Causal inference in functional Magnetic Resonance Imaging a Review of current approaches

Natalia Z. Bielczyk[1,2*], Sebo Uithol[1,2,3], Tim van Mourik[1,2], Martha N. Havenith[1,4], Paul Anderson[1,2], Jeffrey C. Glennon[1,2], Jan K. Buitelaar[1,2]

(1) Donders Institute for Brain, Cognition and Behavior, Kapittelweg 29, 6525 EN Nijmegen, the Netherlands

(2) Radboud University Nijmegen Medical Centre, Geert Grooteplein Zuid 10, 6525 GA Nijmegen, the Netherlands

(3) Charité Universitätsmedizin, Bernstein Centre for Computational Neuroscience, Philippstrasse 13, haus 6, 10119, Berlin, Germany

(4) Radboud University Nijmegen, Comeniuslaan 4, 6525 HP, the Netherlands

∗ Correspondence: natalia.bielczyk@radboudumc.nl

August 23, 2017

Abstract

In the past two decades, functional Magnetic Resonance Imaging has been used to relate neuronal network activity to cognitive processing and behaviour. Recently this approach has been augmented by algorithms that allow us to infer causal links between component populations of neuronal networks. Multiple inference procedures have been proposed to approach this research question but so far, each method has limitations when it comes to establishing whole-brain connectivity patterns. In this paper, we discuss the ways to infer causality in fMRI research. We also formulate recommendations for the future directions in this area.

1 What is causality?

Although inferring causal relations is a fundamental aspect of scientific research, the notion of causation itself can be notoriously difficult to define. The basic idea is straightforward: When process A is the cause of process B, A is necessarily in the past from B, and without A, B would not occur. But in practice, and in dynamic systems such as the brain in particular, the picture is far less clear. First, for any event a large number of (potential) causes can be identified. The efficacy of certain neuronal process in producing behavior is dependent on the state of many other (neuronal) processes, but also on the availability of glucose and oxygen in the brain, etc. In a neuroscientific context, we are generally not interested in most of these causes, but only in a cause that stands out in such a way that it is deemed to provide a substantial part of the explanation, for instance causes that vary with the experimental conditions. However, the contrast between relevant and irrelevant causes (in terms of explanatory power) is arbitrary and strongly dependent on experimental setup, contextual factors, etc. For instance, respiratory movement is typically considered a confound in fMRI experiments, unless the research question concerns the influence of respiration speed on the dynamics of the neuronal networks.

In dynamic systems, causal processes are unlikely to be part of a unidirectional chain of events, but rather a causal web, with often mutual influences between process A and B [87]. As a result, it is hard to maintain the temporal ordering of cause and effect and, indeed, a clear separation between them [122].

Furthermore, causation can never be established directly, just correlation [64]. When a correlation is highly stable, we are inclined to infer a causal link. Additional information

is then needed to assess the direction of the assumed causal link, as correlation indicates for association and not for causation [1]. For example, the motor cortex is always active when a movement is made, so we assume a causal link between the two phenomena. The anatomical and physiological properties of the motor cortex, and the timing of the two phenomena provide clues about the direction of causality (i.e. cortical activity causes the movement, and not the other way around). However, only intervention studies, such as delivering Transcranial Magnetic Simulation (TMS, [78]) pulses over the motor cortex or lesion studies, can confirm the causal link between the activity in the motor cortex and behavior.

Causal studies in fMRI are based on three types of correlations: correlating neuronal activity to 1) mental and behavioral phenomena; 2) physiological state (such as neurotransmitters, hormones, etc.), and 3) neuronal activity in other parts of the brain. In this review we will focus on the last field of research: establishing causal connections between two or more brain areas.

fMRI studies currently use a variety of algorithms to infer causal links [35, 134]. All these methods have different assumptions, advantages and disadvantages (see for instance [144, 139]). In a seminal study by Smith et al., popular approaches to causal processes were compared using synthetic data created with a Dynamic Causal Modeling (DCM, see below) generative model [42]. Surprisingly, most of the methods struggled to perform above chance level, even though the test networks were sparse and the noise levels introduced to the model were low compared to what one would expect in real recordings. This raises the question: given the characteristics of fMRI data (low temporal resolution, slow haemodynamics, low signal-to-noise ratio; see Section 2) and the fact that causal webs in the brain are likely dense and dynamic, is it in principle possible to investigate causality in the brain using MRI?

In this review, we discuss this question. First, we identify seven characteristics of models used to study causality. Then, we compare and contrast the popular approaches to the causal research in fMRI according to these criteria. Our list of features of causality is as follows:

1. *Sign of connections*: Can the algorithm distinguish between excitatory and inhibitory causal relations? In this context, we do not mean *synaptic* effects, but rather an overall driving or attenuating impact of the activity in one brain region on the activity in another region. Certain algorithms only detect the existence of causal influence from the BOLD responses, whereas others can distinguish between these distinct forms of influence.

2. *Strength of connections*: Can the algorithm distinguish between weak and strong connections, apart from indicating the directionality of connections at a certain confidence level?

