

Causality in cognitive neuroscience: concepts, challenges, and distributional robustness

Sebastian Weichwald & Jonas Peters

July 2020

While probabilistic models describe the dependence structure between observed variables, causal models go one step further: they predict, for example, how cognitive functions are affected by external interventions that perturb neuronal activity. In this review and perspective article, we introduce the concept of causality in the context of cognitive neuroscience and review existing methods for inferring causal relationships from data. Causal inference is an ambitious task that is particularly challenging in cognitive neuroscience. We discuss two difficulties in more detail: the scarcity of interventional data and the challenge of finding the right variables. We argue for distributional robustness as a guiding principle to tackle these problems. Robustness (or invariance) is a fundamental principle underlying causal methodology. A causal model of a target variable generalises across environments or subjects as long as these environments leave the causal mechanisms intact. Consequently, if a candidate model does not generalise, then either it does not consist of the target variable's causes or the underlying variables do not represent the correct granularity of the problem. In this sense, assessing generalisability may be useful when defining relevant variables and can be used to partially compensate for the lack of interventional data.

1 Introduction

Cognitive neuroscience aims to describe and understand the neuronal underpinnings of cognitive functions such as perception, attention, or learning. The objective is to characterise brain activity and cognitive functions, and to relate one to the other. The submission guidelines for the Journal of Cognitive Neuroscience, for example, state: “The Journal will not publish research reports that bear solely on descriptions of function without addressing the underlying brain events, or that deal solely with descriptions of neurophysiology or neuroanatomy without regard to function.” We think that understanding this relation

requires us to relate brain events and cognitive function in terms of the cause-effect relationships that govern their interplay. A causal model could, for example, describe how cognitive functions are affected by external interventions that perturb neuronal activity (cf. Section 2.1). Reid et al. (2019) argue that “the ultimate phenomenon of theoretical interest in all FC [functional connectivity] research is understanding the causal interaction among neural entities”.

Causal inference in cognitive neuroscience is of great importance and perplexity. This motivates our discussion of two pivotal challenges. First, the scarcity of interventional data is problematic as several causal models may be equally compatible with the observed data while making conflicting predictions only about the effects of interventions (cf. Section 3.1). Second, the ability to understand how neuronal activity gives rise to cognition depends on finding the right variables to represent the neuronal activity (cf. Section 3.2). Our starting point is the well-known observation that causal models of a target (or response) variable are distributionally robust and thus generalise across environments, subjects, and interventional shifts (Haavelmo, 1944; Aldrich, 1989; Pearl, 2009). Models that do not generalise are either based upon the wrong variables that do not represent causal entities or include variables that are not causes of the target variable. We thus propose to pursue robust (or invariant) models. That way, distributional robustness may serve as a guiding principle towards a causal understanding of cognitive function and may help us tackle both challenges mentioned above.

1.1 Running examples

We consider the following simplified examples. Assume that the consumption of alcohol affects reaction times in a cognitive task. In a randomised controlled trial we find that drinking alcoholic (versus non-alcoholic) beer results in slowed reaction times hereinafter. Therefore, we may write ‘alcohol \rightarrow reaction time’ and call alcohol a cause of reaction time and reaction time an effect of alcohol. Intervening on the cause results in a change in the distribution of the effect. In our example, prohibiting the consumption of any alcoholic beers results in faster reaction times.

In cognitive neuroscience one may wish to describe how the neuronal activity is altered upon beer consumption and how this change in turn affects the reaction time. For this, we additionally require a measurement of neuronal activity, say a functional magnetic resonance imaging (fMRI) scan and voxel-wise blood-oxygen-level dependent (BOLD) signals, that can serve as explanans in a description of the phenomenon ‘alcohol \rightarrow neuronal activity \rightarrow reaction time’. We distinguish the following two scenarios:

Running Example A, illustrated in Figure 1a. A so-called treatment or stimulus variable T (say, consumption of alcohol) affects neuronal activity as measured by a d -dimensional feature vector $\mathbf{X} = [X_1, \dots, X_d]^\top$ and the target variable Y reflects a cognitive function (say, reaction time). We may concisely write $T \rightarrow \mathbf{X} \rightarrow Y$ for a treatment that affects neuronal activity which in turn maintains a cognitive function (this is analogous to the ‘stimulus \rightarrow brain activity \rightarrow response’ set-up considered in Weichwald, Schölkopf, Ball & Grosse-Wentrup, 2014; Weichwald et al., 2015).

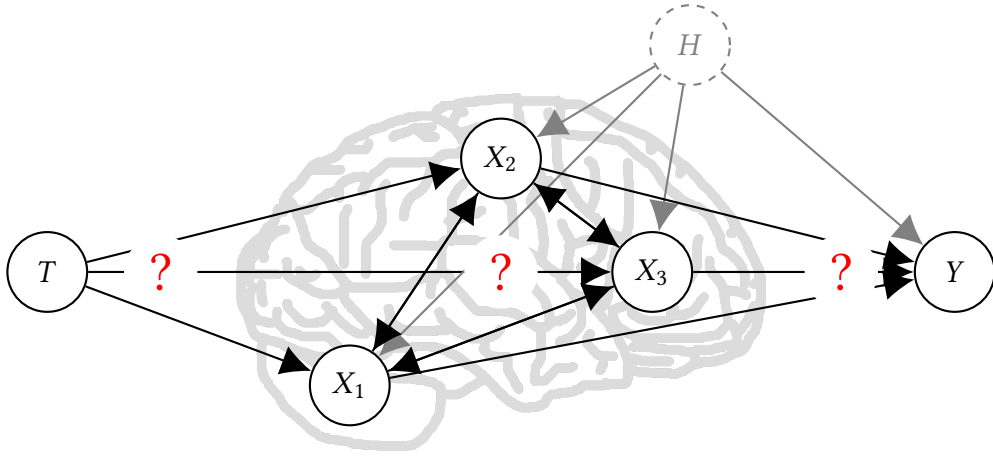
Running Example B, illustrated in Figure 1b. We may wish to describe how neuronal entities cause one another and hence designate one such entity as the target variable Y . In this example, we consider a target variable corresponding to a specific brain signal or region instead of a behavioural or cognitive response.

1.2 Existing work on causality in the context of cognitive neuroscience

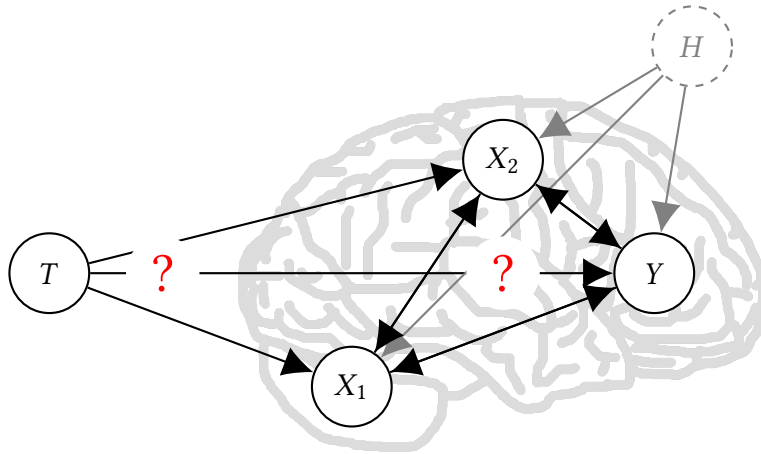
Several methods such as Granger causality or constraint-based methods have been applied to the problem of inferring causality from cognitive neuroscience data. We describe these methods in Section 2.3. In addition, there are ongoing conceptual debates that revolve around the principle of causality in cognitive neuroscience, some of which we now mention. Mehler and Kording (2018) raise concerns about the “lure of causal statements” and expound the problem of confounding when interpreting functional connectivity. Confounders are similarly problematic for multi-voxel pattern analyses (Todd, Nystrom & Cohen, 2013; Woolgar, Golland & Bode, 2014). The causal interpretation of encoding and decoding (forward and backward, univariate and multivariate) models has received much attention as they are common in the analysis of neuroimaging data: Davis et al. (2014) examine the differences between the model types, Haufe et al. (2014) point out that the weights of linear backward models may be misleading, and Weichwald et al. (2015) extend the latter argument to non-linear models and clarify which causal interpretations are warranted from either model type. Feature relevance in mass-univariate and multivariate models can be linked to marginal and conditional dependence statements that yield an enriched causal interpretation when both are combined (Weichwald et al., 2015); this consideration yields refined results in neuroimaging analyses (Huth et al., 2016; Bach, Symmonds, Barnes & Dolan, 2017; Varoquaux et al., 2018) and explains improved functional connectivity results when combining bivariate and partial linear dependence measures (Sanchez-Romero & Cole, 2020). Problems such as indirect measurements and varying temporal delays complicate causal Bayesian network approaches for fMRI (Ramsey et al., 2010; Mumford & Ramsey, 2014). Smith et al. (2011) present a simulation study evaluating several methods for estimating brain networks from fMRI data and demonstrate that identifying the direction of network links is difficult. The discourse on how to leverage connectivity analyses to understand mechanisms in brain networks is ongoing (Valdes-Sosa, Roebroeck, Daunizeau & Friston, 2011; Waldorp, Christoffels & van de Ven, 2011; Smith, 2012; Mill, Ito & Cole, 2017). Many of the above problems and findings are related to the two key challenges that we discuss in Section 3.

1.3 Structure of this work

We begin Section 2 by formally introducing causal concepts. In Section 2.1, we outline why we believe there is a need for causal models in cognitive neuroscience by considering what types of questions could be answered by an OraCLe Modelling (OCM) approach. We discuss the problem of models that are observationally equivalent yet make conflicting predictions about the effects of interventions in Section 2.2. In Section 2.3, we review different causal discovery methods and their underlying assumptions. We focus on two challenges for causality



(a) Running Example A



(b) Running Example B

Figure 1: Illustration of two scenarios in cognitive neuroscience where we seek a causal explanation focusing on a target variable Y that either resembles (a) a cognitive function, or (b) a neuronal entity. The variables T , $\mathbf{X} = [X_1, \dots, X_d]^\top$, and H represent treatment, measurements of neuronal activity, and an unobserved variable, respectively.

in cognitive neuroscience that are expounded in Section 3: (1) the scarcity of interventional data and (2) the challenge of finding the right variables. In Section 4, we argue that one should seek distributionally robust variable representations and models to tackle these challenges. Most of our arguments in this work are presented in an i.i.d. setting and we briefly discuss the implications for time-dependent data in Section 4.5. We conclude in Section 5 and outline ideas that we regard as promising for future research.

2 Causal models and causal discovery

In contrast to classical probabilistic models, causal models induce not only an observational distribution but also a set of so-called interventional distributions. That is, they predict how a system reacts under interventions. We present an introduction to causal models that is based on pioneer work by Pearl (2009) and Spirtes, Glymour and Scheines (2001). Our exposition is inspired by Weichwald (2019, Chapter 2), which provides more introductory intuition into causal models viewed as structured sets of interventional distributions. For both simplicity and focus of exposition, we omit a discussion of counterfactual reasoning and other akin causality frameworks such as the potential outcomes formulation of causality (Imbens & Rubin, 2015). We phrase this article within the framework and terminology of Structural Causal Models (SCMs) (Bollen, 1989; Pearl, 2009).

An SCM over variables $\mathbf{Z} = [Z_1, \dots, Z_d]^\top$ consists of

structural equations that relate each variable Z_k to its parents $\mathbf{PA}(Z_k) \subseteq \{Z_1, \dots, Z_d\}$ and a noise variable N_k via a function f_k such that $Z_k := f_k(\mathbf{PA}(Z_k), N_k)$, and a

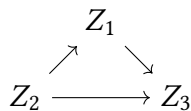
noise distribution \mathbf{P}_N of the noise variables $\mathbf{N} = [N_1, \dots, N_d]^\top$.

We associate each SCM with a directed causal graph where the nodes correspond to the variables Z_1, \dots, Z_d and we draw an edge from Z_i to Z_j whenever Z_i appears on the right hand side of the equation $Z_j := f_j(\mathbf{PA}(Z_j), N_j)$. That is, if $Z_i \in \mathbf{PA}(Z_j)$ the graph contains the edge $Z_i \rightarrow Z_j$. Here, we assume that this graph is acyclic. The structural equations and noise distributions together induce the observational distribution \mathbf{P}_Z of Z_1, \dots, Z_d as simultaneous solution to the equations. (Bongers, Peters, Schölkopf and Mooij (2018) formally define SCMs when the graph includes cycles.)

