, is representative of attention, when compared to invalid cue trials. To synthesize our assumptions and easily draw conclusions, we model our problem using a Bayesian Networks framework. We start by quickly reviewing the paradigm to recover the observed variables and then make further assumptions to parametrize the non-observed variables.

In the beginning of a trial, participants are shown an stimulus (S), instructing them to shift their attention (A) to one of two sides. In the response phase, they are asked to detect the side of a target orientation change (T), by providing a response (R) as quickly as possible. In 80% of the trials, the target orientation change T will occur on the same side of the stimulus S. Stimuli are shown equally often on the two sides.

We assume the participants successfully shift their attention to the cued side with probability π. We additionally consider that each participant has the same π for both sides (and discuss this hypothesis in section 6.1). We expect that the behavioral response will be affected by the shift of atten-

tion only if the target orientation change happens in the attended location. We group the behavioral outcome into two categories, negative or positive, without, for now, dwelling on their meanings. We name the probability of response being positive given that attention was located on the side of orientation change ρ. If attention was located on the opposite side of orientation change, the probability of positive response was $1 - \sigma$.

The previous conditions are summarized in the DAG in Figure 4.2, as well as in the following remark.

Remark 1. In the spatially cued orientation change detection task, let the random variables $S, T, A : \Omega_S \to \{0, 1\}$, and $R : \Omega_+ \to \{0, 1\}$, with respective sample spaces $\Omega_S = \{\text{left}, \text{right}\}$, and $\Omega_+ = \{\text{negative}, \text{positive}\}$ be the variables of interest.

We have, by design, for $x \in \{0, 1\}$,

$$\begin{aligned} P(T = x|S = x) &= 0.8 \\ P(S = x) &= 0.5. \end{aligned}$$

And, by assumption, for $x, y \in \{0, 1\}$, such that $y \neq x$,

$$\begin{aligned} P(A = x|S = x) &= \pi \\ P(R = 1|A = x, T = x) &= \rho \\ P(R = 1|A = x, T = y) &= 1 - \sigma. \end{aligned}$$

Although we are now dealing with four (conditional) probability distributions for R, we expect a behavioral response to be modulated not by stimulus side, but rather, by attentional state. We show that this is still the case in our model, and that response is dependent on stimulus side conditional on the target orientation change. Therefore, we want to show $P(R|S = 0, T) \neq P(R|S = 1, T)$.

$$\begin{aligned} P(R|S, T) &= \sum_A \frac{P(S, T, A, R)}{P(S, T)} \\ &= \sum_A \frac{P(R|T, A)P(T|S)P(A|S)P(S)}{P(T|S)P(S)} \\ &= \sum_A P(R|T, A)P(A|S) \\ &= P(R|T, A = 0)P(A = 0|S) + P(R|T, A = 1)P(A = 1|S) \end{aligned}$$

If we substitute the values for the conditional probability distributions with the parameters from remark 1, we get the following contingency table.

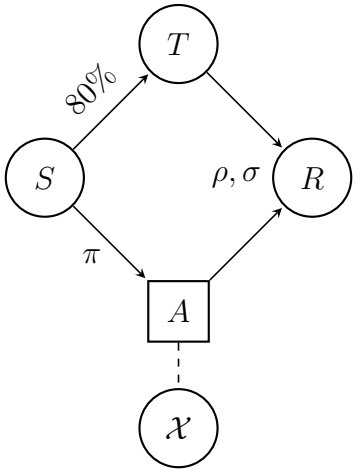

Figure 4.2: DAG modeling the behavioral paradigm. S, T, A, R are binary r.v. , corresponding to Stimulus for attention cue, Target orientation change, Attention shift, and Response, respectively, while $\mathcal{X}$ represents the recorded brain features. We have no direct measures of attention, therefore, A is a hidden variable represented by a square. $\mathcal{X}$ is a (potentially indirect) measure of attention, thus is connected to A by a dashed line.

		$R=0$	$R=1$
$S=0$	$T=0$	$(1-\rho)\pi+\sigma(1-\pi)$	$\rho\pi+(1-\sigma)(1-\pi)$
	$T=1$	$\sigma\pi+(1-\rho)(1-\pi)$	$(1-\sigma)\pi+\rho(1-\pi)$
$S=1$	$T=0$	$(1-\rho)(1-\pi)+\sigma\pi$	$\rho(1-\pi)+(1-\sigma)\pi$
	$T=1$	$\sigma(1-\pi)+(1-\rho)\pi$	$(1-\sigma)(1-\pi)+\rho\pi$

Notice that the inner rows of the contingency table have the same values, and that the same is true for the outer rows. Therefore, it is possible to rewrite the previous contingency table in terms of valid and invalid cue trials, $S = T$ and $S \neq T$, respectively.

	$R=0$	$R=1$
$S=T$	$(1-\rho)\pi+\sigma(1-\pi)$	$\rho\pi+(1-\sigma)(1-\pi)$
$S\neq T$	$(1-\rho)(1-\pi)+\sigma\pi$	$\rho(1-\pi)+(1-\sigma)\pi$

We can conclude that $R \perp\!\!\!\perp S|T$ if $\pi = 0.5$, since for other values of π, $P(R|S = T)$ and $P(R|S \neq T)$ are not generally identical. In other words, only when the stimulus is not able to modulate attention, will R and S be independent conditional on T.

The characteristic time constraints of the paradigm allow us to interpret Figure 4.2 causally, provided that we additionally assume causal sufficiency,

i.e., that there are no non-measured common causes of the variables in the DAG.