3. *Bidirectionality*: Can the algorithm pick up bidirectional connections $X \leftrightarrow Y$, or only indicate the strongest of the two connections $X \rightarrow Y$ and $Y \rightarrow X$? Some algorithms do not *allow* for bidirectional relations, since they cannot deal with cycles in the network.

4. *Immediacy*: Does the algorithm specifically identify direct influences $X \rightarrow Y$, or does it pool across direct and indirect influences Z_i: $X \rightarrow Z_i \rightarrow Y$? While some methods aim to make this distinction, others highlight any influence $X \rightarrow Y$, whenever it is direct or not.

5. *Resilience to confounds*: Does the algorithm correct for possible spurious causal effects from a common source ($Z \rightarrow X$, $Z \rightarrow Y$, so we infer $X \rightarrow Y$ and/or $Y \rightarrow X$),

or other confounders? In general, confounding variables are an issue to all the methods for causal inference, especially when a given study is non-interventional [115], however different methods can suffer from these issues to a different extent.

6. *Type of inference*: Does the algorithm probe causality through classical hypothesis testing or through model comparison? Hypothesis-based algorithms will test a null hypothesis H_0 that there is no causal link between two variables, against a hypothesis H_1 that there is causal link between the two. In contrast, model-comparison-based methods do not have an explicit null hypothesis. Instead, evidence for a predefined set of models is computed. In particular cases, when the investigated network contains only a few nodes and the estimation procedure is computationally cheap, a search through all the connectivity patterns by means of model comparison is possible. In all the other cases, prior knowledge is necessary to select a subset of possible models for model comparison.

7. *Computational cost*: What is the computational complexity of the inference procedure? In the case of model comparison, the computational cost refers to the cost of finding the likelihood of a single model, as the range of possible models depends on the research question. This can lead to practical limitations based on computing power.

8. *Size of the network*: What sizes of network does the method allow for? Some methods are restricted in the number of nodes that it allows, for computational or interpretational reasons.

In the following chapters, the references to this 'causality list' will be marked in the text with lowercase indices.

With respect to assumptions made on the connectivity structure, the approaches discussed here can be divided into three main groups (Fig. 1). The first group comprises multivariate methods that search for directed graphs without imposing any particular structure onto the graph: Granger Causality [127], Transfer Entropy [90], Structural Equation Modeling [91] and Dynamic Causal Modeling [42]. These methods will be referred to as network-wise models throughout the manuscript. The second group of methods is also multivariate, but requires an additional assumption of acyclicity. Models in this group assume that information travels through the brain by feed-forward projections only. As a result, the network can always be represented by a Directed Acyclic Graph (DAG, [143]). Methods in this group include Linear Non-Gaussian Acyclic Models (LiNGAM, [131]) and Bayesian Nets [98], and will be referred to as hierarchical network-wise models throughout the manuscript. The last group of methods, referred to as pairwise methods, use a two-stage procedure: first, a map of nondirectional functional connections is rendered, and second, the directionality in each connection is assessed. Since these methods focus on pairwise connections rather than complete network architectures, they by definition do not impose network assumptions like acyclicity. Patel's tau [102] and Pairwise Likelihood Ratios [67][1] are members of this group.

[1] in this review, we do not include studying a coupling between brain region and the rest of the brain with relation to a particular cognitive task, The Psycho-Physiological Interactions (PPIs [41]), as we are only focused on the methods for assessing causal links within brain networks, and we do not include brain-behavior causal interactions