The following is an example of a linear Gaussian SCM:

$$\begin{aligned} Z_1 &:= f_1(\mathbf{PA}(Z_1), N_1) = Z_2 + N_1 \\ Z_2 &:= f_2(\mathbf{PA}(Z_2), N_2) = N_2 \\ Z_3 &:= f_3(\mathbf{PA}(Z_3), N_3) = Z_1 + 5 \cdot Z_2 + N_3 \end{aligned}$$

with mutually independent standard-normal noise variables N_1, N_2, N_3 . The corresponding graph is



and the SCM induces the observational distribution \mathbf{P}_Z , which is the multivariate Gaussian distribution

$$\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} \sim \mathbf{P}_Z = \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 & 7 \\ 1 & 1 & 6 \\ 7 & 6 & 38 \end{pmatrix} \right). \quad (1)$$

In addition to the observational distribution, an SCM induces interventional distributions. Each intervention denotes a scenario in which we fix a certain subset of the variables to a certain value. For example, the intervention $\text{do}(Z_2 := 0, Z_3 := 5)$ denotes the scenario where we force Z_2 and Z_3 to take on the values 0 and 5, respectively. The interventional distributions are obtained by (a) replacing the structural equations of the intervened upon variables by the new assignment, and (b) considering the distribution induced by the thus obtained new set of structural equations. For example, the distribution under intervention $\text{do}(Z_1 := a)$ for $a \in \mathbb{R}$, denoted by $\mathbf{P}_Z^{\text{do}(Z_1:=a)}$, is obtained by changing the equation $Z_1 := f_1(\mathbf{PA}(Z_1), N_1)$ to $Z_1 := a$. In the above example, we find

$$\mathbf{P}_Z^{\text{do}(Z_1:=a)} = \mathcal{N} \left(\begin{pmatrix} a \\ 0 \\ a \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 5 \\ 0 & 5 & 26 \end{pmatrix} \right),$$

where $X \sim \mathcal{N}(a, 0)$ if and only if $\mathbf{P}(X = a) = 1$. Analogously, for $b \in \mathbb{R}$ and intervention on Z_2 we have

$$\mathbf{P}_Z^{\text{do}(Z_2:=b)} = \mathcal{N} \left(\begin{pmatrix} b \\ b \\ 6 \cdot b \end{pmatrix}, \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix} \right).$$

The distribution of Z_1 differs between the observational distribution and the interventional distribution, that is, $\mathbf{P}_{Z_1} \neq \mathbf{P}_{Z_1}^{\text{do}(Z_2:=b)}$. We call a variable X an (indirect) cause of a variable Y if there exists an intervention on X under which the distribution of Y is different from its distribution in the observational setting. Thus, Z_2 is a cause of Z_1 . The edge $Z_2 \rightarrow Z_1$ in the above causal graph reflects this cause-effect relationship. In contrast, Z_2 remains standard-normally distributed under all interventions $\text{do}(Z_1 := a)$ on Z_1 . Because the distribution of Z_2 remains unchanged under any intervention on Z_1 , Z_1 is not a cause of Z_2 .

In general, interventional distributions do not coincide with the corresponding conditional distributions. In our example we have $\mathbf{P}_{Z_1|Z_1=a} \neq \mathbf{P}_Z^{\text{do}(Z_1:=a)}$ while $\mathbf{P}_{Z_1|Z_2=b} = \mathbf{P}_Z^{\text{do}(Z_2:=b)}$. We further have that the conditional distribution $\mathbf{P}_{Z_3|Z_2,Z_1}$ of Z_3 given its parents Z_1 and Z_2 is invariant under interventions on variables other than Z_3 . We call a model of Z_3 based on Z_1, Z_2 invariant (cf. Section 4.1).

We have demonstrated how an SCM induces a set of observational and interventional distributions. The interventional distributions predict observations of the system upon intervening on some of its variables. As such, a causal model holds additional content compared to a common probabilistic model that amounts to one distribution to describe future observations of the same unchanged system. Sometimes we are only interested in modelling certain interventions or cannot perform others as there may be no well-defined corresponding real-world implementation. For example, we cannot intervene on a person's gender. In

these cases it may be helpful to explicitly restrict ourselves to a set of interventions of interest. Furthermore, the choice of an intervention set puts constraints on the granularity of the model (cf. Section 3.2 and Rubenstein & Weichwald et al., 2017).

2.1 When are causal models important?

We do not always need causal models to answer our research question. For some scientific questions it suffices to consider probabilistic, that is, observational models. For example, if we wish to develop an algorithm for early diagnosis of Alzheimer’s disease from brain scans, we need to model the conditional distribution of Alzheimer’s disease given brain activity. Since this can be computed from the joint distribution, a probabilistic model suffices. If, however, we wish to obtain an understanding that allows us to optimally prevent progression of Alzheimer’s disease by, for example, cognitive training or brain stimulation, we are in fact interested in a causal understanding of the Alzheimer’s disease and require a causal model.

Distinguishing between these types of questions is important as it informs us about the methods we need to employ in order to answer the question at hand. To elaborate upon this distinction, we now discuss scenarios related to our running examples and the relationship between alcohol consumption and reaction time (cf. Section 1.1). Assume we have access to a powerful OraCle Modelling (OCM) machinery that is unaffected by statistical problems such as model misspecification, multiple-testing, or small sample sizes. By asking ourselves, what queries must be answered by OCM for us to ‘understand’ the cognitive function, the difference between causal and non-causal questions becomes apparent.

Assume, firstly, we ran the reaction task experiment with multiple subjects, fed all observations to our OCM machinery, and have Kim visiting our lab today. Since OCM yields us the exact conditional distribution of reaction times $\mathbf{P}_{Y|T=t}$ for Kim having consumed $T = t$ units of alcoholic beer, we may be willing to bet against our colleagues on how Kim will perform in the reaction task experiment they are just about to participate in. No causal model for brain activity is necessary.

Assume, secondly, that we additionally record BOLD responses $\mathbf{X} = [X_1, \dots, X_d]^\top$ at certain locations and times during the reaction task experiment. We can query OCM for the distribution of BOLD signals that we are about to record, that is, $\mathbf{P}_{\mathbf{X}|T=t}$, or the distribution of reaction times given we measure Kim’s BOLD responses $\mathbf{X} = \mathbf{x}$, that is, $\mathbf{P}_{Y|T=t, \mathbf{X}=\mathbf{x}}$. As before, we may bet against our colleagues on how Kim’s BOLD signals will look like in the upcoming reaction task experiment or bet on their reaction time once we observed the BOLD activity $\mathbf{X} = \mathbf{x}$ prior to a reaction cue. Again, no causal model for brain activity is required.

In both of the above situations, we have learned something useful. Given that the data were obtained in an experiment in which alcohol consumption was randomised, we have learned, in the first situation, to predict reaction times after an intervention on alcohol consumption. This may be considered an operational model for alcohol consumption and reaction time. In the second situation, we have learned how the BOLD signal responds to alcohol consumption. Yet, in none of the above situations have we gained understanding of the neuronal underpinnings of the cognitive function and the reaction times. Knowing the conditional distributions $\mathbf{P}_{Y|T=t}$ and $\mathbf{P}_{Y|T=t, \mathbf{X}=\mathbf{x}}$ for any t yields no insight into any of the following questions. Which brain regions maintain fast reaction times? Where in the brain should we release drugs that

excite neuronal activity in order to counterbalance the effect of alcohol? How do we need to update our prediction if we learnt that Kim just took a new drug that lowers blood pressure in the prefrontal cortex? To answer such questions, we require causal understanding.

If we had a causal model, say in form of an SCM, we could address the above questions. An SCM offers an explicit way to model the system under manipulations. Therefore, a causal model can help to answer questions about where to release an excitatory drug. It may enable us to predict whether medication that lowers blood pressure in the prefrontal cortex will affect Kim’s reaction time; in general, this is the case if the corresponding variables appear in the structural equations for Y or any of Y ’s ancestors.

Instead of identifying conditional distributions, one may formulate the problem as a regression task with the aim to learn the conditional mean functions $t \rightarrow \mathbb{E}[\mathbf{X}|T = t]$ and $(t, \mathbf{x}) \rightarrow \mathbb{E}[Y|T = t, \mathbf{X} = \mathbf{x}]$. These functions are then parameterised in terms of t or t and \mathbf{x} . We argue in Section 2.2, point (2), that such parameters do not carry a causal meaning and thus do not help to answer the questions above.

Promoted by slogans such as ‘correlation does not imply causation’ careful and associational language is sometimes used in the presentation of cognitive neuroscience studies. We believe, however, that a clear language that states whether a model should be interpreted causally (that is, as an interventional model) or non-causally (that is, as an observational model) is needed. This will help to clarify both the real world processes the model can be used for and the purported scientific claims.

Furthermore, causal models may generalise better than non-causal models. We expect systematic differences between subjects and between different trials or recording days of the same subject. These different situations, or environments, are presumably not arbitrarily different. If they were, we could not hope to gain any scientific insight from such experiments. The apparent question is, which parts of the model we can expect to generalise between environments. It is well-known that causal models capture one such invariance property, which is implicit in the definition of interventions. An intervention on one variable leaves the assignments of the other variables unaffected. Therefore, the conditional distributions of these other variables, given their parents, are also unaffected by the intervention (Haavelmo, 1944; Aldrich, 1989). Thus, causal models may enable us to formulate more clearly which mechanisms we assume to be invariant between subjects. For example, we may assume that the mechanism how alcohol intake affects brain activity differs between subjects, whereas the mechanism from signals in certain brain regions to reaction time is invariant. We discuss the connection between causality and robustness in Section 4.

2.2 Equivalences of models

Causal models entail strictly more information than observational models. We now introduce the notion of equivalence of models (Pearl, 2009; Peters, Janzing & Schölkopf, 2017; Bongers et al., 2018). This notion allows us to discuss the falsifiability of causal models, which is important when assessing candidate models and their ability to capture cause-effect relationships that govern a cognitive process under investigation.

We call two models observationally equivalent if they induce the same observational distribution. Two models are said to be interventionally equivalent if they induce the same

observational and interventional distributions. As discussed above, for some interventions there may not be a well-defined corresponding experiment in the real world. We therefore also consider interventional equivalence with respect to a restricted set of interventions.

One reason why learning causal models from observational data is difficult is the existence of models that are observationally but not interventionally equivalent. Such models agree in their predictions about the observed system yet disagree in their predictions about the effects of certain interventions. We continue the example from Section 2 and consider the following two SCMs:

$$\begin{array}{ll}
 Z_1 := Z_2 + N_1 & Z_1 := \sqrt{2} \cdot N_1 \\
 Z_2 := N_2 & Z_2 := 1/2 \cdot Z_1 + 1/\sqrt{2} \cdot N_2 \\
 Z_3 := Z_1 + 5 \cdot Z_2 + N_3 & Z_3 := Z_1 + 5 \cdot Z_2 + N_3
 \end{array}$$

where in both cases N_1, N_2, N_3 are mutually independent standard-normal noise variables. The two SCMs are observationally equivalent as they induce the same observational distribution, the one shown in Equation (1). The models are not interventionally equivalent, however, since $\mathbf{P}_{Z_2}^{\text{do}(Z_1:=3)} = \mathcal{N}(0, 1)$ and $\mathbf{P}_{Z_2}^{\text{do}(Z_1:=3)} = \mathcal{N}(3/2, 1/2)$ for the left and right model, respectively. The two models can be told apart when interventions on Z_1 or Z_2 are considered. They are interventionally equivalent with respect to interventions on Z_3 .

The existence of observationally equivalent models that are not interventionally equivalent has several implications. (1) Without assumptions, it is impossible to learn causal structure from observational data. This is not exclusive to causal inference from data and an analogous statement holds true for regression (Györfi, Kohler, Krzyżak & Walk, 2002). The regression problem is solvable only under certain simplicity assumptions, for example, on the smoothness of the regression function, which have been proven useful in real world applications. Similarly, there are several assumptions that can be exploited for causal discovery. We discuss some of these assumptions in Section 2.3. (2) As a consequence, without further restrictive assumptions on the data generating process, the estimated parameters do not carry any causal meaning. For example, given any finite sample from the observational distribution, both of the above SCMs yield exactly the same likelihood. Therefore, the above structures cannot be told apart by a method that employs the maximum likelihood estimation principle. Instead, which SCM and thus which parameters are selected in such a situation may depend on starting values, optimisation technique, or numerical precision. (3) Assume that we are given a probabilistic (observational) model of a data generating process. To falsify it, we may apply a goodness-of-fit test based on an observational sample from that process. An interventional model cannot be falsified based on observational data alone and one has to also take into account the outcome of interventional experiments. This requires that we are in agreement about how to perform the intervention in practice (see also Section 3.2). Interventional data may be crucial in particular for rejecting some of the observationally equivalent models (cf. the example above). The scarcity of interventional data therefore poses a challenge for causality in cognitive neuroscience (cf. Section 3.1).

2.3 Causal discovery

The task of learning a causal model from observational (or a combination of observational and interventional) data is commonly referred to as causal discovery or causal structure learning. We have argued in the preceding section that causal discovery from purely observational data is impossible without any additional assumptions or background knowledge. In this section, we discuss several assumptions that render (parts of) the causal structure identifiable from the observational distribution. In short, assumptions concern how causal links manifest in observable statistical dependences, functional forms of the mechanisms, certain invariances under interventions, or the order of time. We briefly outline how these assumptions can be exploited in algorithms. Depending on the application at hand, one may be interested in learning the full causal structure as represented by its graph or in identifying a local structure such as the causes of a target variable Y . The methods described below cover either of the two cases. We keep the description brief focussing on the main ideas and intuition, while more details can be found in the respective references.