The stimulus could affect response through other ways which are not modulation of covert attention, such as visual processing and memory. The model and results, however, would still be valid, as long as the assumptions in remark 1 remained true. The only difference would be the interpretation of A, that would no longer be identified as a shift in covert attention. Nonetheless, the proper interpretation of A goes beyond the scope of this work and it will not be further explored.

We will now characterize the behavioral response R. Possible behavioral outcomes obtained from the task are detection rate and latency, and both should be affected by attention [Posner et al., 1980]. However, as we will use classification methods for the analysis of statistical dependencies, we need a behavioral outcome R which is categorical and reasonably balanced conditional on S and T, that is, such that the categories have approximately the same number of trials. Following this rationale, we picked as positive response trials, the trials in which the target orientation change was detected and the latency was less than a threshold τ, individually chosen for each participant. The remaining trials were classified as negative response. We refer to this R as balanced response.

The threshold τ was estimated as the most likely detection latency value plus a tolerance. We first estimated the distribution of the detection latencies using gaussian smoothing kernels [Bowman and Azzalini, 1997], and then computed δ_{max} as the latency corresponding to the peak of the distribution. The threshold was defined as $\tau = \delta_{max} + \epsilon$, where $\epsilon \in [50, 100]$ ms was a tolerance parameter, chosen to minimize the difference in the number of trials of the two categories.

For the behavioral analysis, we therefore divided the trials into two conditions, namely, valid and invalid cue, based on whether the attention cue appeared to the same or to the opposite side of the orientation change. We wanted to test if attention would consistently modulate all three of the behavioral features, namely, detection rate (DR), latency (DL) and balanced response (R). Therefore, we employed permutation tests for the null hypotheses,

1. $\mathrm{H0_{DR}}$: DR(valid cue) < DR(invalid cue)
2. $\mathrm{H0_{DL}}$: $\mathbb{E}$[DL|valid cue] > $\mathbb{E}$[DL|invalid cue]
3. $\mathrm{H0_{R}}$: $\mathbb{E}$[R|valid cue] < $\mathbb{E}$[R|invalid cue].

We rejected the null that stimulus does not affect response whenever $p < 0.05$.

4.3 Experimental Data

EEG activity during the trial was recorded at 500 Hz, using actiCAP active electrodes and four BrainAmp DC amplifiers (BrainProducts, Gilching, Germany) with 124 channels. All electrodes were appropriately placed as required for the extended 10-20 system. During the recoding, the electrodes were referenced to the left mastoid. Skin-electrode impedances were kept below 10 kΩ threshold at the start of the experiment. 22 electrodes, in which the impedance exceeded the threshold for any of the participants, were switched off and discarded for analysis in all participants: FpZ, Fp1, Fp2, Fp1p, Fp2p, F1A, F2A, F3A, F4A, F6A, F8A, F6, FT9, FC1, T7, T9, T10, Tp7, Tp9, Tp10, P9, and P10. All discarded electrodes (except FC1) were located either near the ground electrode or in the periphery of the cap, and therefore, were not likely to carry task-relevant information.

For pre-processing, we re-referenced each participant's data to common average, filtered it using a 3rd order high-pass Butterworth filter with 1 Hz cut-off frequency and used principal component analysis to reduce dimensionality from 102 to 64 channels. We performed independent component analysis (ICA) group-wise on the combined data of all participants using the SOBI algorithm [Belouchrani et al., 1993] and rejected independent components (IC) considered non-cortical. An independent component was considered cortical if [Grosse-Wentrup et al., 2011b]: (1) the topography showed a dipolar pattern; (2) the spectrum had $1/f$ shape; (3) eye blinks and movements were not detectable from the time series; (4) no other noise sources, such as 50 Hz line noise, were detectable from the time series.

The brain features of interest were obtained by applying the spatial filters of all cortical ICs to the data of each participant and then computing the log-bandpower in the classical α-range (8-14 Hz) for each IC in ten time windows from the three second attention interval. All windows had a length of one second, with the first one starting 100 ms after beginning of the interval, and the remaining windows shifted by steps of 200 ms. Hence, for each participant and trial, there were ten bandpowered features per IC. We limited our analysis to the α-band in the different time windows, as the dynamics of α-rhythms were found to index visuospatial attention bias and to predict visual target detection in a similar paradigm [Thut et al., 2006].

ICA assumes that each source is observed in all the channels with the same phase, and therefore, might separate phase-locked oscillations into different components [Hyvärinen et al., 2010]. We tested whether this was the case by computing the correlation matrix of α-bandpower averaged across time windows between each component.

4.4 Causal Inference

We used Fisher's Linear Discriminant Analysis (LDA) for classifying the experimental condition using the brain features discussed in the previous section after standardization and outlier rejection. Standardized feature values exceeding three standard deviations were marked as outliers and set to non-informative mean values. Performance accuracy was estimated for each participant using a ten-fold cross-validation and p-values were computed from a binomial test. We always trained our classifiers with balanced data, undersampling the majority class whenever necessary.

On a group-level, we did a single sample Kolmogorov-Smirnov goodness-of-fit test (KS test) to assess the probability that the classification p-values for each IC came from a uniform distribution. We accepted that they did, and that the corresponding feature was not relevant for encoding or decoding, whenever the p-value from the KS test was greater than 0.05.

4.4.1 Discovering Relevant Features

We discussed in section 3.1 the dependence relations that brain features and experimental conditions must satisfy in order to be considered relevant for encoding and decoding. We now discuss the statistical tests which are performed to identify these dependencies.