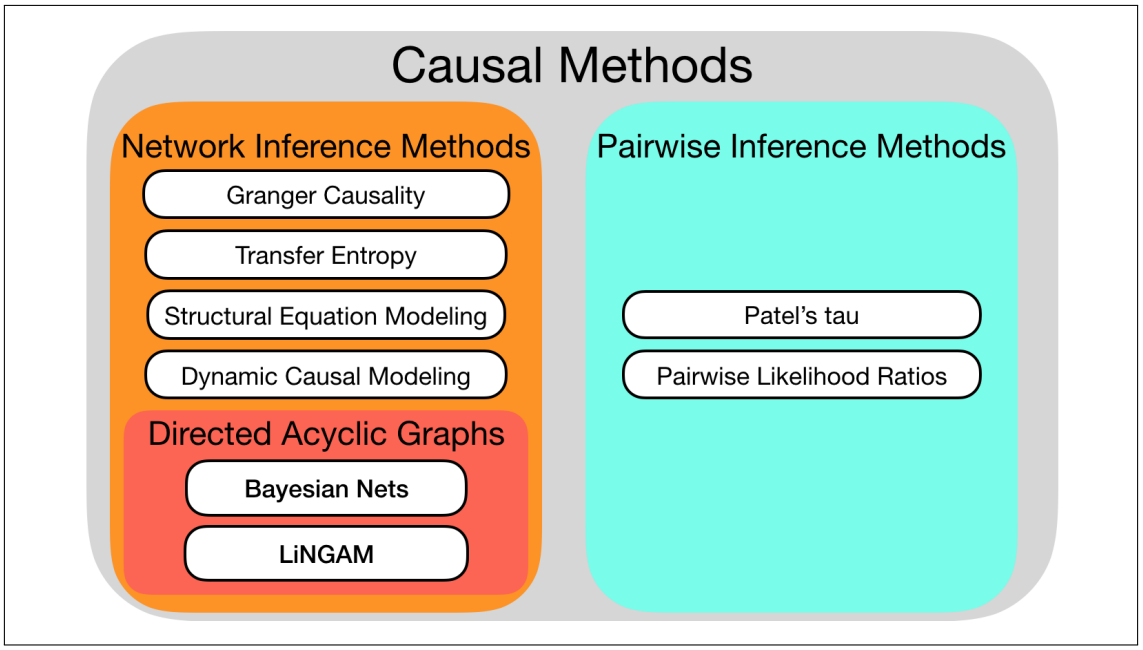

Figure 1: Causal research in functional Magnetic Resonance Imaging. The discussed methods can be divided into two families: Network Inference Methods, which are based on a one-step multivariate procedure, and Pairwise Inference Methods, which are based on a two-step pairwise inference procedures. As pairwise methods by definition establish causal connections on a node-by-node basis, the network as a whole cannot be guaranteed to be of any particular structure.

2 A note on the limitations of fMRI data

The characteristics of fMRI data impose severe restrictions on the possibility of finding causal relations using MRI. We discuss three of them.

2.1 Temporal resolution and haemodynamics

First, and best known, the temporal resolution of the image acquisition in MR imaging is generally restricted to a sampling rate $< 1[Hz]$. Recently, multiband fMRI protocols have gained in popularity [33], which increases the upper limit for the scanning frequency to up to $10[Hz]$, albeit at the cost of a severely decreased signal-to-noise ratio. However, no imaging protocol (including multiband imaging) can overcome the limitation of the recorded signal itself: the lagged change in blood oxygenation, which peaks $3-6[s]$ after neuronal firing in the adult human brain [2]. The haemodynamic response thus acts as a low-pass filter which results in high correlations between activity in consecutive frames [109]. Since the haemodynamic lags (understood as the peaks of the haemodynamic response) are region- and subject- specific [27], it is difficult to infer causality between two time series with potentially different haemodynamic lags [9].

2.2 Signal-to-noise ratio

Second, fMRI data is characterized by a relatively low signal-to-noise ratio. In grey matter, the recorded haemodynamic response changes by 1-2% at field strengths of $1.5-2.0[T]$ ([100, 12]), and by 5-6% at field strengths of $4.0[T]$. Moreover, typical fMRI protocols generate relatively short time series. For example, the new Human Connectome Project resting state datasets [31] do not contain more than a few hundred to maximally few

thousand samples. The short BOLD time series limits the option of improving signal-to-noise ratios through averaging across samples.

2.3 Caveats associated with region definition

Third, in order to propose a causal model, one first needs to define the nodes of the network. A single voxel does not represent a biologically meaningful portion of the brain [136]. Therefore, before attempting to establish causal connection in the network, one needs to integrate the BOLD time series over regions of interest (ROIs): groups of voxels that are assumed to share a common signal with a neuroscientific meaning. Choosing the optimal regions of interest for a study is a complex problem [105, 89, 142], and there are a few distinct strategies to address it. First, one can define regions of interest on the basis of brain anatomy. However, a consequence of this strategy could be that BOLD activity related to the cognitive process of interest will be mixed with other, unrelated activity within the ROIs. This is particularly likely to happen given that brain structure is not exactly replicable across individuals, so that a specific area cannot be defined reliably based on location alone. As indicated in the seminal computational study by Smith et al. [134], and also in a recent study by Bielczyk et al. [9], such signal mixing is detrimental to causal inference and causes all the existing methods for causal inference in fMRI to underperform. Second, parcellation into ROIs can be performed in a functional, data-driven fashion[2]. Another possibility to reduce the effect of mixing signals is to perform Principal Component Analysis (PCA, [72, 132]) and separate the BOLD time series within each anatomical region into a sum of orthogonal signals (eigenvariates) and choose only the signal with the highest contribution to the BOLD (the first eigenvariate, [42]), instead of averaging activity over full anatomical regions. Finally, one can build ROIs on the basis of patterns of activation only (task localizers [32, 60]). However, this approach cannot be applied to resting state research. In this work, we assume that the definition of ROIs has been performed by the researcher prior to the causal inference, and we do not discuss it any further.

3 Network-wise methods

The first group of models that we discuss in this review involves multivariate methods: methods that simultaneously assess all causal links in the network - specifically, Granger Causality, Transfer Entropy, Structural Equation Modeling and Dynamic Causal Modeling. These methods do not pose any constraints on the connectivity structure. Granger Causality, Transfer Entropy and Structural Equation Modeling infer causal processes through classical hypothesis testing. As there are no limits to the size of the analysed network, these methods allow for (relatively) hypothesis-free discovery. Dynamic Causal Modeling on the other hand, compares a number of predefined causal structures in networks of only a few nodes. As such, it requires a specific hypothesis based on prior knowledge.