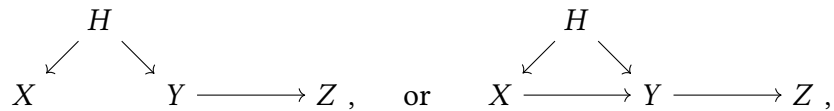
Randomisation. The often called ‘gold standard’ to establishing whether T causes Y is to introduce controlled perturbations, that is, targeted interventions, to a system. Without randomisation, a dependence between T and Y could stem from a confounder between T and Y or from a causal link from Y to T . If T is randomised it is no further governed by the outcome of any other variable or mechanism. Instead, it only depends on the outcome of a randomisation experiment, such as the roll of a die. If we observe that under the randomisation, Y depends on T , say the higher T the higher Y , then there must be a (possibly indirect) causal influence from T to Y . In our running examples, this allows us to conclude that the amount of alcoholic beer consumed causes reaction times (cf. Section 1.1). When falsifying interventional models, it suffices to consider randomised experiments as interventions (Peters et al., 2017, Proposition 6.48). In practice, however, performing randomised experiments is often infeasible due to cost or ethical concerns, or impossible as, for example, we cannot randomise gender nor fully control neuronal activity in the temporal lobe. While it is sometimes argued that the experiment conducted by James Lind in 1747 to identify a treatment for scurvy is among the first randomised controlled trials, the mathematical theory and methodology was popularised by Ronald A. Fisher in the early 20th century (Conniffe, 1991).

Constraint-based methods. Constraint-based methods rely on two assumptions that connect properties of the causal graph with conditional independence statements in the induced distribution. The essence of the first assumption is sometimes described as Reichenbach’s common cause principle (Reichenbach, 1956): If X and Y are dependent, then there must be some cause-effect structure that explains the observed dependence, that is, either X causes Y , or Y causes X , or another unobserved variable H causes both X and Y , or some combination of the aforementioned. This principle is formalised by the Markov condition (see for example Lauritzen, 1996). This assumption is considered to be mild. Any distribution induced by an acyclic SCM satisfies the Markov condition with respect to the corresponding graph (Lauritzen, Dawid, Larsen & Leimer, 1990; Pearl, 2009). The second assumption (often referred to as faithfulness), states that any (conditional) independence between random variables is im-

plied by the graph structure (Spirtes et al., 2001). For example, if two variables are independent, then neither does cause the other nor do they share a common cause. Both assumptions together establish a one-to-one correspondence between conditional independences in the distribution and graphical separation properties between the corresponding nodes.

The back-bone of the constraint-based causal discovery algorithms such as the PC algorithm is to test for marginal and conditional (in)dependences in observed data and to find all graphs that encode the same list of separation statements (Spirtes et al., 2001; Pearl, 2009). This allows us to infer a so-called Markov equivalence class of graphs: all of its members encode the same set of conditional independences. It has been shown that two directed acyclic graphs (assuming that all nodes are observed) are Markov equivalent if and only if they have the same skeleton and v-structures $\rightarrow \circ \leftarrow$ (Verma & Pearl, 1990). Allowing for hidden variables, as done by the FCI algorithm, for example, enlarges the class of equivalent graphs and the output is usually less informative (Spirtes et al., 2001).

The following example further illustrates the idea of a constraint-based search. For simplicity, we assume a linear Gaussian setting, so that (conditional) independence coincides with vanishing (partial) correlation. Say we observe X , Y , and Z . Assume that the partial correlation between X and Z given Y vanishes while none of the other correlations and partial correlations vanish. Under the Markov and faithfulness assumptions there are multiple causal structures that are compatible with those constraints, such as $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \leftarrow Z$, $X \leftarrow Y \rightarrow Z$, or



where H is unobserved. Still, the correlation pattern rules out certain other causal structures. For example, neither $X \rightarrow Y \leftarrow Z$ nor $X \leftarrow H \rightarrow Y \leftarrow Z$ can be the correct graph structure since either case would imply that X and Z are uncorrelated (and $X \perp\!\!\!\perp Z \mid Y$ is not satisfied).

Variants of the above setting were considered in neuroimaging where a randomised experimental stimulus or time-ordering was used to further disambiguate between the remaining possible structures (Grosse-Wentrup, Janzing, Siegel & Schölkopf, 2016; Weichwald, Gretton, Schölkopf & Grosse-Wentrup, 2016a; Weichwald, Grosse-Wentrup & Gretton, 2016b; Mastakouri, Schölkopf & Janzing, 2019). Constraint-based causal inference methodology also clarifies the interpretation of encoding and decoding analyses in neuroimaging and has informed a refined understanding of the neural dynamics of probabilistic reward prediction and an improved functional atlas (Weichwald et al., 2015; Bach et al., 2017; Varoquaux et al., 2018).

Direct applications of this approach in cognitive neuroscience are difficult, not only due to the key challenges discussed in Section 3, but also due to indirect and spatially smeared neuroimaging measurements that effectively spoil conditional independences. In the linear setting, there are recent advances that explicitly tackle the problem of inferring the causal structure between latent variables, say the neuronal entities, based on observations of recorded variables (Silva, Scheine, Glymour & Spirtes, 2006). Further practical challenges include the difficulty of testing for non-parametric conditional independence (Shah & Peters, 2020) and near-faithfulness violations (Uhler, Raskutti, Bühlmann & Yu, 2013).

Score-based methods. Instead of directly exploiting the (conditional) independences to inform our inference about the causal graph structure, score-based methods assess different graph structures by their ability to fit observed data (see for example Chickering, 2002). This approach is motivated by the idea that graph structures that encode the wrong (conditional) independences will also result in bad model fit. Assuming a parametric model class, we can evaluate the log-likelihood of the data and score different candidate graph structures by the Bayesian Information Criterion, for example. The number of possible graph structures to search over grows super-exponentially. That combinatorial difficulty can be dealt with by applying greedy search procedures that usually, however, do not come with finite sample guarantees. Alternatively, Zheng, Dan, Aragam, Ravikumar and Xing (2020) exploit an algebraic characterisation of graph structures to maximise a score over acyclic graphs by solving a continuous optimisation problem. The score-based approach relies on correctly specifying the model class. Furthermore, in the presence of hidden variables, the search space grows even larger and model scoring is complicated by the need to marginalise over those hidden variables (Jabbari, Ramsey, Spirtes & Cooper, 2017).

Restricted structural causal models. Another possibility is to restrict the class of functions in the structural assignments and the noise distributions. Linear non-Gaussian acyclic models (Shimizu, Hoyer, Hyvärinen & Kerminen, 2006), for example, assume that the structural assignments are linear and the noise distributions are non-Gaussian. As for independent component analysis, identifiability of the causal graph follows from the Darmois-Skitovich theorem (Darmois, 1953; Skitovič, 1962). Similar results hold for nonlinear models with additive noise (Hoyer, Janzing, Mooij, Peters & Schölkopf, 2008; Zhang & Hyvärinen, 2009; Peters, Mooij, Janzing & Schölkopf, 2014; Bühlmann, Peters & Ernest, 2014) or linear Gaussian models when the error variances of the different variables are assumed to be equal (Peters & Bühlmann, 2014). The additive noise assumption is a powerful, yet restrictive, assumption that may be violated in practical applications.

Dynamic causal modelling (DCM). We may have prior beliefs about the existence and direction of some of the edges. Incorporating these by careful specification of the priors is an explicit modelling step in DCM (Valdes-Sosa et al., 2011). Given such a prior, we may prefer one model over the other among the two observationally equivalent models presented in Section 2.2, for example. Since the method’s outcome relies on this prior information, any disagreement on the validity of that prior information necessarily yields a discourse about the method’s outcome (Lohmann, Erfurth, Müller & Turner, 2012). Further, a simulation study raised concerns regarding the validity of the model selection procedure in DCM (Friston, Harrison & Penny, 2003; Lohmann et al., 2012; Friston, Daunizeau & Stephan, 2013; Breakspear, 2013; Lohmann, Müller & Turner, 2013).

Granger causality. Granger causality is among the most popular approaches for the analysis of connectivity between time-evolving processes. It exploits the existence of time and the fact that causes precede their effects. Together with its non-linear extensions it has been considered for the analysis of neuroimaging data with applications to electro-encephalography

(EEG) and fMRI data (Marinazzo, Pellicoro & Stramaglia, 2008; Marinazzo, Liao, Chen & Stramaglia, 2011; Stramaglia, Wu, Pellicoro & Marinazzo, 2012; Stramaglia, Cortes & Marinazzo, 2014). The idea is sometimes wrongly described as follows: If including the past of Y_t improves our prediction of X_t compared to a prediction that is only based on the past of X_t alone, then Y Granger-causes X . Granger (1969) himself put forward a more careful definition that includes a reference to all the information in the universe: If the prediction of X_t based on all the information in the universe up to time t is better than the prediction where we use all the information in the universe up to time t apart from the past of Y_t , then Y Granger-causes X . In practice, we may instead resort to a multivariate formulation of Granger causality. If all relevant variables are observed (often referred to as causal sufficiency), there is a close correspondence between Granger causality and the constraint-based approach (Peters et al., 2017, Chapter 10.3.3). Observing all relevant variables, however, is a strong assumption which is most likely violated for data sets in cognitive neuroscience. While Granger causality may be combined with a goodness-of-fit test to at least partially detect the existence of confounders (Peters, Janzing & Schölkopf, 2013), it is commonly applied as a computationally efficient black box approach that always outputs a result. In the presence of instantaneous effects (for example, due to undersampling) or hidden variables, these results may be erroneous (see, for example, Sanchez-Romero et al., 2019).

Inferring causes of a target variable. We now consider a problem that is arguably simpler than inferring the full causal graph: identifying the causes of some target variable of interest. As outlined in the running examples in Section 1.1, we assume that we have observations of the variables T, Y, X_1, \dots, X_d , where Y denotes the target variable. Assume that there is an unknown structural causal model that includes the variables T, Y, X_1, \dots, X_d and that describes the data generating process well. To identify the variables among X_1, \dots, X_d that cause Y , it does not suffice to regress Y on X_1, \dots, X_d . The following example of an SCM shows that a good predictive model for Y is not necessarily a good interventional model for Y . Consider

$$\begin{aligned} X_1 &:= N_1 \\ Y &:= X_1 + N_Y \\ X_2 &:= 10 \cdot Y + N_2 \end{aligned} \quad \begin{array}{c} X_1 \rightarrow Y \rightarrow X_2 \end{array}$$

where N_1, N_2, N_Y are mutually independent standard-normal noise variables. X_2 is a good predictor for Y , but X_2 does not have any causal influence on Y : the distribution of Y is unchanged upon interventions on X_2 .

Recently, causal discovery methods have been proposed that aim to infer the causal parents of Y if we are given data from different environments, that is, from different experimental conditions, repetitions, or different subjects. These methods exploit a distributional robustness property of causal models and are described in Section 4.

Cognitive function versus brain activity as the target variable. When we are interested in inferring direct causes of a target variable Y , it can be useful to include background knowledge. Consider our Running Example A (cf. Section 1.1 and Figure 1a) with reaction

time as the target variable and assume we are interested in inferring which of the variables measuring neuronal activity are causal for the reaction time Y . We have argued in the preceding paragraph that if a variable X_j is predictive of Y , it does not necessarily have to be causal for Y . Assuming, however, that we can exclude that the cognitive function ‘reaction time’ causes brain activity (for example, because of time ordering), we obtain the following simplification: every X_j that is predictive of Y , must be an indirect or direct cause of Y , confounded with Y , or a combination of both. This is different if our target variable is a neuronal entity as in Running Example B (cf. Figure 1b). Here, predictive variables can be either ancestors of Y , confounded with Y , descendants of Y , or some combination of the aforementioned (these statements follow from the Markov condition).

3 Two challenges for causality in cognitive neuroscience

Performing causal inference on measurements of neuronal activity comes with several challenges, many of which have been discussed in the literature (cf. Section 1.2). In the following two subsections we explicate two challenges that we think deserve special attention. In Section 4, we elaborate on how distributional robustness across environments, such as different recording sessions or subjects, can serve as a guiding principle for tackling those challenges.

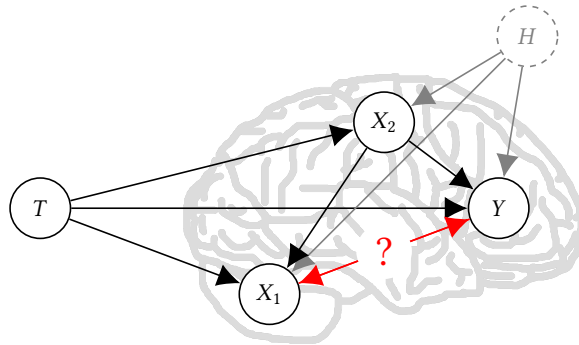
3.1 Challenge 1: The scarcity of targeted interventional data

In Section 2.2 we discussed that different causal models may induce the same observational distribution while they make different predictions about the effects of interventions. That is, observationally equivalent models need not be interventionally equivalent. This implies that some models can only be refuted when we observe the system under interventions which perturb some specific variables in our model. In contrast to broad perturbations of the system, we call targeted interventions those for which the intervention target is known and for which we can list the intervened-upon variables in our model, say “ X_1, X_3, X_8 have been intervened upon.” Even if some targeted interventions are available, there may still be multiple models that are compatible with all observations obtained under those available interventions. In the worst case, a sequence of up to d targeted interventional experiments may be required to distinguish between the possible causal structures over d observables X_1, \dots, X_d when the existence of unobserved variables cannot be excluded while assuming Markovianity, faithfulness, and acyclicity (Eberhardt, 2013). In general, the more interventional scenarios are available to us, the more causal models we can falsify and the further we can narrow down the set of causal models compatible with the data.