Features relevant for encoding are the ones in which we can reject the null hypothesis, $\mathrm{H0}_{enc} : X_i \perp\!\!\!\perp C$, for brain feature X_i and experimental condition C. From a classification perspective, a feature X_i is relevant for encoding if a classifier with input X_i can classify C above chance level. Note that this is a sufficient, but not a necessary, condition for rejecting $\mathrm{H0}_{enc}$, since, for example, a linear classifier might be unable to capture nonlinear dependencies in the data. More powerful tests can be obtained by using more general classifiers, as random forests [Breiman, 2001], or nonlinear independence tests, such as the Hilbert-Schmidt independence criterion [Gretton et al., 2008].

Features relevant for decoding are the ones in which we can reject the null hypothesis $\mathrm{H0}_{dec} : X_i \perp\!\!\!\perp C|\mathcal{X}\backslash X_i$, where $\mathcal{X} = \{X_1, \ldots, X_n\}$. Intuitively, X_i is relevant for decoding if it carries information about C which is not contained in the remaining features. If $\mathrm{H0}_{dec}$ is true, $P(C = c|\mathcal{X}) = P(C = c|\mathcal{X} \setminus X_i)$ for every c under the true distribution P. Therefore, the best possible (Bayes) classifier does not change after removing X_i, i.e., $\arg\max_c P(c|\mathcal{X}) = \arg\max_c P(c|\mathcal{X} \setminus X_i)$. Hence, if $\mathrm{H0}_{dec}$ is true, the performance accuracy of the Bayes classifier should be the same whether or not X_i is in the input set, as long as all remaining conditions stay the same. A practical way of testing this hypothesis without changing the dimensionality of the input set is to do

a permutation test. This test consists of permuting X_i with respect to the experimental conditions a large number of times, e.g. 1000, and rejecting $H0_{dec}$ if, and only if, there was an increase in performance accuracy in only a small fraction of times, e.g., less than 5%.

For the stimulus-based analysis, $C = S$, we separated the trials in which the attention cue appeared to the left, or to the right hemifields. For the response-based analysis, $C = R$, we first rejected all trials in which participants pressed a button before stimulus changed, or not at all. We then separated the remaining trials according to whether the balanced response was positive or negative, as described in section 4.2.

4.4.2 Discovering Genuine Causes

In section 3.4, we showed that if response and stimulus were d-separated by a brain feature, there could be no hidden common cause between response and brain feature. In our setting, however, this d-separation could never occur, since part of the dependence between response and stimulus is mediated by the target orientation change, T (see Figure 4.2). Therefore, the conditioning set for d-separation must additionally contain T. In other words, if

1. $R \not\perp\!\!\!\perp S$
2. $R \perp\!\!\!\perp S | T, X_i$,

X_i is a genuine cause of R.

We first select the features $X_i \in \mathcal{X}^{+enc(S,R)} \setminus \mathcal{X}^{+enc(S,R)}_{-dec(S,R)}$ to test for conditional independence. If we do not find any genuine cause in the set, we stop, otherwise we continue with the test in $\mathcal{X}^{+enc(S,R)}_{-dec(S,R)}$.

Testing conditional independence in the case of mixed discrete and continuous data was considered by [Spirtes et al., 2000] one of the open problems in causal inference. To address this, we employ the Kernel-Based Conditional Independence (KCI) Test [Zhang et al., 2011], which uses kernel methods to derive a test statistic for conditional independence and its asymptotic distribution. The KCI test is well-suited to our needs, since it can handle mixed data and is sensitive to non-linear dependencies.

We therefore use the KCI test, after dividing trials into valid and invalid cue trials to condition on T, to test $R \not\perp\!\!\!\perp S | T, X_i$.

Chapter 5

Results

5.1 Participants

Eight healthy adults, four female (mean age 24.5 ± 3.5 years), participated in the experiment. All participants were fluent speakers of English, and had normal or corrected-to-normal vision; two of them were left-handed. None of them, except participant 1, had performed the task before. All of them were payed for their participation and completed a total of 360 trials.

5.2 Behavioral Results

Detection rate was higher on valid cue trials than on invalid cue trials for all participants (group-level average of 87% and 57%, respectively). However, for participants 4, 6, and 8, this difference was not significant, according to a one-sided two-sample permutation test (see table 5.1), with performance difference not greater than 6% for any of them.

Detection was faster on valid than on invalid cue trials in the group (group-level average of latency difference was 250 ms), except in participant 8, for whom detection was 20 ms slower, on average. The permutation test found no significant difference between detection latencies for the two conditions for participants 2, 4, 6, and 8 (see table 5.1), with latency difference never greater than 100 ms.

Positive balanced response, i.e., in which the target was detected and latency was smaller than a threshold, occurred more frequently on valid than on invalid cue trials (group-level average of 61% and 37%). Participant 8 was the only one who had positive response more often (5%) on invalid cue trials. For participants 2, 6, and 8, the difference in positive balanced response was not significant according to a permutation test (see table 5.1).

Table 5.1: p-Values that quantify the probability that, under the null hypothesis of no modulation of attention, detection rate increased, detection latency decreased, and rate of positive response increased in valid versus invalid cue trials for each participant. Values exceeding 0.05 are highlighted in gray.

	Participant							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Rate	0	0	0	0.159	0	0.428	0	0.473
Latency	0	0.104	0	0.177	0	0.140	0.007	0.503
Response	0	0.059	0	0.001	0	0.426	0.039	0.794

Left-right asymmetries in the processing of targets were small relative to the behavioral differences from valid and invalid cue trials. Mean difference of detection latencies between sides across participants was around 20 ms, with values not exceeding 90 m. Detection rate difference between sides across subjects was never greater than 6%, with average 1%.