3.1 Granger causality

Clive Granger introduced Granger Causality (GC) in the field of economics [51]. GC has found its way into many other disciplines, including fMRI research [114, 13, 127, 135]. GC is based on prediction [28]: the signal in a certain region is dependent on its past values. Therefore, a time series $Y(t)$ at time point t can be partly predicted by its past values

[2]There are multiple strategies for functional parcellation of the brain into ROIs: it can be implemented either through Principal Component Analysis [30], a hierarchical Independent Component Analysis known as Instantaneous Correlation Parcellation [146], by probabilistic clustering [71, 145, 5, 6, 11] or by the new, semi-automated classification technique by Glasser et al. [48]

$Y(t-i)$. A signal in an upstream region is followed by the same signal in a downstream region with a certain temporal lag. Therefore, if prediction of $Y(t)$ improves when past values of another signal $X(t-i)$ are taken into account, X is said to Granger-cause Y. Time series $X(t)$ and $Y(t)$ can be multivariate, therefore they will be further referred to as $\vec{X}(t)$, $\vec{Y}(t)$.

$Y(t)$ is represented as an *autoregressive process*: it is being predicted by a linear combination of its past states and a Gaussian noise[3]. This model is compared to a model including the past values of $X(t)$:

$$H_0 : \vec{Y}(t) = \sum_{i=1}^{N} \mathbf{B}_{yi} \vec{Y}(t-i) + \sigma(t) \tag{1}$$

$$H_1 : \vec{Y}(t) = \sum_{i=1}^{N} \mathbf{B}_{yi} \vec{Y}(t-i) + \sum_{i=1}^{N} \mathbf{B}_{xi} \vec{X}(t-i) + \vec{\sigma}(t) \tag{2}$$

Theoretically, this autoregressive (AR) model can take any order N (which can be optimized using, e.g., Bayesian Information Criterion [124]), but in fMRI research it is usually set to $N = 1$ [127], i.e. a lag that is equal to the repetition time (TR). The difference in explained variance between both models can be statistically tested, usually by means of an F-test or permutation tests.

By fitting the parameters of the AR model, which include the influence magnitudes $\mathbf{B}_{yi}, \mathbf{B}_{xi}$, the sign$_1$ as well as the strength$_2$ of the causal direction can be readily assessed with GC. Like all the methods in this chapter, GC does not impose any constraints on the network architecture and therefore can yield bidirectional connections$_3$. As a multivariate method, GC fits the whole connectivity structure at once. Therefore, ideally, it indicates the direct causal connections only$_4$, whereas the indirect connections should be captured only through higher order paths in the graph revealed in the GC analysis. However, this is not enforced directly by the method. In fact, in the original formulation of the problem by Granger, GC between X and Y works based on the assumption that the input of all the other variables in the environment potentially influencing X and Y has been removed [51]. In theory, this would provide resilience to confounds$_5$. However, in reality this assumption is most often not feasible in fMRI [53]. In a result, direct and indirect causality between X and Y are in fact pooled. In GC, the significance of results is achieved through classical hypothesis testing$_6$. Since the temporal resolution of fMRI is so low, only first order AR models with a time-lag equal to 1 TR are applicable. Therefore, there is no need to optimize either the temporal lag or the model order, and as such the computational cost of GC estimation procedure in fMRI is low$_7$. The AR model imposes a mathematical restriction on the size of the network though$_8$: the number of regions divided by the number of shifts can never exceed the number of time points (degrees of freedom).

The applicability of GC to fMRI data has been heavily debated [141]. Firstly, the application of GC requires certain additional assumptions such as signal stationarity[4], which does not always hold in fMRI data. Theoretical work by Seth et al. [128], and work by Roebroeck et al. [113], suggest that despite the limitations related to slow haemodynamics, GC is still informative about the directionality of causal links in the brain [127]. The face validity of GC analysis was also recently empirically validated using joint fMRI and MEG recordings [95], with the causal links inferred with GC matching the ground truth confirmed by MEG. On the other hand, recent experimental findings report that GC predominantly

[3]There is also an equivalent of GC in the frequency domain, spectral GC [46, 47], but this method will not be covered in this review

[4]Stationarity means that the joint probability distribution in the signal does not change over time. This also implies that mean, variance and other moments of the distribution of the samples in the signal do not change over time

identifies major arteries and veins as causal hubs [150]. This result can be associated with a regular pulsating behaviour with different phases in the arteries across the brain. This is a well-known effect and is even explicitly targeted with physiological noise estimates such as RETROICOR [49].

Another point of concern is the time lag in fMRI data, which restricts the possible scope of AR models that can be fit in the GC procedure. Successful implementations of GC in EEG/MEG research typically involve lags of less than a hundred milliseconds [61]. In contrast, for fMRI the minimal lag is one full TR, which is typically between $0.7[s]$ and $3.0[s]$ (although new acceleration protocols allow for further reduction of TR). What is more, the HRF may well vary across regions [56, 24], revealing apparent causal connections: when the HRF in one region is faster than in another, the temporal precedence of the peak will easily be mistaken for causation. The estimated directionality can in the worst case, even be reversed, when the region with the slower HRF in fact causes the faster one [9]. Furthermore, the BOLD signal might be non-invertible into the neuronal time series [127], which can affect GC analysis regardless whether it is performed on the BOLD time series or the deconvolved signal.