Therefore, the scarcity of targeted interventional data is a barrier to causal inference in cognitive neuroscience. Our ability to intervene on neural entities such as the BOLD level or oscillatory bandpower in a brain region is limited and so is our ability to either identify the right causal model from interventional data or to test causal hypotheses that are made in the literature. One promising avenue are non-invasive brain stimulation techniques such as transcranial magnetic or direct/alternating current stimulation which modulate neural activity by creating a field inside the brain (Nitsche et al., 2008; Herrmann, Rach, Neuling &

Strüber, 2013; Bestmann & Walsh, 2017; Kar, Ito, Cole & Krekelberg, 2020). Since the stimulation acts broadly and its neurophysiological effects are not yet fully understood, transcranial stimulation cannot be understood as targeted intervention on some specific neuronal entity in our causal model (Antal & Herrmann, 2016; Vosskuhl, Strüber & Herrmann, 2018). The inter-individual variability in response to stimulation further impedes its direct use for probing causal pathways between brain regions (López-Alonso, Cheeran, Rio-Rodriguez & Fernández-del-Olmo, 2014). Bergmann and Hartwigsen (2020) review the obstacles to inferring causality from non-invasive brain stimulation studies and provide guidelines to attenuate the aforementioned. Invasive stimulation techniques, such as deep brain stimulation relying on electrode implants (Mayberg et al., 2005), may enable temporally and spatially more fine-grained perturbations of neural entities. Dubois et al. (2017) exemplify how to revise causal structures inferred from observational neuroimaging data on a larger cohort through direct stimulation of specific brain regions and concurrent fMRI on a smaller cohort of neurosurgical epilepsy patients. In non-human primates, concurrent optogenetic stimulation with whole-brain fMRI had been used to map the wiring of the medial prefrontal cortex (Liang et al., 2015; Lee et al., 2010). Yet, there are ethical barriers to large-scale invasive brain stimulation studies and it may not be exactly clear how an invasive stimulation corresponds to an intervention on, say, the BOLD response measured in some voxels. We thus believe that targeted interventional data will remain a scarcity due to physical and ethical limits to non-invasive and invasive brain stimulation.

Consider the following variant of our Running Example B (cf. Section 1.1). Assume that (a) the consumption of alcoholic beer T slows neuronal activity in the brain regions X_1 , X_2 , and Y , (b) X_2 is a cause of X_1 , and (c) X_2 is a cause of Y . Here, (a) could have been established by randomising T , whereas (b) and (c) may be background knowledge. Nothing is known, however, about the causal relationship between X_1 and Y (apart from the confounding effect of X_2). The following graph summarises these causal relationships between the variables:



Assume we establish on observational data that there is a dependence between X_1 and Y and that we cannot render these variables conditionally independent by conditioning on any combination of the remaining observable variables T and X_2 . Employing the widely accepted Markov condition, we can conclude that either $X_1 \rightarrow Y$, $X_1 \leftarrow Y$, $X_1 \leftarrow H \rightarrow Y$ for some unobserved variable H , or some combination of the aforementioned settings. Without any further assumptions, however, these models are observationally equivalent. That is, we cannot refute any of the above possibilities based on observational data alone. Even randomising

T does not help: The above models are interventionally equivalent with respect to interventions on T . We could apply one of the causal discovery methods described in Section 2.3. All of these methods, however, employ further assumptions on the data generating process that go beyond the Markov condition. We may deem some of those assumptions implausible given prior knowledge about the system. Yet, in the absence of targeted interventions on X_1 , X_2 or Y , we can neither falsify candidate models obtained by such methods nor can we test all of the underlying assumptions. In Section 4.2, we illustrate how we may benefit from heterogeneity in the data, that is, from interventional data where the intervention target is unknown.

3.2 Challenge 2: Finding the right variables

Causal discovery often starts by considering observations of some variables Z_1, \dots, Z_d among which we wish to infer cause-effect relationships, thereby implicitly assuming that those variables are defined or constructed in a way that they can meaningfully be interpreted as causal entities in our model. This, however, is not necessarily the case in neuroscience. Without knowing how higher-level causal concepts emerge from lower levels, for example, it is hard to imagine how to make sense and use of a causal model of the 86 billion neurons in a human brain (Herculano-Houzel, 2012). One may hypothesise that a model of averaged neuronal activity in distinct functional brain regions may be pragmatically useful to reason about the effect of different treatments and to understand the brain. For such an approach we need to find the right transformation of the high-dimensional observed variables to obtain the right variables for a causal explanation of the system.

The problem of relating causal models with different granularity and finding the right choice of variable transformations that enable causal reasoning has received attention in the causality literature also outside of neuroscience applications. Eberhardt (2016) fleshes out an instructive two-variable example that demonstrates that the choice of variables for causal modelling may be underdetermined even if interventions were available. For a wrong choice of variables our ability to causally reason about a system breaks. An example of this is the historic debate about whether a high cholesterol diet was beneficial or harmful with respect to heart disease. It can be partially explained by an ambiguity of how exactly total cholesterol is manipulated. Today, we know that low-density lipoproteins and high-density lipoproteins have opposing effects on heart disease risk. Merging these variables together to total cholesterol does not yield a variable with a well-defined intervention: Referring to an intervention on total cholesterol does not specify what part of the intervention is due to a change in low-density lipoproteins (LDL) versus high-density lipoproteins (HDL). As such, only including total cholesterol instead of LDL and HDL may therefore be regarded as a too coarse-grained variable representation that breaks a model's causal semantics, that is, the ability to map every intervention to a well-defined interventional distribution (Spirtes & Scheines, 2004; Steinberg, 2007; Truswell, 2010).

Yet, we may sometimes prefer to transform micro variables into macro variables. This can result in a concise summary of the causal information that abstracts away detail, is easier to communicate and operationalise, and more effectively represents the information necessary for a certain task (Hoel, Albantakis & Tononi, 2013; Hoel, 2017; Weichwald, 2019); for ex-

ample, a causal model over 86 billion neurons may be unwieldy for a brain surgeon aiming to identify and remove malignant brain tissue guided by the cognitive impairments observed in a patient. Rubenstein & Weichwald et al. (2017) formalise a notion of exact transformations that ensures causally consistent reasoning between two causal models where the variables in one model are transformations of the variables in the other. Roughly speaking, two models are considered causally consistent if the following two ways to reason about how the distribution of the macro-variables changes upon a macro-level intervention agree with one another: (a) find an intervention on the micro-variables that corresponds to the considered macro-level intervention, and consider the macro-level distribution implied by the micro-level intervention, and (b) obtain the interventional distribution directly within the macro-level structural causal model sidestepping any need to refer to the micro-level. If the two resulting distributions agree with one another for all (compositions of) interventions, then the two models are said to be causally consistent and we can view the macro-level as an exact transformation of the micro-level causal model that preserves its causal semantics. A formal exposition of the framework and its technical subtleties can be found in the aforementioned work. Here, we revisit a variant of the cholesterol example for an illustration of what it entails for two causal models to be causally consistent and illustrate a failure mode: Consider variables L (LDL), H (HDL), and D (disease), where $D := H - L + N_D$ for L, H, N_D mutually independent random variables. Then a model based on the transformed variables $T = L + H$ and $D \equiv D$ is in general not causally consistent with the original model: For $(l_1, h_1) \neq (l_2, h_2)$ with $l_1 + h_1 = l_2 + h_2$ the interventional distributions induced by the micro-level model corresponding to setting $L := l_1$ and $H := h_1$ or alternatively $L := l_2$ and $H := h_2$ do in general not coincide due to the differing effects of L and H on D . Both interventions correspond to the same level of T and the intervention setting $T := t$ with $t = l_1 + h_1 = l_2 + h_2$ in the macro-level model. Thus, the distributions obtained from reasoning (a) and (b) above do not coincide. If, on the other hand, we had $\tilde{D} := H + L + N_D$, then we could indeed use a macro-level model where we consider $T = H + L$ to reason about the distribution of \tilde{D} under the intervention $\text{do}(T := t)$ without running into conflict with the interventional distributions implied by all corresponding interventions in the micro-level model. This example can analogously be considered in the context of our running examples (cf. Section 1.1): Instead of LDL, HDL, and disease one could alternatively think of some neuronal activity (L) that delays motor response, some neuronal activity (H) that increases attention levels, and the detected reaction time (D) assessed by subjects performing a button press; the scenario then translates into how causal reasoning about the cause of slowed reaction times is hampered once we give up on considering H and L as two separate neural entities and instead try to reason about the average activity T . Janzing, Rubenstein and Schölkopf (2018) observe similar problems for causal reasoning when aggregating variables and show that the observational and interventional stationary distributions of a bivariate autoregressive processes cannot in general be described by a two-variable causal model. A recent line of research focuses on developing a notion of approximate transformations of causal models (Beckers & Halpern, 2019; Beckers, Eberhardt & Halpern, 2019). While there exist first approaches to learn discrete causal macro-variables from data (Chalupka, Perona & Eberhardt, 2015; Chalupka, Eberhardt & Perona, 2016), we are unaware of any method that is generally applicable and learns causal variables from complex high-dimensional data.

In cognitive neuroscience, we commonly treat large-scale brain networks or brain systems as causal entities and then proceed to infer interactions between those (Yeo et al., 2011; Power et al., 2011). Smith et al. (2011) demonstrate that this should be done with caution: Network identification is strongly susceptible to slightly wrong or different definitions of the regions of interest (ROIs) or the so-called atlas. Analyses based on Granger causality depend on the level of spatial aggregation and were shown to reflect the intra-areal properties instead of the interactions among brain regions if an ill-suited aggregation level is considered (Chicharro & Panzeri, 2014). Currently, there does not seem to be consensus as to which macroscopic entities and brain networks are the right ones to (causally) reason about cognitive processes (Uddin, Yeo & Spreng, 2019). Furthermore, the observed variables themselves are already aggregates: A single fMRI voxel or the local field potential at some cortical location reflects the activity of thousands of neurons (Logothetis, 2008; Einevoll, Kayser, Logothetis & Panzeri, 2013); EEG recordings are commonly considered a linear superposition of cortical electromagnetic activity which has spurred the development of blind source separation algorithms that try to invert this linear transformation to recover the underlying cortical variables (Nunez & Srinivasan, 2006).

4 Causality and leveraging robustness

4.1 Robustness of causal models

The concept of causality is linked to invariant models and distributional robustness. Consider again the setting with a target variable Y and covariates X_1, \dots, X_d , as described in the running examples in Section 1.1. Suppose that the system is observed in different environments. Suppose further that the generating process can be described by an SCM, that $\text{PA}(Y) \subseteq \{X_1, \dots, X_d\}$ are the causal parents of Y , and that the different environments correspond to different interventions on some of the covariates, while we neither (need to) know the interventions' targets nor its precise form. In our reaction time example, the two environments may represent two subjects (say, a left-handed subject right after having dinner and a trained race car driver just before a race) that differ in the mechanisms for X_1 , X_3 , and X_7 . Then the joint distribution over Y, X_1, \dots, X_d may be different between the environments and also the marginal distributions may vary. Yet, if the interventions do not act directly on Y the causal model is invariant in the following sense: the conditional distribution of $Y \mid \mathbf{X}_{\text{PA}(Y)}$ is the same in all environments. In the reaction time examples this could translate to the neuronal causes that facilitate fast (versus slow) reaction times to be the same across subjects. This invariance can be formulated in different ways. For example, we have for all k and ℓ , where k and ℓ denote the indices of two environments, and for almost all x

$$(Y^k \mid \mathbf{X}_{\text{PA}(Y)}^k = x) = (Y^\ell \mid \mathbf{X}_{\text{PA}(Y)}^\ell = x) \quad \text{in distribution.} \quad (2)$$

Equivalently,

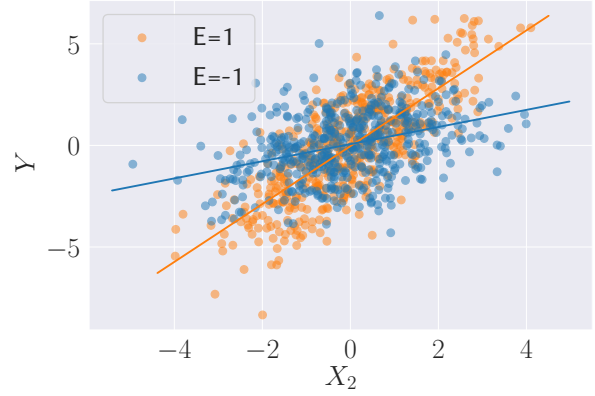
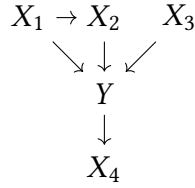
$$E \perp\!\!\!\perp Y \mid \mathbf{X}_{\text{PA}(Y)}, \quad (3)$$

where the variable E represents the environment. In practice, we often work with model classes such as linear or logistic regression for modelling the conditional distribution $Y | \mathbf{X}_{\text{PA}(Y)}$. For such model classes, the above statements simplify. In case of linear models, for example, Equations (2) and (3) translate to regression coefficients and error variances being equal across different environments.

For an example, consider a system that, for environment $E = 1$, is governed by the following structural assignments

SCM for $E = 1$:

$$\begin{aligned} X_1 &:= N_1 \\ X_2 &:= 1 \cdot X_1 + N_2 \\ X_3 &:= N_3 \\ Y &:= X_1 + X_2 + X_3 + N_Y \\ X_4 &:= Y + N_2 \end{aligned}$$



with N_1, N_2, N_3, N_4, N_Y mutually independent and standard-normal, and where environment $E = -1$ corresponds to an intervention changing the weight of X_1 in the assignment for X_2 to -1 . Here, for example, $\{X_1, X_2, X_3\}$ and $\{X_1, X_2, X_4\}$ are so-called invariant sets: the conditionals $Y|X_1, X_2, X_3$ and $Y|X_1, X_2, X_4$ are the same in both environments. The invariant models $Y|X_1, X_2, X_3$ and $Y|X_1, X_2, X_4$ generalise to a new environment $E = -2$, which changes the same weight to -2 , in that they would still predict well. Note that $Y|X_1, X_2, X_4$ is a non-causal model. The lack of invariance of $Y|X_2$ is illustrated by the different regression lines in the scatter plot on the right.