5.3 Feature Extraction

The topographies of the 10 cortical ICs that we did not reject are shown in Figure 5.1. We note that ICs 2 and 3, located over the occipital cortex, are indicative of visual processing, while ICs 7 and 8 over the sensorimotor cortex, represent sensorimotor processes. In contrast, ICs 4, 5, and 9 are likely generated from the anterior cingulate cortex, the precuneus and the intersection between cuneus and precuneus, respectively, and, therefore may be linked to activity in fronto-parietal executive attention networks. ICs 1, 6, and 10 may contain residual of demixing from other independent components.

We computed, for each participant, the correlation matrix of the α-bandpower, averaged over time windows, in each of the ICs (see Figure 5.2). A pattern of highly correlated independent components is present, with ICs 1, 2, 3, 5, 6, 9, and 10 showing a high correlation in most of the subjects.

5.4 Classification Results

For the stimulus-based analysis, $C = S$, we separated the trials in which the attention cue appeared to the left or to the right hemifields. For each

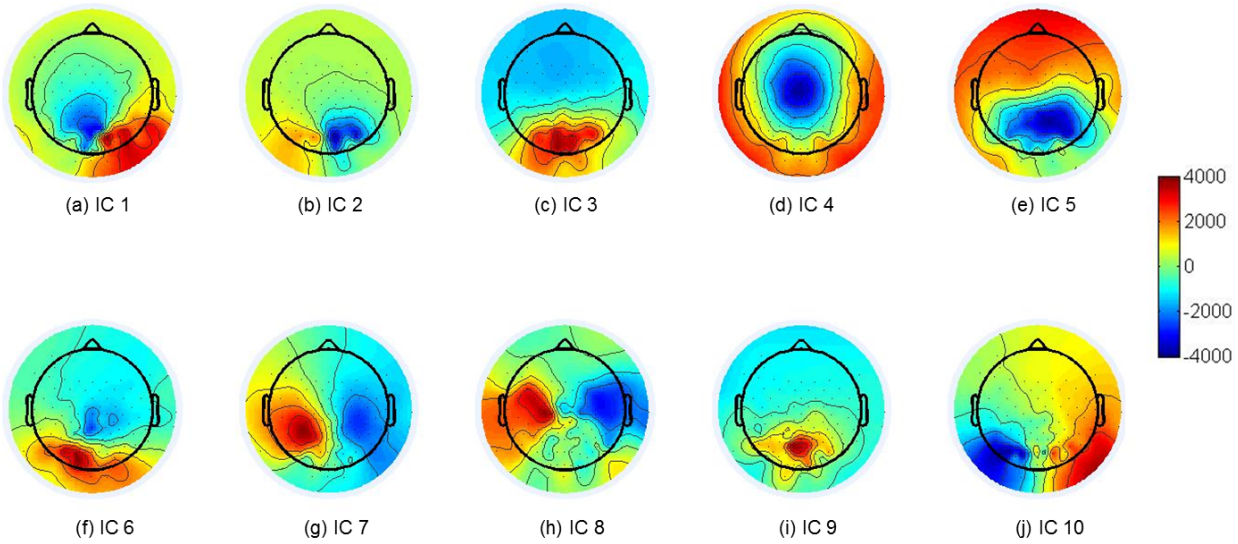

Figure 5.1: Topographies for each cortical IC.

participant, there were 170 trials in each condition. For the response-based analysis, $C = R$, we first rejected all trials in which participants pressed a button before stimulus changed, or not at all. We then separated the remaining trials according to whether the balanced response was positive or negative, as described in section 4.2. The average number of trials in each condition across participants was 140, with minimum 120 trials in one of the participants. We thus obtained one feature vector for each independent component for each of the eight participants, $\{X_1, \ldots, X_{10}\}$, with each X_i being a matrix with ten rows (time windows) and 340 columns (samples) in the stimulus condition, and around 280 columns in the response condition.

Performance accuracy for both stimulus and response classification were above chance level for every participant (see table 5.2). The worst classification performances averaged across stimulus and response were obtained by participants 2 and 6.

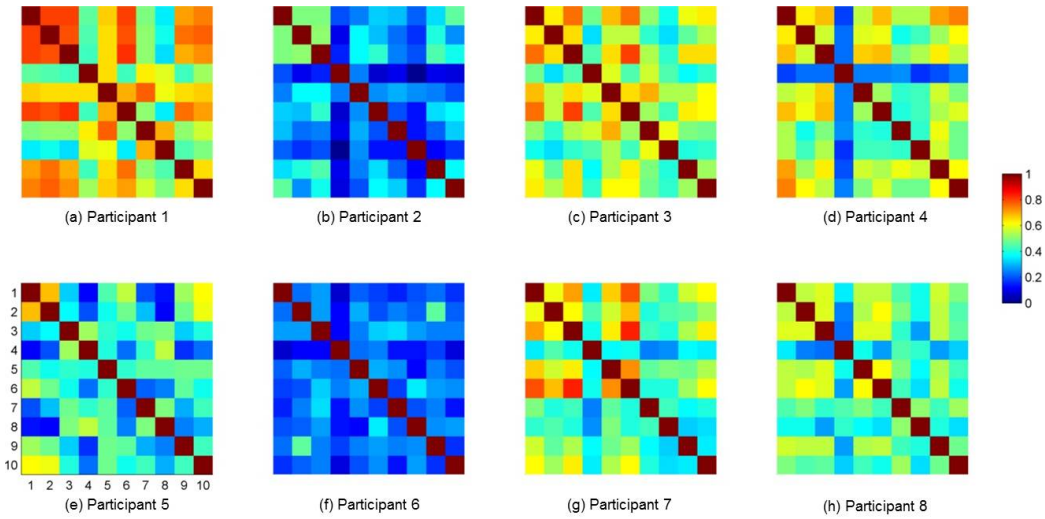

Figure 5.2: Correlation of α-bandpower, averaged over time windows, for each subject.