3.2 Transfer Entropy

Transfer Entropy (TE [121]) is another data-driven technique, equivalent to Granger Causality under Gaussian assumptions [3], and asymptotically equivalent to GC for general Markovian (non-linear, non-Gaussian) systems [**?**]. In other words, TE is a non-parametric form of GC (or, GC is a parametric form of TE). It was originally defined for pairwise analysis, and later extended to multivariate analysis [83, 96]. TE is based on the concept of *Shannon entropy* [129]. Shannon entropy $H(x)$ quantifies the information contained in a signal of unknown spectral properties as the amount of uncertainty, or unpredictability. For example, a binary signal that only gets values of 0 with a probability p, and values of 1 with a probability $1-p$, is most unpredictable when $p = 0.5$. This is because there is always exactly a 50% chance of correctly predicting the next sample. Therefore, being informed about the next sample in a binary signal of $p = 0.5$ reduces the amount of uncertainty to a higher extent than being informed about the next sample in a binary signal of, say, $p = 0.75$. This can be interpreted as a larger amount of information contained in the first signal as compared to the latter. The formula which quantifies the information content according to this rule reads as follows:

$$H(X) = -\sum_i P(x_i)\log_2 P(x_i) \tag{3}$$

where x_i are the possible values in the signal (for the binarized signal, there are only two possible values: 0 and 1).

TE builds up on the concept of Shannon entropy by extension to *conditional Shannon entropy*: it describes the amount of uncertainty reduced in future values of Y by knowing the past values of X along with the past values of Y:

$$TE_{X\to Y} = H(Y|Y_{t-\tau}) - H(Y|X_{t-\tau}, Y_{t-\tau}) \tag{4}$$

where τ denotes the time lag.

In theory, TE requires no assumptions about the properties of the data, not even signal stationarity although in most real-world applications, stationarity is required to almost the same extent as in GC. TE is model-free and as such it does need a priori definition of the causal process, and it may work for both linear and nonlinear interactions between the nodes.

TE can distinguish the signum of connections$_1$, as the drop in the Shannon entropy can be both positive and negative. Furthermore, the absolute value of the drop in the

Shannon entropy can provide a measure of the connection strength$_2$. TE can also distinguish bidirectional connections, as in this case, both $TE_{X\to Y}$ and $TE_{Y\to X}$ will be nonzero$_3$. However, the absolute value of the drop can provide a measure of the connection strength$_2$. Immediacy and resilience to confounds in TE depends on the implementation to a large extent: using a simple Pearson's correlation to compute functional connectivity increases the amount of spurious (indirect) connections, whereas partial correlation is meant to pick up on direct connections only$_5$. The inference in TE is performed through classical hypothesis testing$_6$ and is highly cost-efficient$_7$. As in GC, the maximum number of regions in the network divided by the number of shifts can never exceed the number of time points (degrees of freedom)$_8$.

TE is a straightforward and computationally cheap method [148]. However, it struggled when applied to synthetic fMRI benchmark datasets [134]. One reason for this could be the time lag embedded in the inference procedure, which is an obstacle to TE in fMRI research for the same reasons as for GC: it requires at least one full TR. TE is nevertheless gaining interest in the field of fMRI [130, 84, 101, 18, 96].

3.3 Structural Equation Modeling

Structural Equation Modeling (SEM, [91]) is a simplified version of Granger Causality. This method was originally applied to a few disciplines: economics, psychology and genetics [152], and was only recently adapted for fMRI research [91]. SEM can be considered a predecessor to Dynamic Causal Modelling [42]. SEM is used to study effective connectivity in cognitive paradigms, e.g., on motor coordination [79, 155], as well as in search for biomarkers of psychiatric disorders [120, 17]. It was also used for investigating heritability of large scale, resting state connectivity patterns [17].

The idea is to express every ROI time series in a network by a *linear combination* of all the time series (with the addition of noise), which implies no time lag in the communication. These signals are combined in a mixing matrix $\mathbf{B}$:

$$\vec{X}(t) = \mathbf{B}\vec{X}(t) + \vec{\sigma}(t) \tag{5}$$

where $\vec{\sigma}$ denotes the noise, and the assumption is that each univariate component $X_i(t)$ is a mixture of the remaining components $X_j(t)$, $j \neq i$. This is a simple multivariate regression equation, and the causal inference is based on search for the regression coefficients which correspond to the maximum likelihood (ML) solution: a set of model parameters $\mathbf{B}$ that gives the highest probability of the observed data. In case of SEM, we are looking for $\mathbf{B}$ parameters by minimizing the term

$$\|\vec{X} - \mathbf{B}\vec{X}\|^2 \tag{6}$$

Under the assumption of normality of the noise, there is a closed-form solution to this problem which gives the ML solution for parameters $\mathbf{B}$, referred to as Ordinary Least Squares (OLS) estimator for β [58]. In general, this estimator will give nonzero values to all parameters B.