The validity of (2) and (3) follows from the fact that the interventions do not act on Y directly and can be proved using the equivalence of Markov conditions (Lauritzen, 1996; Peters, Bühlmann & Meinshausen, 2016, Section 6.6). Here, we try to argue that it also makes sense intuitively. Suppose that someone proposes to have found a complete causal model for a target variable Y , using certain covariates \mathbf{X}_S (for Y , we may again think of the reaction time in Example A). Suppose that fitting that model for different subjects yields significantly different model fits – maybe even with different signs for the causal effects from variables in \mathbf{X}_S to Y such that $E \perp\!\!\!\perp Y | \mathbf{X}_S$ is violated. In this case, we would become sceptical about whether the proposed model is indeed a complete causal model. Instead, we might suspect that the model is missing an important variable describing how reaction time depends on brain activity.

In practice, environments can represent different sources of heterogeneity. In a cognitive neuroscience setting, environments may be thought of as different subjects who react differently, yet not arbitrarily so (cf. Section 2.1), to varying levels of alcohol consumption. Likewise, different experiments that are thought to involve the same cognitive processes may be thought of as environments; for example, the relationship ‘neuronal activity \rightarrow reaction time’ (cf. Example A, Section 1.1) may be expected to translate from an experiment that compares

reaction times after consumption of alcoholic versus non-alcoholic beers to another experiment where subjects are exposed to Burgundy wine versus grape juice. The key assumption is that the environments do not alter the mechanism of Y —that is, $f_Y(\mathbf{PA}(Y), N_Y)$ —directly or, more formally, there are no interventions on Y . To test whether a set of covariates is invariant, as described in (2) and (3), no causal background knowledge is required.

The above invariance principle is also known as ‘modularity’ or ‘autonomy’. It has been discussed not only in the field of econometrics (Haavelmo, 1944; Aldrich, 1989; Hoover, 2008), but also in philosophy of science. Woodward (2005) discusses how the invariance idea rejects that ‘either a generalisation is a law or else is purely accidental’. In our notion, the criteria (2) and (3) depend on the environments E . In particular, a model may be invariant with respect to some changes, but not with respect to others. In this sense, robustness and invariance should always be thought with respect to a certain set of changes. Woodward (2005) introduces the possibility to talk about various degrees of invariance, beyond the mere existence or absence of invariance, while acknowledging that mechanisms that are sensitive even to mild changes in the background conditions are usually considered as not scientifically interesting. Cartwright (2003) analyses the relationship between invariant and causal relations using linear deterministic systems and draws conclusions analogous to the ones discussed above. In the context of the famous Lucas critique (Lucas, 1976), it is debated to which extent invariance can be used for predicting the effect of changes in economic policy (Cartwright, 2009): Economy consists of many individual players who are capable of adapting their behaviour to a change in policy. In cognitive neuroscience, we believe that the situation is different. Cognitive mechanisms do change and adapt, but not necessarily arbitrarily quickly. Some cognitive mechanism of an individual at the same day can be assumed to be invariant with respect to changes in the visual input, say. Depending on the precise setup, however, we may expect moderate changes of the mechanisms, say, for example, the development of cognitive function in children or learning effects. In other settings, where mechanisms may be subject to arbitrary large changes, scientific insight seems impossible (see Section 2.1).

Recently, the principle of invariance has also received increasing attention in the statistics and machine learning community (Schölkopf et al., 2012; Peters et al., 2016; M. Arjovsky & Lopez-Paz, 2019). It can also be applied to models that do not have the form of an SCM. Examples include dynamical models that are governed by differential equations (Pfister, Bauer & Peters, 2019a).

4.2 Distributional robustness and scarcity of interventional data

The idea of distributional robustness across changing background conditions may help us to falsify causal hypotheses, even when interventional data is difficult to obtain, and in this sense may guide us towards models that are closer to the causal ground truth. For this, suppose that the data are obtained in different environments and that we expect a causal model for Y to yield robust performance across these environments (see Section 4.1). Even if we lack targeted interventional data in cognitive neuroscience and thus cannot test a causal hypothesis directly, we can test the above implication. We can test the invariance, for example, using conditional independence tests or specialised tests for linear models (Chow, 1960). We can, as a surrogate, hold out one environment, train our model on the remaining environ-

ments, and evaluate how well that model performs on the held-out data (cf. Figure 2); the reasoning is that a non-invariant model may not exhibit robust predictive performance and instead yield a bad predictive performance for one or more of the folds. If a model fails the above then either (1) we included the wrong variables, (2) we have not observed important variables, or (3) the environment directly affects Y . Tackling (1), we can try to refine our model and search for different variable representations and variable sets that render our model invariant and robust in the post-analysis. In general, there is no way to recover from (2) and (3), however.

While a model that is not invariant across environments cannot be the complete causal model (assuming the environments do not act directly on the target variable), it may still have non-trivial prediction performance and predict better than a simple baseline method in a new, unseen environment. The usefulness of a model is questionable, however, if its predictive performance on held-out environments is not significantly better than a simple baseline. Conversely, if our model shows robust performance on the held-out data and is invariant across environments, it has the potential of being a causal model (while it need not be; see Section 4.1 for an example). Furthermore, a model that satisfies the invariance property is interesting in itself as it may enable predictions in new, unseen environments. For this line of argument, it does not suffice to employ a cross-validation scheme that ignores the environment structure and only assesses predictability of the model on data pooled across environments. Instead, we need to respect the environment structure and assess the distributional robustness of the model across these environments.

For an illustration of the interplay between invariance and predictive performance, consider a scenario in which $X_1 \rightarrow Y \rightarrow H \rightarrow X_2$, where H is unobserved. Here, we regard different subjects as different environments and suppose that (unknown to us) the environment acts on H : One may think of a variable E pointing into H . Let us assume that our study contains two subjects, one that we use for training and another one that we use as held-out fold. We compare a model of the form $\widehat{Y} = f(\mathbf{X}_{\text{PA}(Y)}) = f(X_1)$ with a model of the form $\widetilde{Y} = g(\mathbf{X}) = g(X_1, X_2)$. On a single subject, the latter model including all observed variables has more predictive power than the former model that only includes the causes of Y . The reason is that X_2 carries information about H , which can be leveraged to predict Y . As a result, $g(X_1, X_2)$ may predict Y well (and even better than $f(X_1)$) on the held-out subject if it is similar to the training subject in that the distribution of H does not change between the subjects. If, however, H was considerably shifted for the held-out subject, then the performance of predicting Y by $g(X_1, X_2)$ may be considerably impaired. Indeed, the invariance is violated and we have $E \not\perp Y | X_1, X_2$. In contrast, the causal parent model $f(X_1)$ may have worse accuracy on the training subject but satisfies invariance: Even if the distribution of H is different for held-out subjects compared to the training subject, the predictive performance of the model $f(X_1)$ does not change. We have $E \perp Y | X_1$.

In practice, we often consider more than two environments. We hence have access to several environments when training our model, even if we leave out one of the environments to test on. In principle, we can thus already during training distinguish between invariant and non-invariant models. While some methods have been proposed that explicitly make use of these different environments during training time (cf. Section 4.4), we regard this as a

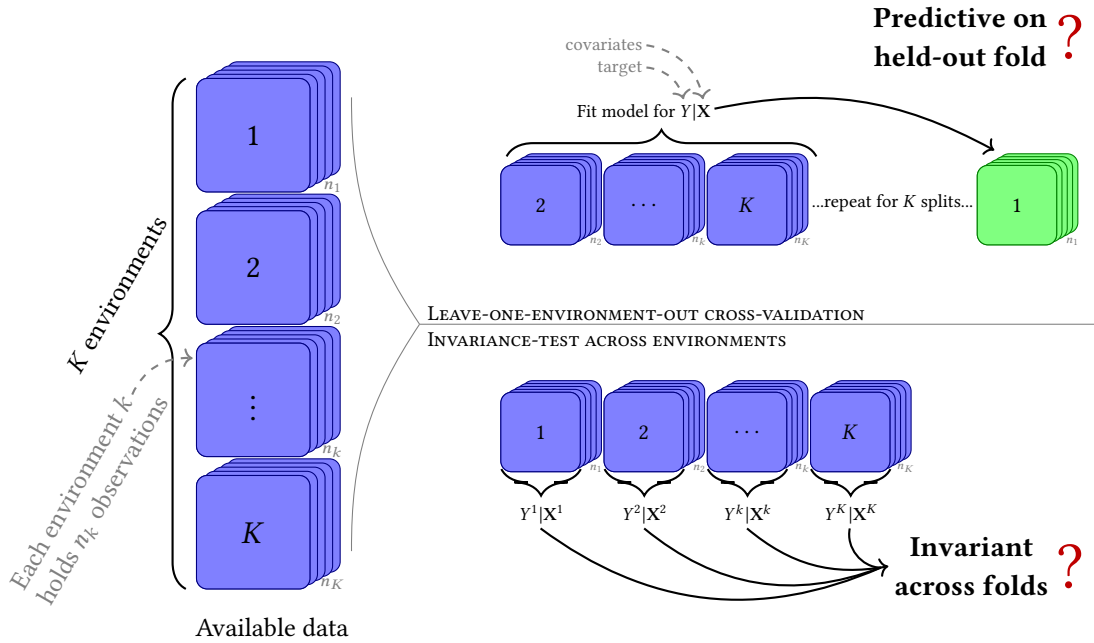


Figure 2: Illustration of a cross-validation scheme across K environments (cf. Section 4.2). Environments can correspond to recordings on different days, during different tasks, or on different subjects, and define how the data is split into folds for the cross-validation scheme. We propose to assess a model by (a) leave-one-environment-out cross-validation testing for robust predictive performance on the held-out fold and (b) an invariance-test across environments assessing whether the model is invariant across folds. The cross-validation scheme (a) is repeated K times, so that each environment acts as a held-out fold once. Models whose predictive performance does not generalise to held-out data or that are not invariant across environments can be refuted as non-causal. For linear models, for example, invariance across environments can be assessed by evaluating to which extent regression coefficients and error variances differ across folds (cf. Section 4.2).

mainly unexplored but promising area of research. In Section 4.2.1, we present a short analysis of classifying motor imagery conditions on EEG data that demonstrates how leveraging robustness may yield models that generalise better to unseen subjects.

In summary, employing distributional robustness as guiding principle prompts us to reject models as non-causal if they are not invariant or if they do not generalise better than a simple baseline to unseen environments, such as sessions, days, neuroimaging modalities, subjects, or other slight variations to the experimental setup. Models that are distributionally robust and do generalise to unseen environments are not necessarily causal but satisfy the prerequisites for being interesting candidate models when it comes to capturing the underlying causal mechanisms.

4.2.1 Exemplary proof-of-concept EEG analysis: leave-one-environment-out cross-validation

Here, we illustrate the proposed cross-validation scheme presented in Figure 2 on motor imagery EEG data due to Tangermann et al. (2012). The data consist of EEG recordings of 9 subjects performing multiple trials of 4 different motor imagery tasks. For each subject 22-channel EEG recordings at 250 Hz sampling frequency are available for 2 days with 6 runs of 48 trials each. We analysed the publicly available data that is bandpass filtered between 0.5 and 100 Hz and 50 Hz notch filtered. The data was further preprocessed by re-referencing to common average reference (car) and projecting onto the orthogonal complement of the null component. Arguably, the full causal structure in this problem is unknown. Instead of assessing the causal nature of a model directly, we therefore evaluate whether distributional robustness of a model across training subjects may help to find models that generalise better to new unseen subjects.