Table 5.2: Performance Accuracy for Stimulus and Response using all ICs

	Participant							
PE*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Stimulus	76.7650	64.7060	71.1760	71.7650	75.5880	70.5880	82.6470	74.7060
Response	79.3330	70.3570	75.6250	71.4290	68.7500	64.2310	68.8460	73.7500

5.4.1 Encoding Model

All features were considered relevant for encoding of stimulus and response on a group-level analysis (see table 5.3). However, for participants 2, 4, 6 and 8 there were more than one features irrelevant for encoding of stimulus, while for encoding of response, participants 2, 5, and 6 had more than one irrelevant features.

5.4.2 Decoding Model

No feature was considered relevant for decoding of stimulus or response on a single or group-level analysis (see table 5.4).

5.5 Analysis of Results

From the behavioral data, it is possible to conclude that participants 1, 3, 5 and 7 had their behavior modulated by attention, i.e., $S \not\perp\!\!\!\perp R$. Participant 1 showed the best results - likely, due to his/her previous experience with the paradigm. Participants 4 and 8 were extraordinarily proficient at the task, with high detection rate and low latency in both conditions, and were probably performing at a level that was too easy. Participant 2 had a significant difference in detection rate, but not in detection latency and would likely provide better results with more training time.

From the classification results, it follows that all brain features are relevant for encoding of stimulus and response, and that none of the features are relevant for decoding of stimulus or response,

$$X_1, X_2, \ldots, X_{10} \in X^{+enc(S,R)}_{-dec(S,R)}. \tag{5.1}$$

However, this result is not consistent with the causal interpretation rule S6 from table 3.1. In fact, rule S6 implies that every feature would be an indirect effect of stimulus with respect to the other features, and that none of them could be a direct effect. This inconsistency likely stems from the fact that permutation tests for conditional independence are biased towards conditional dependence [Strobl et al., 2008], or from the high correlation values between some α-bandpowered sources shown in Figure 5.2. Indeed, if two features X_i, X_j are highly correlated, removing X_i from the classifier input set will not decrease classification accuracy, as long as X_j remains in the set. We, therefore, cannot use the results from the decoding model in the causal analysis.

Table 5.3: p-Values quantifying the probability of $S \perp\!\!\!\perp X_i$ (Stimulus) and $R \perp\!\!\!\perp X_i$ (Response), respectively, for each subject and feature. Values exceeding 0.05 are highlighted in gray. p-Values of the Kolmogorov-Smirnov goodness-of-fit test, quantifying the probability that the classification p-values from each IC were drawn from a uniform distribution, are denoted by KSp.

		Participant								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	KSp
Stimulus	**IC 1**	0.0098	0.0028	0.0000	0.0028	0.0000	0.0001	0.0000	0.0000	0.0000
	IC 2	0.0098	0.0001	0.0578	0.0171	0.0014	0.0000	0.0000	0.0578	0.0000
	IC 3	0.0001	0.1061	0.0000	0.0098	0.0000	0.1061	0.0000	0.0000	0.0000
	IC 4	0.0000	0.8939	0.0001	0.0053	0.0000	0.8726	0.0000	0.0223	0.0000
	IC 5	0.0073	0.1061	0.0039	0.1274	0.0001	0.8217	0.0001	0.0007	0.0000
	IC 6	0.0001	0.0028	0.0001	0.0053	0.0223	0.0367	0.0000	0.0005	0.0000
	IC 7	0.0000	0.0367	0.0130	0.0578	0.0000	0.0000	0.0000	0.0001	0.0000
	IC 8	0.0000	0.0875	0.0005	0.0000	0.0000	0.3128	0.0005	0.0367	0.0000
	IC 9	0.0001	0.0171	0.0000	0.0003	0.0171	0.0000	0.0288	0.0000	0.0000
	IC 10	0.0010	0.0715	0.0367	0.0000	0.0053	0.0578	0.0000	0.0001	0.0000
Response	**IC 1**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0008	0.0005	0.0012	0.0000
	IC 2	0.0000	0.0024	0.0000	0.0002	0.0811	0.2476	0.0003	0.0040	0.0000
	IC 3	0.0000	0.2954	0.0000	0.0000	0.0001	0.7116	0.0000	0.0001	0.0000
	IC 4	0.0000	0.0000	0.0000	0.0000	0.0010	0.0026	0.0149	0.4743	0.0000
	IC 5	0.0000	0.1281	0.0001	0.0000	0.6925	0.0149	0.0026	0.0002	0.0000
	IC 6	0.0001	0.0134	0.0000	0.0000	0.0000	0.0359	0.0002	0.0000	0.0000
	IC 7	0.0367	0.2186	0.0000	0.0011	0.0000	0.2884	0.0000	0.0000	0.0001
	IC 8	0.0000	0.0001	0.0000	0.0319	0.0021	0.0272	0.0026	0.0000	0.0000
	IC 9	0.0003	0.0035	0.0005	0.0007	0.0000	0.4262	0.0000	0.0118	0.0000
	IC 10	0.0000	0.1047	0.0059	0.0000	0.0007	0.0054	0.0002	0.0005	0.0000

Table 5.4: p-Values quantifying the probability of $S \perp\!\!\!\perp \mathcal{X} \setminus X_i$ (Stimulus) and $R \perp\!\!\!\perp \mathcal{X} \setminus X_i$ (Response), respectively, for each subject and feature. Values exceeding 0.05 are highlighted in gray. p-Values of the Kolmogorov-Smirnov goodness-of-fit test, quantifying the probability that the classification p-values from each IC were drawn from a uniform distribution, are denoted by KSp.