While the Least Squares approximation gives connection strength estimates in β, it does not provide a measure of confidence. It therefore cannot contrast between a weak connection and no connection for a small $\mathbf{B}_{ij}$ value. This issue can be overcome in three ways. First, one can perform permutation testing[5]. Second, one can perform causal inference through model comparison: various models are fitted one by one, and the variance of the residual noise resulting from different model fits is compared, using either an F-test,

[5] generate a null distribution of $\mathbf{B}$ values by shuffling node labels across subjects and fitting $\mathbf{B}$ values to these shuffled datasets, and find confidence intervals on the basis of the desired probability of errors I and II

or goodness of fit (GFI [155]). Highly optimized software packages such as LiSREL [74] allow for an exploratory analysis with SEM by comparing millions of models against each other [70]. Third, one can fit the $\mathbf{B}$ matrix with new methods including regularization that forces sparsity of the solution [69], and therefore eliminates weak and noise-induced connections from the connectivity matrix.

Furthermore, in SEM applications to fMRI datasets, it is a common practice to establish the presence of connections with use of anatomical information derived, e.g., from Diffusion Tensor Imaging [107]. In that case, SEM inference focuses on estimating the strength of causal effects and not on identifying the causal structure.

SEM does not constrain the weight of connections, therefore it can retrieve both excitatory and inhibitory connections$_1$ as well as bidirectional connections$_3$. The connection coefficients $\mathbf{B}_{ij}$ can take any rational numbers and as such, they can reflect the strength of the connections$_2$. As with GC, SEM was designed to reflect direct connections$_4$: if regions X_i and X_j are connected only through a polysynaptic causal web, $\mathbf{B}_{ij}$ should come out as zero, and the polysynaptic connection should be retrievable from the path analysis. Again similar to GC, SEM is resilient to confounds only under the assumption that the model represents an isolated system, and all the relevant variables present in the environment are taken into account$_5$. Moreover, in order to obtain the maximum likelihood solution for $\mathbf{B}$ parameters, one needs to make a range of assumptions on the properties of the noise in the network. Typically, a Gaussian white noise is assumed, although background noise in the brain is most probably scale-free [59]. Inference can be performed either through the classical hypothesis testing (as the computationally cheap version) or through model comparison (as the computationally heavier version)$_{6,7}$.

In summary, SEM is a straightforward approach: it simplifies the causal inference by reducing the complex network with a low-pass filter at the output to a very simple linear system, but this simplicity comes at the cost of a number of assumptions.

3.4 Dynamic Causal Modeling

Both the aforementioned network-wise methods were developed in other disciplines, and only later applied to fMRI data. Yet, using prior knowledge about the properties of fMRI datasets can prove useful when searching for causal interactions. Dynamic Causal Modeling (DCM [42]) is a hypothesis testing tool which uses state space equations reflecting the structure of fMRI datasets. This technique was also implemented for other neural recording methods: EEG and MEG [76]). DCM is well received within the neuroimaging community (the original article by Friston et al. [42] gained over 2,700 citations at the time of submitting this manuscript).

In this work, we describe the original work by Friston et al. [39] because, despite multiple recent developments [77, 140, 88, 137, 82, 23, 125, 43, 57, 37, 110, 106, 36], it remains the most popular version of DCM in the fMRI community. The idea of DCM is as follows. First, one needs to build a generative forward model (Fig. 2). This model has two levels of description: the neuronal level, not directly observed in the experiment (Fig. 2, (iii)), and the haemodynamic level observed in the experiment (Fig. 2, (v)). This model reflects scientific evidence on how the BOLD response is generated from neuronal activity.

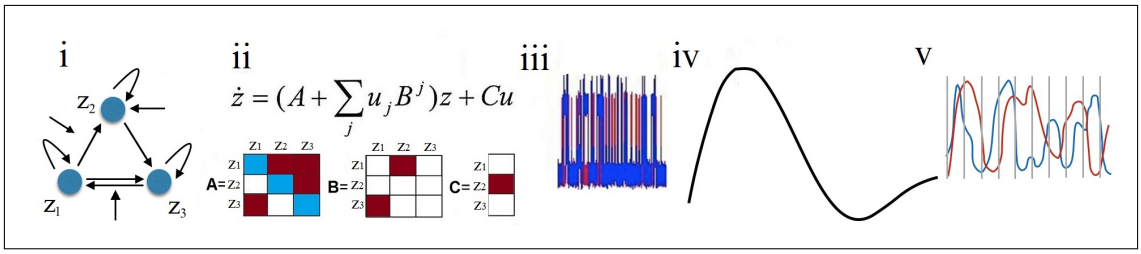

Figure 2: The full pipeline for the DCM forward model. The model involves three node network stimulated during the cognitive experiment **(i)**. The parameter set describing the dynamics in this network includes a fixed connectivity matrix (A), modulatory connections (B), and inputs to the nodes (C) **(ii)**. In the equation describing the fast neuronal dynamics, z denotes the dynamics in the nodes, and u is an experiment-related input. Red: excitatory connections. Blue: inhibitory connections. The dynamics in this network can be described with use of ordinary differential equations. The outcome is the fast neuronal dynamics **(iii)**. The neuronal time series is then convolved with the haemodynamic response function **(iv)** in order to obtain the BOLD response **(v)**, which may be then subsampled (vertical bars). This is the original, bilinear implementation of DCM [42]. Now, more complex versions of DCM with additional features are available, such as spectral DCM [43], stochastic DCM [23], nonlinear DCM [137], two-state DCM [88], large DCMs [125, 37] etc.