120 models were derived on the training data comprising recordings of the first 5 subjects. Models relied on 6 different sets of extracted timeseries components: the re-referenced EEG channels, 3 different sets of 5 PCA components with varying variance-explained ratios, and 2 sets of 5 corICA components using neighbouring covariance pairs and a partition size of 15 seconds (cf. Pfister, Weichwald, Bühlmann and Schölkopf (2019b) for more details on corICA). Signals were bandpass filtered in 4 different frequency bands (8 – 30, 8 – 20, 20 – 30, and 58 – 80). For each trial and feature set, bandpower features for classification were obtained as the log-variance of the bandpass-filtered signal during seconds 3 – 6 of each trial. For each of the $6 \cdot 4 = 24$ configurations of trial features, we fitted 5 different linear discriminant analysis classifiers without shrinkage, with automatic shrinkage based on the Ledoit-Wolf lemma, and shrinkage parameter settings 0.2, 0.5, and 0.8. These 120 pipelines were fitted once on the entire training data and classification accuracies and areas under the receiver operating curve scores obtained on 4 held-out subjects (y -axes in Figure 3). Classifier performance was cross-validated on the training data following the following three different cross-validation schemes (cross-validation scores are shown on the x -axes in Figure 3):

loso-cv *Leave-one-subject-out cross-validation* is the proposed cross-validation scheme. We hold out data corresponding to each training subject once, fit an LDA classifier on the remaining training data, and assess the models accuracy on the held-out training

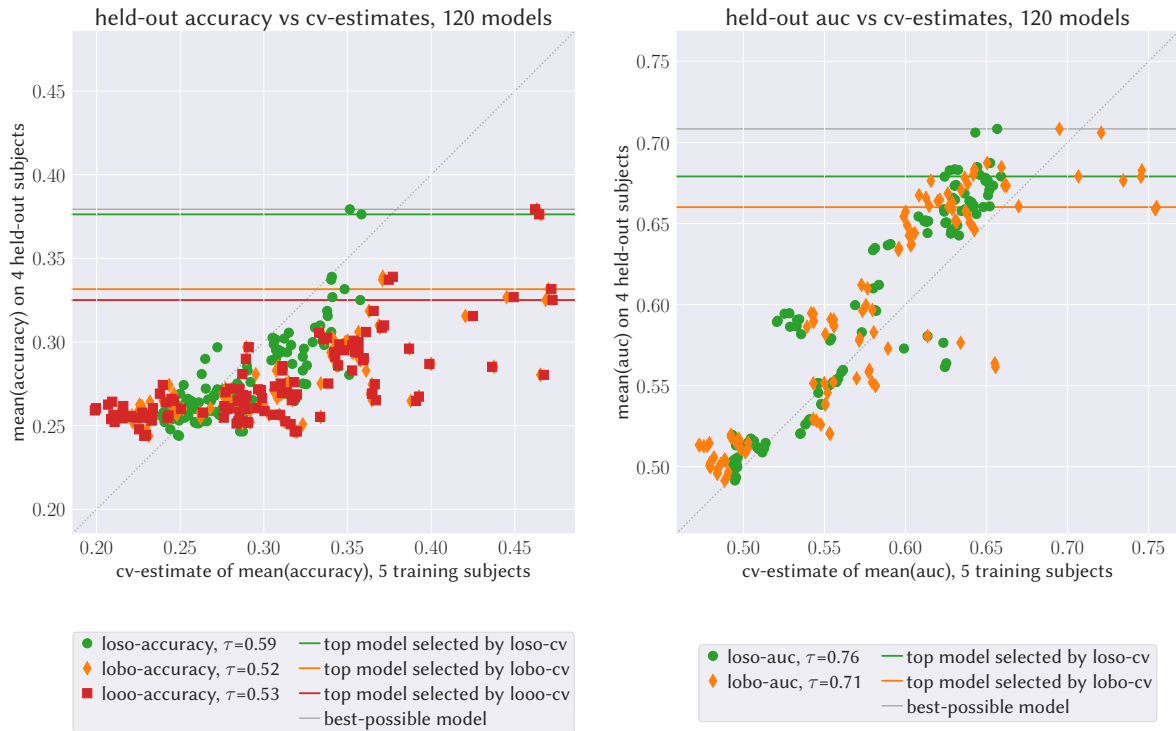


Figure 3: We compare 120 models for the prediction of 4 motor imagery tasks that leverage different EEG components, bandpower features in different frequency bands, and different classifiers. The left and right panel consider classification accuracy or AUC averaged over four held-out subjects as performance measure, respectively. The leave-one-subject-out (loso) cross-validation accuracy on 5 training subjects captures how robust a model is across training subjects. This leave-one-environment-out cross-validation scheme (see Figure 2) seems indeed able to identify models that generalise slightly better to new unseen environments (here the 4 held-out subjects) than a comparable 7-fold leave-one-block-out (lobo)-cv or the leave-one-observation-out (looo) scheme. This is reflected in the Kendall’s τ rank correlation and the scores of the top-ranked models. All top-ranked models outperform random guessing on the held-out subjects (which corresponds to 25% and 50% in the left and right figure, respectively). The displacement along the x -axis of the lobo- and loso-cv scores indicates the previously reported overestimation of held-out performance when using those cross-validation schemes.

subject. The average of those cross-validation scores reflects how robustly each of the 120 classifier models performs across environments (here subjects).

lobo-cv *Leave-one-block-out cross-validation* is a 7-fold cross-validation scheme that is similar to the above loso-cv scheme, where the training data is split into random 7 blocks of roughly equal size. Not respecting the environment structure within the training data, this cross-validation scheme does not capture a models robustness across environments.

looo-cv *Leave-one-observation-out cross-validation* leaves out a single observation and is equivalent to lobo-cv with a block size of one.

In Figure 3 we display the results of the different cross-validation schemes and the Kendall's τ rank correlation between the different cv-scores derived on the training data and a model's classification performance on the four held-out subjects. The loso-cv scores correlate more strongly with held-out model performance and thereby slightly better resolve the relative model performance. Considering the held-out performance for the models with top cv-scores, we observe that selecting models based on the loso-cv score may indeed select models that tend to perform slightly better on new unseen subjects. Furthermore, comparing the displacement of the model scores from the diagonal shows that the loso-cv scheme's estimates are less-biased than the lobo and looo cross-validation scores, when used as an estimate for the performance on held-out subjects; this is in line with Varoquaux et al. (2017).

4.3 Robustness and variable constructions

Whether distributional robustness holds can depend on whether we consider the right variables. This is shown by the following example. Assume that the target Y is caused by the two brain signals X_1 and X_2 via

$$Y := \alpha X_1 + \beta X_2 + N_Y,$$

for some $\alpha \neq 0, \beta \neq 0$, and noise variable N_Y . Assume further that the environment influences the covariates X_1 and X_2 via $X_1 := X_2 + E + N_1$ and $X_2 := E + N_2$, but does not influence Y directly. Here, X_1 and X_2 may represent neuronal activity in two brain regions that are causal for reaction times while E may indicate the time of day or respiratory activity. We then have the invariance property

$$E \perp\!\!\!\perp Y \mid X_1, X_2.$$

If, however, we were to construct or—due to limited measurement ability—only be able to observe $\tilde{X} := X_1 + X_2$, then whenever $\alpha \neq \beta$ we would find that

$$E \not\perp\!\!\!\perp Y \mid \tilde{X}.$$

This conditional dependence is due to many value pairs for (X_1, X_2) leading to the same value of \tilde{X} : Given $\tilde{X} = \tilde{x}$, the value of E holds information about whether say $(X_1, X_2) =$

$(\tilde{x}, 0)$ or $(X_1, X_2) = (0, \tilde{x})$ is more probable and thus—since X_1 and X_2 enter Y with different weights—holds information about Y ; E and Y are conditionally dependent given \tilde{X} . Thus, the invariance may break down when aggregating variables in an ill-suited way. This example is generic in that the same conclusions hold for all assignments $X_1 := f_1(X_2, E, N_1)$ and $X_2 := f_2(E, N_2)$, as long as causal minimality, a weak form of faithfulness, is satisfied (Spirtes et al., 2001).

Rather than taking the lack of robustness as a deficiency, we believe that this observation has the potential to help us finding the right variables and granularity to model our system of interest. If we are given several environments, the guiding principle of distributional robustness can nudge our variable definition and ROI definition towards the construction of variables that are more suitable for causally modelling some cognitive function. If some ROI activity or some EEG bandpower feature does not satisfy any invariance across environments then we may conclude that our variable representation is misaligned with the underlying causal mechanisms or that important variables have not been observed (assuming that the environments do not act on Y directly).

This idea can be illustrated by a thought experiment that is a variation of the LDL-HDL example in Section 3.2: Assume we wish to aggregate multiple voxel-activities and represent them by the activity of a ROI defined by those voxels. For example, let us consider the reaction time scenario (Example A, Section 1.1) and voxels X_1, \dots, X_d . Then we may aggregate the voxels $\bar{X} = \sum_{i=1}^d X_i$ to obtain a macro-level model in which we can still sensibly reason about the effect of an intervention on the treatment variable T onto the distribution of \bar{X} , the ROIs average activity. Yet, the model is in general not causally consistent with the original model. First, our ROI may be chosen too coarse such that for $\hat{\mathbf{x}} \neq \tilde{\mathbf{x}} \in \mathbb{R}^d$ with $\sum_{i=1}^d \hat{x}_i = \sum_{i=1}^d \tilde{x}_i = \bar{x}$ the interventional distributions induced by the micro-level model corresponding to setting all $X_i := \tilde{x}_i$ or alternatively to $X_i := \hat{x}_i$ do not coincide—for example, a ROI that effectively captures the global average voxel-activity cannot resolve whether a higher activity is due to increased reaction-time-driving neuronal entities or due to some upregulation of other neuronal processes unrelated to the reaction time, such as respiratory activity. This ROI would be ill-suited for causal reasoning and non-robust as there are two micro-level interventions that imply different distributions on the reaction times while corresponding to the same intervention setting $\bar{X} := \bar{x}$ with $\bar{x} = \sum_{i=1}^d \hat{x}_i = \sum_{i=1}^d \tilde{x}_i$ in the macro-level model. Second, our ROI may be defined too fine grained such that, for example, the variable representation does only pick up on the left-hemisphere hub of a distributed neuronal process relevant for reaction times. If the neuronal process has different laterality in different subjects, then predicting the effects of interventions on only the left-hemispherical neuronal activity cannot be expected to translate to all subjects. Here, a macro-variable that averages more voxels, say symmetric of both hemispheres, may be more robust to reason about the causes of reaction times than the more fine grained unilateral ROI. In this sense, seeking for variable constructions that enable distributionally robust models across subjects, may nudge us to meaningful causal entities. The spatially refined and finer resolved cognitive atlas obtained by Varoquaux et al. (2018), whose map definition procedure was geared towards an atlas that would be robustness across multiple studies and 196 different experimental conditions, may be seen as an indicative manifestation of the above reasoning.

4.4 Existing methods exploiting robustness

We now present some existing methods that explicitly consider the invariance of a model. While many of these methods are still in their infancy when considering real world applications, we believe that further development in that area could play a vital role when tackling causal questions in cognitive neuroscience.

Robust Independent Component Analysis. Independent component analysis (ICA) is commonly performed in the analysis of magneto- and electro-electroencephalography (MEG and EEG) data in order to invert the inevitable measurement transformation that leaves us with observations of a linear superposition of underlying cortical (and non-cortical) activity. The basic ICA model assumes our vector of observed variables X is being generated as $X = AS$ where A is a mixing matrix and $S = [S_1, \dots, S_d]^T$ is a vector of unobserved mutually independent source signals. The aim is to find the unmixing matrix $V = A^{-1}$. If we perform ICA on individual subjects' data separately, the resulting unmixing matrices will often differ between subjects. This not only hampers the interpretation of the resulting sources as some cortical activity that we can identify across subjects, it also hints—in light of the above discussion—at some unexplained variation that is due to shifts in background conditions between subjects such as different cap positioning or neuroanatomical variation. Instead of simply pooling data across subjects, Pfister et al. (2019b) propose a methodology that explicitly exploits the existence of environments, that is, the fact that EEG samples can be grouped by subjects they were recorded from. This way, the proposed confounding-robust ICA (coroICA) procedure identifies an unmixing of the signals that generalises to new subjects. The additional robustness resulted, for their considered example, in improved classification accuracies on held-out subjects and can be viewed as a first-order adjustment for subject specific differences. The application of ICA procedures to pooled data will generally result in components that do not robustly transfer to new subjects and are thus necessarily variables that do not lend themselves for a causal interpretation. The coroICA procedure aims to exploit the environments to identify unmixing matrices that generalise across subjects.

Causal discovery with exogenous variation. Invariant causal prediction, proposed by Peters et al. (2016), aims at identifying the parents of Y within a set of covariates X_1, \dots, X_d . We have argued that the true causal model is invariant across environments, see Equation (2), if the data are obtained in different environments and the environment does not directly influence Y . That is, when enumerating all invariant models by searching through subsets of X_1, \dots, X_d , one of these subsets must be the set of causal parents of Y . As a result, the intersection $\widehat{S} = \bigcap_{S: S \text{ invariant}} S$ of all sets of covariates that yield invariant models is guaranteed to be a subset of the causes $\text{PA}(Y)$ of Y . (Here, we define the intersection over the empty index set as the empty set.) Testing invariance with a hypothesis test to the level α , say $\alpha = 0.05$, one obtains that \widehat{S} is contained in the set of parents of Y with high probability

$$\mathbf{P} \left(\widehat{S} \subseteq \text{PA}(Y) \right) \geq 1 - \alpha.$$

Under faithfulness, the method can be shown to be robust against violation of the above assumptions. If the environment acts on Y directly, for example, there is no invariant set and in the presence of hidden variables, the intersection \widehat{S} of invariant models can still be shown to be a subset of the ancestors of Y with large probability.

It is further possible to model the environment as a random variable (using an indicator variable, for example), that is often called a context variable. One can then exploit the background knowledge of its exogeneity to identify the full causal structure instead of focussing on identifying the causes of a target variable. Several approaches have been suggested (for example, Spirtes et al., 2001; Eaton & Murphy, 2007; Zhang, Huang, Zhang, Glymour & Schölkopf, 2017; Mooij, Magliacane & Claassen, 2020). Often, these methods first identify the target of the intervention and then exploit known techniques of constraint- or score-based methods. Some of the above methods also make use of time as a context variable or environment (Zhang et al., 2017; Pfister, Bühlmann & Peters, 2018).