		Participant								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	KSp
Stimulus	**IC 1**	0.2370	0.3630	0.7590	0.0600	0.6140	0.8630	0.7250	0.1070	0.9881
	IC 2	0.4480	0.5470	0.9650	0.3700	0.3860	0.7930	0.7790	0.1420	0.6373
	IC 3	0.3790	0.7420	0.5210	0.1070	0.2300	0.6430	0.8190	0.1860	0.9165
	IC 4	0.4260	0.7460	0.9450	0.2720	0.0920	0.6470	0.6810	0.1900	0.9844
	IC 5	0.5160	0.5060	0.7810	0.2570	0.4280	0.3150	0.9840	0.1310	0.6912
	IC 6	0.4530	0.8490	0.8100	0.2690	0.2390	0.8990	0.6570	0.3360	0.6666
	IC 7	0.3660	0.8520	0.4950	0.6030	0.3630	0.5800	0.7640	0.2570	0.5803
	IC 8	0.1590	0.4990	0.4260	0.5250	0.4170	0.9670	0.8920	0.3070	0.7352
	IC 9	0.2320	0.7600	0.8650	0.3520	0.1810	0.8210	0.4720	0.0860	0.9773
	IC 10	0.2370	0.3630	0.7590	0.0600	0.6140	0.8630	0.7250	0.1070	0.9881
Response	**IC 1**	0.3060	0.8480	0.2500	0.8730	0.6100	0.2680	0.6970	0.9460	0.6134
	IC 2	0.2650	0.6430	0.8430	0.9350	0.8600	0.6560	0.5750	0.8040	0.0536
	IC 3	0.2330	0.8230	0.7150	0.9030	0.7390	0.5000	0.7070	0.8580	0.0478
	IC 4	0.1500	0.8840	0.5110	0.7560	0.6360	0.3130	0.5780	0.7290	0.5616
	IC 5	0.3490	0.8610	0.6990	0.3590	0.4110	0.4800	0.7560	0.5490	0.2242
	IC 6	0.1870	0.6920	0.8480	0.8110	0.3710	0.3000	0.6750	0.3760	0.8971
	IC 7	0.2090	0.8680	0.1350	0.8100	0.6600	0.5560	0.5720	0.6160	0.3664
	IC 8	0.2330	0.5490	0.3390	0.8700	0.6490	0.3410	0.6350	0.5200	0.6961
	IC 9	0.0630	0.9730	0.6230	0.9040	0.5480	0.6230	0.4010	0.6060	0.3982
	IC10	0.3220	0.9430	0.6150	0.8650	0.8570	0.5650	0.7990	0.3530	0.3078

5.6 Causal Analysis

Although our behavioral model from section 4.2 includes the target orientation change T as a link between S and R, for this section, we consider that R is conditioned on T, making the brain the only possible medium of information transfer between S and R. Having identified the sets of relevant features for encoding of stimulus and response, we can apply the interpretation rules from table 3.1. From rule S1, it follows that S is causal to $\{X_1, \ldots, X_{10}\}$, while from rule R1, each X_i is either a cause for S or there is a non-observed common cause to X_i and R. For participants 1, 3, 5, and 7, we additionally have $S \not\perp\!\!\!\perp R$, which, along with randomization of S, implies that there is at least one directed path from S to R, observed or not.

Although enhancement of target detection was evident in only half of the participants, it was possible to classify stimulus and response above chance in all of them. If we accept the null $S \perp\!\!\!\perp R$ in participants 2, 4, 6 and 8, since all features were relevant for encoding of both stimulus and response, we can conclude that there is at least one hidden common cause for each of the brain features and the response, which is not modulated by stimulus. Possible candidates for this common cause would be EEG correlates of the alerting attentional network [Raz and Buhle, 2006], of arousal, or of tiredness, which would modulate detection of an upcoming target, regardless of the cued side.

Assuming that the causal structure among brain networks did not change among participants, there is evidence that these hidden common causes would also be present in participants 1, 3, 5, and 7. Therefore, the test for genuine causes would not produce any positive result. However, we were not able to test this prediction due to time constraints. We nevertheless provided all the framework necessary for the search of genuine causes in future works.

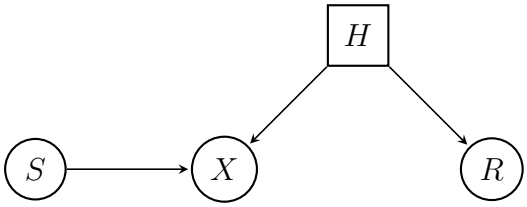

Figure 5.3: Inferred causal structure for participants 2, 4, 6, and 8. All features X were relevant for encoding of stimulus and response and, therefore, were caused by S and either affected R, or there was a hidden common cause to X and R. Since in these participants we could not conclude that $S \not\perp\!\!\!\perp R$, assuming independency between S and R, there can be no active path between stimulus and response, implying the existence of a hidden common cause.