At the neuronal level of the DCM generative model, simple interactions between brain areas are posited, either bilinear [42] or nonlinear [137]. The haemodynamic level is more complex and follows the biologically informed Balloon-Windkessel model [16]. This model is also being iteratively updated based on new experimental findings, for instance to mimic adaptive decreases to sustained inputs during stimulation or the post-stimulus undershoot [57].

The Balloon-Windkessel model [16] describes the BOLD signal observed in fMRI experiments as a function of neuronal activity but also region-specific and subject-specific physiological features such as the time constant of signal decay, the rate of flow-dependent elimination, and the haemodynamic transit time or resting oxygen fraction. Effectively, this haemodynamic response is a convolution with a linear kernel (Fig. 2, (iv)) which typically peaks at $4-6[s]$ after the neuronal activity takes place, to match the lagged oxygen consumption in the neuronal tissue mentioned in Section 2.1.

In this paper, the deterministic, bilinear single node per region DCM will be described [42]. The DCM procedure starts with defining hypotheses based on observed activations, which involves defining which regions are included in the network (usually on the basis of activations found through the General Linear Model [40]) and then and defining a model space based on the hypotheses. In the latter model selection phase, a range of literature-informed connectivity patterns and inputs in the networks (referred to as 'models') are posited (Fig. 2, (i)). Subsequently, for every model one needs to set priors on the parameters of interest: connectivity strengths and input weights in the model (Fig. 2, (ii)) and the haemodynamic parameters. The priors for haemodynamic parameters are experimentally informed Gaussian distributions [42]. The priors for connectivity strengths are Gaussian probability distributions centered at zero[6] (which is often referred to as *conservative shrinkage* priors). The user usually does not need to specify the priors, as they are already implemented in the DCM algorithms.

Next, an iterative procedure is used to find the model evidence by minimizing a cost function, a so-called *negative free energy* [45]. A cost function is a goodness of the model,

[6]Except for self-connections, which are negative log-transformed so they are negative-constrained, and the prior is centered on log(0)

i.e. balance between fit and complexity (which penalizes for correlations between parameters, and for moving away from the prior distributions). During the iterative procedure, the prior probability distributions gradually shift their mean and standard deviation, and converge towards the final posterior distributions.

In DCM, causality is modeled as a set of upregulating or downregulating connections between nodes. During the inference procedure, conservative shrinkage priors can shift towards both positive and negative values, which can be interpreted as effective excitation or effective inhibition (except for self connections, which are always only negative[7], Fig. 2, (ii), connections denoted in blue)$_1$. During the inference procedure, the neural and hemodynamic parameters of all models postulated for model comparison are optimized$_2$. The models can contain both uni- and bidirectional connections [147, 15]$_3$. The estimated model evidence can then be compared$_6$. As such, the original DCM [42] is a hypothesis-testing tool working only through model comparison. However, now, linear version of DCM dedicated to exploratory research in large networks is also available [37]. Testing the immediacy$_4$ and resilience to confounds$_5$ in DCM is possible through creating separate models and comparing their evidence. For instance, one can compare the evidence for $X \rightarrow Y$ with evidence for $X \rightarrow Z \rightarrow Y$ in order to test whether or not the connection $X \rightarrow Y$ is direct or rather mediated by another region Z. Note that this strategy requires an explicit specification of the alternative models and it cannot take hidden causes into consideration[8]. However, including extra regions in order to increase resilience to confounds is not necessarily a good idea. Considering the potentially large number of fitted parameters per region[9], this may result in a combinatorial explosion. Also, models with different nodes are, strictly, not comparable in DCM [42]. Extending the models by adding additional nodes not only increases the computation time considerably$_7$ but also the risk of overfitting the data. The original DCM [42] is therefore restricted to small networks of a few nodes$_8$[10].

The proper application of DCM needs a substantial amount of expertise [138]. The regions of interest can be found in a data-driven fashion, through a preliminary classical General Linear Model analysis, but the model specification requires prior knowledge of the research problem [75]. Given the number of possible combinations of connections, even for a network as small as three to four regions, there are millions of possible models. Apart from very particular cases in sensory systems, the underlying connectivity of the human brain is largely unknown, which can make it hard to reduce the model space to only a few options. Therefore, the classic DCM cannot be used for exploratory research.