Anchor regression. We argued above that focusing on invariance has an advantage when inferring causal structure from data. If we are looking for generalisability across environments, however, focusing solely on invariance may be too restrictive. Instead, we may select the most predictive model among all invariant models. The idea of anchor regression is to explicitly trade off invariance and predictability (Rothenhäusler, Meinshausen, Bühlmann & Peters, 2018). For a target variable Y , predictor variables \mathbf{X} , and so-called anchor variables $\mathbf{A} = [A_1, \dots, A_q]^\top$ that represent the different environments and are normalised to have unit variance, the anchor regression coefficients are obtained as solutions to the following minimisation problem

$$\widehat{b}_\gamma := \arg \min_{b \in \mathbb{R}^d} \underbrace{\mathbb{E} [(Y - b^\top \mathbf{X})^2]}_{\text{prediction}} + \gamma \underbrace{\mathbb{E} [\|\mathbf{A}(Y - b^\top \mathbf{X})\|_2^2]}_{\text{invariance}}.$$

Higher parameters γ steer the regression towards more invariant predictions (converging against the two stage least squares solutions in identifiable instrumental variable settings). For $\gamma = 0$ we recover the ordinary least square solution. The solution \widehat{b}_γ can be shown to have the best predictive power under shift interventions up to a certain strength that depends on γ . As before, the anchor variables can code time, environments, subjects, or other factors, and we thus obtain a regression that is distributionally robust against shifts in those factors.

4.5 Time series data

So far, we have mainly focused on the setting of i.i.d. data. Most of the causal inference literature dealing with time dependency considers discrete-time models. This comes with additional complications for causal inference. For example, there are ongoing efforts to adapt causal inference algorithms and account for sub- or sup-sampling and temporal aggregation (Danks & Plis, 2013; Hyttinen, Plis, Jarvisalo, Eberhardt & Danks, 2016). Problems of temporal aggregation relate to Challenge 2 of finding the right variables, which is a conceptual problem in time series models that requires us to clarify our notion of intervention for

time-evolving systems (Rubenstein & Weichwald et al., 2017). When we observe time series with non-stationarities we may consider these as resulting from some unknown shift interventions. That is, non-stationarities over time may be due to shifts in the background conditions and as such can be understood as shifts in environments analogous to the i.i.d. setting. This way, we may again leverage the idea of distributional robustness for inference on time-evolving systems for which targeted interventional data is scarce. Extensions of invariant causal prediction to time series data that aim to leverage such variation have been proposed by Christiansen and Peters (2018), Pfister et al. (2018) and the ICA procedure described in Section 4.4 also exploits non-stationarity over time. SCMs extend to continuous-time models (Peters, Bauer & Pfister, 2020), where the idea to trade off prediction and invariance has been applied to the problem of inferring chemical reaction networks (Pfister et al., 2019a).

A remark is in order if we wish to describe time-evolving systems by one causal summary graph where each time series component is collapsed into one node: For this to be reasonable, we need to assume a time-homogeneous causal structure. Furthermore, it requires us to carefully clarify its causal semantics: While summary graphs can capture the existence of cause-effect relationships, they do in general not correspond to a structural causal model that admits a causal semantics nor enables interventional predictions that are consistent with the underlying time-resolved structural causal model (Rubenstein & Weichwald et al., 2017; Janzing et al., 2018). That is, the wrong choice of time-agnostic variables and corresponding interventions may be ill-suited to represent the cause-effect relationships of a time-evolving system (cf. Challenge 2 and Rubenstein & Weichwald et al., 2017; Janzing et al., 2018).

5 Conclusion and future work

Causal inference in cognitive neuroscience is ambitious. It is important to continue the open discourse about the many challenges, some of which are mentioned above. Thanks to the open and critical discourse there is great awareness and caution when interpreting neural correlates (Rees, Kreiman & Koch, 2002). Yet, “FC [functional connectivity] researchers already work within a causal inference framework, whether they realise it or not” (Reid et al., 2019).

In this article we have provided our view on the numerous obstacles to a causal understanding of cognitive function. If we, explicitly or often implicitly, ask causal questions, we need to employ causal assumptions and methodology. We propose to exploit that causal models using the right variables are distributionally robust. In particular, we advocate distributional robustness as a guiding principle for causality in cognitive neuroscience. While causal inference in general and in cognitive neuroscience in particular is a challenging task, we can at least exploit the rational to refute models and variables as non-causal that are frail to shifts in the environment. This guiding principle does not necessarily identify causal variables nor causal models, but it nudges our search into the right direction away from frail models and non-causal variables. While we presented first attempts that aim to leverage observations obtained in different environments (cf. Section 4.4), this article poses more questions for future research than it answers.

We believe that procedures that exploit environments during training are a promising

avenue for future research. While we saw mildly positive results in our case study, further research needs to show whether this trend persists in studies with many subjects. It may be possible to obtain improvements when combining predictive scores on held-out-subjects with other measures of invariance and robustness. The development of such methods may be spurred and guided by field-specific benchmarks (or competitions) that assess models' distributional robustness across a wide range of scenarios, environments, cognitive tasks, and subjects.

When considering robustness or invariance across trainings environments, the question arises how the ability to infer causal structure and the generalisation performance to unseen environments depend on the number of training environments. While first attempts have been made to theoretically understand that relation (Rothenhäusler et al., 2018; Christiansen, Pfister, Jakobsen, Gnecco & Peters, 2020), most of the underlying questions are still open. We believe that an answer would depend on the strength of the interventions, the sample size, the complexity of the model class and, possibly, properties of the (meta-)distribution of environments.

We believe that advancements regarding the errors-in-variable problem may have important implications for cognitive neuroscience. Nowadays, we can obtain neuroimaging measurements at various spatial and temporal resolutions using, among others, fMRI, MEG and EEG, positron emission tomography, or near-infrared spectroscopy (Filler, 2009; Poldrack, 2018). Yet, all measurement modalities are imperfect and come with different complications. One general problem is that the observations are corrupted by measurement noise. The errors-in-variables problem complicates even classical regression techniques where we wish to model $Y \approx \beta X^* + \epsilon$ but only have access to observations of a noise-corrupted $X = X^* + \eta$ (Schennach, 2016). This inevitably carries over and hurdles causal inference as the measurement noise spoils conditional independence testing, biases any involved regression steps, and troubles additive noise approaches that aim to exploit noise properties for directing causal edges and methods testing for invariance. First steps addressing these issues in the context of causal discovery have been proposed by (Zhang et al., 2018; Blom, Klimovskaia, Magliacane & Mooij, 2018; Scheines & Ramsey, 2016).

Summarising, we believe that there is a need for causal models if we aim to understand the neuronal underpinnings of cognitive function. Only causal models equip us with concepts that allow us to explain, describe, predict, manipulate, deal and interact with, and reason about a system and that allow us to generalise to new, unseen environments. A merely associational model suffices to predict naturally unfolding disease progression, for example. We need to obtain understanding in form of a causal model if our goal is to guide rehabilitation after cognitive impairment or to inform the development of personalised drugs that target specific neuronal populations. Distributional robustness and generalisability to unseen environments is an ambitious goal, in particular in biological systems and even more so in complex systems such as the human brain. Yet, it may be the only and most promising way forward.

Acknowledgments

The authors thank the anonymous reviewers for their constructive and helpful feedback on an earlier version of this manuscript. SW was supported by the Carlsberg Foundation.

References

- Aldrich, J. (1989). Autonomy. *Oxford Economic Papers*, 41, 15–34.
- Antal, A. & Herrmann, C. S. (2016). Transcranial alternating current and random noise stimulation: possible mechanisms. *Neural Plasticity*, 2016.
- Bach, D. R., Symmonds, M., Barnes, G. & Dolan, R. J. (2017). Whole-Brain Neural Dynamics of Probabilistic Reward Prediction. *Journal of Neuroscience*, 37(14), 3789–3798.
- Beckers, S., Eberhardt, F. & Halpern, J. Y. (2019). Approximate Causal Abstraction. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press.
- Beckers, S. & Halpern, J. Y. (2019). Abstracting Causal Models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI Press.
- Bergmann, T. O. & Hartwigsen, G. (2020). Inferring Causality from Noninvasive Brain Stimulation in Cognitive Neuroscience. *Journal of Cognitive Neuroscience*, 1–29. Forthcoming.
- Bestmann, S. & Walsh, V. (2017). Transcranial electrical stimulation. *Current Biology*, 27(23), R1258–R1262.
- Blom, T., Klimovskaia, A., Magliacane, S. & Mooij, J. M. (2018). An upper bound for random measurement error in causal discovery. In *Proceedings of the 34th annual conference on Uncertainty in Artificial Intelligence (UAI)*.
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.
- Bongers, S., Peters, J., Schölkopf, B. & Mooij, J. M. (2018). Theoretical Aspects of Cyclic Structural Causal Models. *arXiv preprint arXiv:1611.06221*.
- Breakspear, M. (2013). Dynamic and stochastic models of neuroimaging data: A comment on Lohmann et al. *NeuroImage*, 75, 270–274.
- Bühlmann, P., Peters, J. & Ernest, J. (2014). CAM: Causal Additive Models, high-dimensional order search and penalized regression. *Annals of Statistics*, 42, 2526–2556.
- Cartwright, N. (2003). Two theorems on invariance and causality. *Philosophy of Science*, 70(1), 203–224.
- Cartwright, N. (2009). Causality, invariance and policy. In M. Klopotek, A. Przepiorkowski, S. T. Wierzchon & K. Trojanowski (Eds.), *The oxford handbook of philosophy of economics* (pp. 410–423). New York, NY: Oxford University Press.
- Chalupka, K., Eberhardt, F. & Perona, P. (2016). Multi-Level Cause-Effect Systems. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (Vol. 51, pp. 361–369). Proceedings of Machine Learning Research. PMLR.
- Chalupka, K., Perona, P. & Eberhardt, F. (2015). Visual Causal Feature Learning. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press.
- Chicharro, D. & Panzeri, S. (2014). Algorithms of causal inference for the analysis of effective connectivity among brain regions. *Frontiers in Neuroinformatics*, 8, 64.
- Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3, 507–554.
- Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28, 591–605.
- Christiansen, R. & Peters, J. (2018). Invariant Causal Prediction in the Presence of Latent Variables. *arXiv preprint arXiv:1808.05541*.

- Christiansen, R., Pfister, N., Jakobsen, M., Gnecco, N. & Peters, J. (2020). The difficult task of distribution generalization in nonlinear models. *arXiv preprint arXiv:2006.07433*.
- Conniffe, D. (1991). R.A. Fisher and the development of statistics - a view in his centenary year. *Journal of the Statistical and Social Inquiry Society of Ireland*, 26(3), 55–108.
- Danks, D. & Plis, S. (2013). Learning causal structure from undersampled time series.
- Darmois, G. (1953). Analyse générale des liaisons stochastiques. *Revue de l'Institut International de Statistique*, 21, 2–8.
- Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D. & Poldrack, R. A. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *NeuroImage*, 97, 271–283.
- Dubois, J., Oya, H., Tyszka, J. M., Howard, M., Eberhardt, F. & Adolphs, R. (2017). Causal mapping of emotion networks in the human brain: Framework and initial findings. *Neuropsychologia*.
- Eaton, D. & Murphy, K. P. (2007). Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the 11th international conference on artificial intelligence and statistics (AISTATS)* (pp. 107–114).
- Eberhardt, F. (2013). Experimental indistinguishability of causal structures. *Philosophy of Science*, 80(5), 684–696.
- Eberhardt, F. (2016). Green and grue causal variables. *Synthese*, 193(4), 1029–1046.
- Einevoll, G. T., Kayser, C., Logothetis, N. K. & Panzeri, S. (2013). Modelling and analysis of local field potentials for studying the function of cortical circuits. *Nature Reviews Neuroscience*, 14(11), 770–785.
- Filler, A. (2009). The history, development and impact of computed imaging in neurological diagnosis and neurosurgery: CT, MRI, and DTL. *Nature Precedings*.
- Friston, K. J., Harrison, L. & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4), 1273–1302.
- Friston, K., Daunizeau, J. & Stephan, K. E. (2013). Model selection and gobbledygook: Response to Lohmann et al. *NeuroImage*, 75, 275–278.
- Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3), 424–438.
- Grosse-Wentrup, M., Janzing, D., Siegel, M. & Schölkopf, B. (2016). Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. *NeuroImage*, 125, 825–833.
- Györfi, L., Kohler, M., Krzyżak, A. & Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer.
- Haavelmo, T. (1944). The Probability Approach in Econometrics. *Econometrica*, 12, S1–S115 (supplement).
- Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B. & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87, 96–110.
- Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences*, 109(Supplement 1), 10661–10668.