Chapter 6

Discussion

In this work, we explored a framework for detecting causal relations among variables characteristic of a neuroimaging experiment.

In the theoretical part, we provided an introduction to causal Bayesian Networks and reviewed how this framework can be applied to neuroimaging. We extended the results on a set of causal interpretation rules from a previous study [Weichwald et al., 2015] to include an additional experimental configuration, and provided a causal inference test able to detect the absence of hidden common causes between recorded brain activity and behavior.

In the experimental part, we investigated the relevant cortical sources for the encoding of visuospatial attention shifts using classification methods trained on bandpassed cortical independent components from EEG. Covert shifts of visuospatial attention were induced by presenting brief visual cues to either the left or right hemifields of vision, following a paradigm based on Posner's central cue experiment [Posner et al., 1980]. There was a behaviorally measurable enhancement of processing of targets at the cued position in half of the subjects. In the EEG, we found that the α-bandpower from every cortical independent components was individually able to classify the cued side above chance level. Additionally, α-activity from each source could also predict, above chance level, whether a target would be detected under a time limit. None of the features were considered relevant for the decoding of stimulus or response in any of the subjects. A possible explanation for this is the high correlation of the α-activity among components for every subject. Alternatively, the permutation method for decoding, when combined with an LDA classifier, might not be suitable for this analysis, due to bias toward the conditional dependence [Strobl et al., 2008]. Due to time constraints, we were not able to test whether any of the features were genuine causes of behavioral response, but this could be implemented in a straightforward way in a follow-up study.

Our results replicate previous findings of α modulation during voluntary covert shifts of visuospatial attention [Thut et al., 2006, Van Gerven et al., 2009, Kelly et al., 2005]. In comparison to our work, we used independent component analysis to blindly separate sources of cortical activity and found that attention shifts were associated with α activity in all cortical sources over different time windows throughout the trial. This contrasts with previous studies which reported attentional modulation of α mainly in parieto-occipital EEG electrodes [Thut et al., 2006, Rihs et al., 2007]. By using sliding time windows, we are likely more sensitive to dynamic changes in activity over the different networks, which might be related to the orienting of attention [Posner and Boies, 1971]. We obtained a mean classification rate across participants of 73.5% for stimulus and 71.5%, comparable to previous studies [Kelly et al., 2005, Van Gerven et al., 2009], without requiring extensive previous training, individual feature selection, or computationally demanding classifiers.

6.1 Assumptions and Limitations

In section 4.2, we assume that each participant has the same probability for shifting attention to both sides and that attention affects response in the same way, regardless of the side. The behavioral results show that there is little difference of performance in the two sides for each participant, when compared to the behavioral enhancement due to attention. Therefore, there is evidence supporting this assumption in our paradigm.

Causal inference in the framework of Causal Bayesian Networks (CBN) rests on the untestable assumptions of faithfulness and causal markov condition (CMC). Although reasonable, there are theoretical and experimental situations in which these conditions do not hold. For example, in certain quantum systems, the CMC seems to be violated [Elby, 1992], while a sequence of faithful distributions can converge to an unfaithful one [Robins et al., 2003]. Nevertheless, in most real world applications, both conditions are likely satisfied [Spirtes et al., 2000].

Additional limitations to causal inference using CBNs have a statistical nature. CBN methods are based in conditional-independence tests, which might produce different results if there is more data, or if different tests, with more power, with different bias, or sensitive to non-linear dependence in the data, are used. Additionally, one must interpret the lack of evidence against independence as evidence in favor of it.

Although one must be careful about the assumptions and limitations of causal inference, one must accept them in order to make causal claims.

6.2 Future Directions

A fundamental challenge in causal inference within neuroimaging is the lack of an objective ground truth against which it is possible to test whether a method performs better or worse than any other. For those of us who are interested in characterizing the neural mechanisms underlying cognition, the best causal inference method is the one which better consistently predicts the results of controlled interventions on the brain features. Therefore, it is not yet possible to refrain from interventional experiments to infer causation.

Validation of causal methods in neuroimaging could be theoretically achieved with brain-computer interfaces (BCI). In the BCI setting, an intervention could be seen as an instruction to a participant to self-regulate a particular brain feature. A good causal inference method would be able to predict which brain features are likely to change the distribution of a behavioral response, when up or downregulated, and which are not. In our paradigm, for example, we could ask the subject to downregulate α-bandpower of specific independent components, and test whether the mean positive response rate becomes significantly larger than without this intervention. Efficient causal inference methods would be specially useful for the development of new therapeutic BCI, such as for stroke rehabilitation [Grosse-Wentrup et al., 2011a].

It can be argued, however, that BCI studies do not contain pearlian interventions [Pearl, 2000], unless it is confirmed that self-regulation will not affect more than one brain features jointly. We suggest that a more appropriate framework for validating causal models would be invasive multiple single-cell recordings, in which the consequences of interventions, such as lesions, pharmacological inhibitors, and optogenetics, can be measured in a straightforward manner. Causal inference in this scenario could guide the design of new insightful experiments for better understanding how the brain causes perception and cognition.

In the end, the importance of Causal Inference in Neuroimaging will be demonstrated by the scientific progress it enables.

Bibliography

[Bahramisharif et al., 2010] Bahramisharif, A., Van Gerven, M., Heskes, T., and Jensen, O. (2010). Covert attention allows for continuous control of brain–computer interfaces. *European Journal of Neuroscience*, 31(8):1501–1508.

[Belouchrani et al., 1993] Belouchrani, A., Abed-meraim, K., Cardoso, J. F., and Moulines, E. (1993). Second order blind separation of temporally correlated sources.

[Bowman and Azzalini, 1997] Bowman, A. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis*. Number 18 in Oxford statistical science series. Clarendon Press, Oxford.

[Brainard, 1997] Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4):433–436.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

[Cohen and Maunsell, 2010] Cohen, M. R. and Maunsell, J. H. (2010). A neuronal population measure of attention predicts behavioral performance on individual trials. *J. Neurosci.*, 30(45):15241–15253.

[Dawid, 2010] Dawid, A. P. (2010). Beware of the dag! In Guyon, I., Janzing, D., and Schölkopf, B., editors, *NIPS Causality: Objectives and Assessment*, volume 6 of *JMLR Proceedings*, pages 59–86. JMLR.org.

[Dudley, 2002] Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press, 2nd edition.

[Elby, 1992] Elby, A. (1992). Should We Explain the EPR Correlations Causally? *Philosophy of Science*, 59(1):16–25.

[Gretton et al., 2008] Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2008). A kernel statistical test of independence. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in neural information processing systems 20*, pages 585–592, Red Hook, NY, USA. Max-Planck-Gesellschaft, Curran.

[Grosse-Wentrup et al., 2011a] Grosse-Wentrup, M., Mattia, D., and Oweiss, K. (2011a). Using brain-computer interfaces to induce neural plasticity and restore function. *Journal of Neural Engineering*, 8(2):1–5.

[Grosse-Wentrup et al., 2011b] Grosse-Wentrup, M., Schölkopf, B., and Hill, J. (2011b). Causal influence of gamma oscillations on the sensorimotor rhythm. *NeuroImage*, 56(2):837 – 842. Multivariate Decoding and Brain Reading.

[Hyvärinen et al., 2010] Hyvärinen, A., Ramkumar, P., Parkkonen, L., and Hari, R. (2010). Independent component analysis of short-time fourier transforms for spontaneous eeg/meg analysis. *NeuroImage*, 49(1):257–271.

[Kelly et al., 2005] Kelly, S., Lalor, E., Reilly, R., and Foxe, J. (2005). Independent brain computer interface control using visual spatial attention-dependent modulations of parieto-occipital alpha. In *Neural Engineering, 2005. Conference Proceedings. 2nd International IEEE EMBS Conference on*, pages 667–670.

[Kleiner et al., 2007] Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., and Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, 36(14):1.

[Klimesch et al., 1998] Klimesch, W., Doppelmayr, M., Russegger, H., Pachinger, T., and Schwaiger, J. (1998). Induced alpha band power changes in the human eeg and attention. *Neuroscience Letters*, 244(2):73 – 76.

[Naselaris et al., 2011] Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fmri. *NeuroImage*, 56(2):400 – 410. Multivariate Decoding and Brain Reading.

[Pearl, 1985] Pearl, J. (1985). A constraint-propagation approach to probabilistic reasoning. In Kanal, L. N. and Lemmer, J. F., editors, *UAI*. Elsevier.

[Pearl, 2000] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA.

[Pearl and Paz, 1985] Pearl, J. and Paz, A. (1985). *Graphoids: A graph-based logic for reasoning about relevance relations*. University of California (Los Angeles). Computer Science Department.

[Posner and Boies, 1971] Posner, M. I. and Boies, S. J. (1971). Components of attention. *Psychological review*, 78(5):391.

[Posner et al., 1980] Posner, M. I., Snyder, C. R., and Davidson, B. J. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, 32(1):3–25.

[Raz, 2004] Raz, A. (2004). Anatomy of attentional networks. *The Anatomical Record Part B: The New Anatomist*, 281B(1):21–36.

[Raz and Buhle, 2006] Raz, A. and Buhle, J. (2006). Typologies of attentional networks. *Nature Reviews Neuroscience*, 7(5):367–379.

[Reichenbach, 1956] Reichenbach, H. (1956). *The direction of time*. University of Los Angeles Press, Berkeley.

[Rihs et al., 2007] Rihs, T. A., Michel, C. M., and Thut, G. (2007). Mechanisms of selective inhibition in visual spatial attention are indexed by α-band eeg synchronization. *European Journal of Neuroscience*, 25(2):603–610.

[Robins et al., 2003] Robins, J. M., Scheines, R., Spirtes, P., and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, 90(3):491–515.

[Sauseng et al., 2005] Sauseng, P., Klimesch, W., Stadler, W., Schabus, M., Doppelmayr, M., Hanslmayr, S., Gruber, W. R., and Birbaumer, N. (2005). A shift of visual spatial attention is selectively associated with human eeg alpha activity. *European Journal of Neuroscience*, 22(11):2917–2926.

[Spirtes et al., 2000] Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press, 2nd edition.

[Strobl et al., 2008] Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9.

[Thut et al., 2006] Thut, G., Nietzel, A., Brandt, S. A., and Pascual-Leone, A. (2006). α-band electroencephalographic activity over occipital cortex

indexes visuospatial attention bias and predicts visual target detection. *The Journal of Neuroscience*, 26(37):9494–9502.

[Van Gerven et al., 2009] Van Gerven, M., Bahramisharif, A., Heskes, T., and Jensen, O. (2009). Selecting features for bci control based on a covert spatial attention paradigm. *Neural Networks*, 22(9):1271–1277.

[Weichwald et al., 2015] Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., and Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59.

[Zhang et al., 2011] Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In Cozman, F. G. and Pfeffer, A., editors, *UAI*, pages 804–813. AUAI Press.