- Herrmann, C. S., Rach, S., Neuling, T. & Strüber, D. (2013). Transcranial alternating current stimulation: a review of the underlying mechanisms and modulation of cognitive processes. *Frontiers in human neuroscience*, 7, 279.
- Hoel, E. P. (2017). When the Map Is Better Than the Territory. *Entropy*, 19(5).
- Hoel, E. P., Albantakis, L. & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49), 19790–19795.
- Hoover, K. D. (2008). Causality in economics and econometrics. In S. N. Durlauf & L. E. Blume (Eds.), *The new palgrave dictionary of economics* (2nd). Basingstoke, UK: Palgrave Macmillan.
- Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J. & Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NeurIPS)* (pp. 689–696). NeurIPS Foundation.
- Huth, A. G., Lee, T., Nishimoto, S., Bilenko, N. Y., Vu, A. T. & Gallant, J. L. (2016). Decoding the Semantic Content of Natural Movies from Human Brain Activity. *Frontiers in Systems Neuroscience*, 10, 81.
- Hyttinen, A., Plis, S., Jarvisalo, M., Eberhardt, F. & Danks, D. (2016). Causal discovery from subsampled time series data by constraint optimization. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models* (Vol. 52, pp. 216–227). Proceedings of Machine Learning Research. PMLR.
- Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
- Jabbari, F., Ramsey, J., Spirtes, P. & Cooper, G. (2017). Discovery of causal models that contain latent variables through bayesian scoring of independence constraints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 142–157). Springer.
- Janzing, D., Rubenstein, P. & Schölkopf, B. (2018). Structural causal models for macro-variables in time-series. *arXiv preprint arXiv:1804.03911*.
- Kar, K., Ito, T., Cole, M. W. & Krekelberg, B. (2020). Transcranial alternating current stimulation attenuates BOLD adaptation and increases functional connectivity. *Journal of Neurophysiology*, 123(1), 428–438.
- Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.
- Lauritzen, S. L., Dawid, A. P., Larsen, B. N. & Leimer, H.-G. (1990). Independence properties of directed markov fields. *Networks*, 20, 491–505.
- Lee, J. H., Durand, R., Gradinaru, V., Zhang, F., Goshen, I., Kim, D.-S., Fenno, L. E., Ramakrishnan, C. & Deisseroth, K. (2010). Global and local fMRI signals driven by neurons defined optogenetically by type and wiring. *Nature*, 465(7299), 788–792.
- Liang, Z., Watson, G. D., Alloway, K. D., Lee, G., Neuberger, T. & Zhang, N. (2015). Mapping the functional network of medial prefrontal cortex by combining optogenetics and fMRI in awake rats. *NeuroImage*, 117, 114–123.
- Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, 453(7197), 869–878.
- Lohmann, G., Erfurth, K., Müller, K. & Turner, R. (2012). Critical comments on dynamic causal modelling. *NeuroImage*, 59(3), 2322–2329.

- Lohmann, G., Müller, K. & Turner, R. (2013). Response to commentaries on our paper: Critical comments on dynamic causal modelling. *NeuroImage*, 75, 279–281.
- López-Alonso, V., Cheeran, B., Rio-Rodríguez, D. & Fernández-del-Olmo, M. (2014). Inter-individual variability in response to non-invasive brain stimulation paradigms. *Brain stimulation*, 7(3), 372–380.
- Lucas, R. J. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1(1), 19–46.
- M. Arjovsky, I. G., L. Bottou & Lopez-Paz, D. (2019). Invariant Risk Minimization. *arXiv preprint arXiv:1907.02893*.
- Marinazzo, D., Pellicoro, M. & Stramaglia, S. (2008). Kernel-Granger causality and the analysis of dynamical networks. *Physical Review E*, 77(5), 056215.
- Marinazzo, D., Liao, W., Chen, H. & Stramaglia, S. (2011). Nonlinear connectivity by Granger causality. *NeuroImage*, 58(2), 330–338.
- Mastakouri, A., Schölkopf, B. & Janzing, D. (2019). Selecting causal brain features with a single conditional independence test per feature. In *Advances in Neural Information Processing Systems 32 (NeurIPS)* (pp. 12532–12543). NeurIPS Foundation.
- Mayberg, H. S., Lozano, A. M., Voon, V., McNeely, H. E., Seminowicz, D., Hamani, C., Schwab, J. M. & Kennedy, S. H. (2005). Deep Brain Stimulation for Treatment-Resistant Depression. *Neuron*, 45(5), 651–660.
- Mehler, D. M. A. & Kording, K. P. (2018). The lure of causal statements: rampant mis-inference of causality in estimated connectivity. *arXiv preprint arXiv:1812.03363*.
- Mill, R. D., Ito, T. & Cole, M. W. (2017). From connectome to cognition: the search for mechanism in human functional brain networks. *NeuroImage*, 160, 124–139.
- Mooij, J. M., Magliacane, S. & Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99), 1–108.
- Mumford, J. A. & Ramsey, J. D. (2014). Bayesian networks for fMRI: A primer. *NeuroImage*, 86, 573–582.
- Nitsche, M. A., Cohen, L. G., Wassermann, E. M., Priori, A., Lang, N., Antal, A., Paulus, W., Hummel, F., Boggio, P. S., Fregni, F. & Pascual-Leone, A. (2008). Transcranial direct current stimulation: State of the art 2008. *Brain Stimulation*, 1(3), 206–223.
- Nunez, P. L. & Srinivasan, R. (2006). *Electric Fields of the Brain: The neurophysics of EEG* (2nd ed.). Oxford University Press.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.
- Peters, J., Bauer, S. & Pfister, N. (2020). Causal models for dynamical systems. *arXiv preprint arXiv:2001.06208*.
- Peters, J. & Bühlmann, P. (2014). Identifiability of Gaussian Structural Equation Models with Equal Error Variances. *Biometrika*, 101(1), 219–228.
- Peters, J., Bühlmann, P. & Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (with discussion)*, 78(5), 947–1012.
- Peters, J., Janzing, D. & Schölkopf, B. (2013). Causal Inference on Time Series using Structural Equation Models. In *Advances in Neural Information Processing Systems 26 (NeurIPS)* (pp. 585–592). NeurIPS Foundation.

- Peters, J., Janzing, D. & Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.
- Peters, J., Mooij, J. M., Janzing, D. & Schölkopf, B. (2014). Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research*, 15, 2009–2053.
- Pfister, N., Bauer, S. & Peters, J. (2019a). Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51), 25405–25411.
- Pfister, N., Bühlmann, P. & Peters, J. (2018). Invariant Causal Prediction for Sequential Data. *Journal of the American Statistical Association*, 114(527), 1264–1276.
- Pfister, N., Weichwald, S., Bühlmann, P. & Schölkopf, B. (2019b). Robustifying Independent Component Analysis by Adjusting for Group-Wise Stationary Noise. *Journal of Machine Learning Research*, 20(147), 1–50. (Co-first authorship between NP and SW)
- Poldrack, R. A. (2018). *The new mind readers: what neuroimaging can and cannot reveal about our thoughts*. Princeton University Press.
- Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L. et al. (2011). Functional network organization of the human brain. *Neuron*, 72(4), 665–678.
- Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A. & Glymour, C. (2010). Six problems for causal inference from fMRI. *NeuroImage*, 49(2), 1545–1558.
- Rees, G., Kreiman, G. & Koch, C. (2002). Neural correlates of consciousness in humans. *Nature Reviews Neuroscience*, 3, 261–270.
- Reichenbach, H. (1956). *The Direction of Time*. University of California Press.
- Reid, A. T., Headley, D. B., Mill, R. D., Sanchez-Romero, R., Uddin, L. Q., Marinazzo, D., Lurie, D. J., Valdés-Sosa, P. A., Hanson, S. J., Biswal, B. B. et al. (2019). Advancing functional connectivity research from association to causation. *Nature neuroscience*, 1(10).
- Rothenhäusler, D., Meinshausen, N., Bühlmann, P. & Peters, J. (2018). Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*.
- Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M. & Schölkopf, B. (2017). Causal Consistency of Structural Equation Models. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*. (Co-first authorship between PKR and SW)
- Sanchez-Romero, R. & Cole, M. W. (2020). Combining Multiple Functional Connectivity Methods to Improve Causal Inferences. *Journal of Cognitive Neuroscience*, 1–15. Forthcoming.
- Sanchez-Romero, R., Ramsey, J. D., Zhang, K., Glymour, M. R. K., Huang, B. & Glymour, C. (2019). Estimating feedforward and feedback effective connections from fMRI time series: Assessments of statistical methods. *Network Neuroscience*, 3(2), 274–306.
- Scheines, R. & Ramsey, J. (2016). Measurement error and causal discovery. In *CEUR workshop proceedings* (Vol. 1792, pp. 1–7).
- Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics*, 8, 341–377.
- Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K. & Mooij, J. M. (2012). On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)* (pp. 1255–1262). Omnipress.

- Shah, R. & Peters, J. (2020). The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *Annals of Statistics (accepted)*, *ArXiv e-prints (1804.07203)*.
- Shimizu, S., Hoyer, P. O., Hyvärinen, A. & Kerminen, A. (2006). A Linear Non-Gaussian Acyclic model for Causal Discovery. *Journal of Machine Learning Research*, 7(10), 2003–2030.
- Silva, R., Scheine, R., Glymour, C. & Spirtes, P. (2006). Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7, 191–246.
- Skitovič, V. P. (1962). Linear combinations of independent random variables and the normal distribution law. *Selected Translations in Mathematical Statistics and Probability*, 2, 211–228.
- Smith, S. M. (2012). The future of fMRI connectivity. *Neuroimage*, 62(2), 1257–1266.
- Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D. & Woolrich, M. W. (2011). Network modelling methods for fMRI. *Neuroimage*, 54(2), 875–891.
- Spirtes, P., Glymour, C. N. & Scheines, R. (2001). *Causation, Prediction, and Search* (2nd ed.). MIT Press.
- Spirtes, P. & Scheines, R. (2004). Causal Inference of Ambiguous Manipulations. *Philosophy of Science*, 71(5), 833–845.
- Steinberg, D. (2007). *The Cholesterol Wars: The Skeptics vs the Preponderance of Evidence*. Academic Press.
- Stramaglia, S., Wu, G.-R., Pellicoro, M. & Marinazzo, D. (2012). Expanding the transfer entropy to identify information circuits in complex systems. *Physical Review E*, 86(6), 066211.
- Stramaglia, S., Cortes, J. M. & Marinazzo, D. (2014). Synergy and redundancy in the Granger causal analysis of dynamical networks. *New Journal of Physics*, 16(10), 105003.
- Tangermann, M., Müller, K.-R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K., Müller-Putz, G., Nolte, G., Pfurtscheller, G., Preissl, H., Schalk, G., Schlögl, A., Vidaurre, C., Waldert, S. & Blankertz, B. (2012). Review of the BCI Competition IV. *Frontiers in Neuroscience*, 6, 55.
- Todd, M. T., Nystrom, L. E. & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage*, 77, 157–165.
- Truswell, A. S. (2010). *Cholesterol and Beyond: The Research on Diet and Coronary Heart Disease 1900–2000*. Springer.
- Uddin, L. Q., Yeo, B. T. T. & Spreng, R. N. (2019). Towards a Universal Taxonomy of Macro-scale Functional Human Brain Networks. *Brain Topography*, 32(6), 926–942.
- Uhler, C., Raskutti, G., Bühlmann, P. & Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *Annals of Statistics*, 41(2), 436–463.
- Valdes-Sosa, P. A., Roebroeck, A., Daunizeau, J. & Friston, K. (2011). Effective connectivity: influence, causality and biophysical modeling. *Neuroimage*, 58(2), 339–361.
- Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y. & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145, 166–179. Individual Subject Prediction.
- Varoquaux, G., Schwartz, Y., Poldrack, R. A., Gauthier, B., Bzdok, D., Poline, J.-B. & Thirion, B. (2018). Atlases of cognition with large-scale human brain mapping. *PLOS Computational Biology*, 14(11), 1–18.

- Verma, T. & Pearl, J. (1990). Equivalence and Synthesis of Causal Models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence (UAI)*.
- Voskuhl, J., Strüber, D. & Herrmann, C. S. (2018). Non-invasive brain stimulation: a paradigm shift in understanding brain oscillations. *Frontiers in human neuroscience*, 12, 211.
- Waldorp, L., Christoffels, I. & van de Ven, V. (2011). Effective connectivity of fMRI data using ancestral graph theory: Dealing with missing regions. *NeuroImage*, 54(4), 2695–2705.
- Weichwald, S. (2019). *Pragmatism and Variable Transformations in Causal Modelling* (Doctoral dissertation, ETH Zurich).
- Weichwald, S., Gretton, A., Schölkopf, B. & Grosse-Wentrup, M. (2016a). Recovery of non-linear cause-effect relationships from linearly mixed neuroimaging data. In *Pattern Recognition in Neuroimaging (PRNI), 2016 International Workshop on*. IEEE.
- Weichwald, S., Grosse-Wentrup, M. & Gretton, A. (2016b). MERLiN: Mixture Effect Recovery in Linear Networks. *IEEE Journal of Selected Topics in Signal Processing*, 10(7), 1254–1266.
- Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T. & Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110, 48–59.
- Weichwald, S., Schölkopf, B., Ball, T. & Grosse-Wentrup, M. (2014). Causal and anti-causal learning in pattern recognition for neuroimaging. In *Pattern Recognition in Neuroimaging (PRNI), 2014 International Workshop on*. IEEE.
- Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.
- Woolgar, A., Golland, P. & Bode, S. (2014). Coping with confounds in multivoxel pattern analysis: What should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013. *NeuroImage*, 98, 506–512.
- Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H. & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3), 1125–1165.
- Zhang, K., Gong, M., Ramsey, J., Batmanghelich, K., Spirtes, P. & Glymour, C. (2018). Causal discovery with linear non-gaussian models under measurement error: structural identifiability results. In *Proceedings of the 34th annual conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press.
- Zhang, K., Huang, B., Zhang, J., Glymour, C. & Schölkopf, B. (2017). Causal discovery from nonstationary/heterogeneous data: skeleton estimation and orientation determination. In *Proceedings of the 26th international joint conference on artificial intelligence (IJCAI)* (pp. 1347–1353).
- Zhang, K. & Hyvärinen, A. (2009). On the Identifiability of the Post-Nonlinear Causal Model. In *Proceedings of the Twenty-Fifth Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press.
- Zheng, X., Dan, C., Aragam, B., Ravikumar, P. & Xing, E. P. (2020). Learning sparse nonparametric DAGs. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*. (to appear).