However, there are new approaches to DCM which allow more more exploratory take on the causal discovery in fMRI. Firstly, *family-wise* approaches [103] group large families of similar models together in order to test a particular hypothesis. For instance, one can compare the joint evidence behind all the possible models that contain connection $X \rightarrow Y$ with the joint evidence behind all the possible models that contain connection $Y \rightarrow X$ (Fig. 2, (i)). Secondly, in order to extend the scope of application of the DCM analysis to larger networks, recently the new, large-scale DCM framework for resting state fMRI has been proposed [111]. This framework uses the new, spectral DCM [43] designed for resting state fMRI and able to handle dozens of nodes in the network. Spectral DCM

[7] This self inhibition is mathematically motivated: the system characterizing the fast dynamics of the neuronal network must be stable, and this requires the diagonal terms of the adjacency matrix A (Fig. 2, (ii)) to be negative

[8] in this work, we refer to the original DCM implementation [42], but there are also implementations of DCM involving hidden states, such as [22]

[9] The minimum number of nodes per region is two hemodynamic parameters and one input/output to connect to the rest of network

[10] As mentioned previously, today, large DCMs dedicated to exploratory research in large networks are also available [125, 37]

is then combined with functional connectivity priors in order to estimate the effective connectivity in the large-scale resting state networks.

There are a few points that need particular attention when interpreting the results of the DCM analysis. First, the negative free energy is a trade-off between model fit and model complexity. This means that in case the data quality is poor and the DCM is not able to fit the data well to any of the prespecified models, it will only regard the number of parameters in the model. In that case, simpler models will be preferred over more complex ones. As such, it is hard to distinguish between model winning because it best fits the data, or because it contains few connections. Therefore, it is recommended to iterate the process of choosing the model space until the winning model is in the middle of the range in terms of complexity.

Second, within a winning model, not every connection is necessarily significant. For every connection, the Variational Bayes algorithm gives a posterior probability distribution, based on which a p-value can be computed. In case when some of these are not significant, no conclusion about these connections can be drawn and a simpler model without these connections is likely to win. Therefore, one should add this simpler version to the pool of models for model comparison, and recompute the modeling procedure. However, this simpler model is often not included in the analysis, nor is the significance of the individual connections always reported ([86, 62, e.g.]). In general, it is advisable to check the amount of explained variance at the end of the DCM analysis, which should determine whether or not the winning model explains significant portion of variance in the data.

The most popular implementation of the DCM estimation procedure is based on Variational Bayes (VB, [10]) which is a deterministic algorithm. Recently, also Markov-Chain Monte Carlo (MCMC, [10, 126] was implemented for DCM. MCMC is more computationally costly than VB because it is stochastic, but is also more likely to converge to the global rather than local minimum of the free energy landscape because its outcome is less dependent on the initial condition.

DCM is clearly tailored for fMRI and accounts for haemodynamic response. It is able to test explicit hypotheses, and has been found to produce highly reproducible results [123, 116, 7]. It has been proven to be reliable when directly compared against GC and SEM [104]. Also, the DCM procedure can provide complimentary information to GC [39]: GC models dependency among observed BOLD responses, whereas DCM models coupling among the hidden states generating observations. DCM seems to be equally effective as GC in certain circumstances, such as when the haemodynamic response function (HRF) is deconvolved from the data [24, 118, 117, 149]. On the other hand, its proper use requires knowledge on the biology and on the inference procedure. DCM also has limitations in terms of the size of the possible models, and it implies a danger of overfitting the data and of insufficient exploration of the model space in many studies. Therefore, it has also gained some critics over the years [85].

However, DCM was further developed into multiple procedures including more sophisticated generative models than the original model discussed here. The field of DCM research in fMRI is still growing [44]. The DCM generative model is continuously being updated, in terms of the structure of the forward model ([57], the estimation procedure[126]), and the scope of the possible applications [44].

4 Hierarchical network-wise models

The second group of methods consists of hierarchical network-wise models. These are also based on multivariate methods but with one additional constraint: the network can only include forward projections (and therefore, no closed cycles). Consequently, the resulting models have a hierarchical structure with feed forward distribution of information through

the network.

4.1 LiNGAM

The Linear Non-Gaussian Acyclic Model (LiNGAM, [131]) is an example of a data driven approach working under the assumption of acyclicity [143]. The model itself is simple: every time course within an ROI $X_i(t)$ is considered to be a linear combination of all other signals with no time lag:

$$\vec{X}(t) = \mathbf{B}\vec{X}(t) + \vec{\sigma}(t) \tag{7}$$

in which $\mathbf{B}$ denotes a matrix containing the connectivity weights, and $\vec{\sigma}$ denotes noise. The model is in principle the same as in SEM (Section 3.3), but the difference lies in the inference procedure: whereas in SEM, inference is based on minimizing the *variance* of the residual noise under the assumption of independence and Gaussianity, LiNGAM finds connections based on the *dependence* between residual noise components $\vec{\sigma}(t)$ and regressors $\vec{X}(t)$.

The rationale of this method is as follows. Let us assume that the network is noisy, and every time series within the network is associated with a background noise uncorrelated with the signal in that node. An example of such a mixture of signal with noise is given in Fig. 3A. Then, let us assume that $\hat{X}(t)$ - which is a mixture of signal $X(t)$ and noise $\sigma_X(t)$ - causes $Y(t)$. Then, as it cannot distinguish between the signal and the noise, Y becomes a function of both these components. $Y(t)$ is also associated with noise $\sigma_Y(t)$, however, as there is no causal link $Y \rightarrow X$, $X(t)$ is not dependent on the noise component $\sigma_Y(t)$. Therefore, if Y depends on the $\sigma_X(t)$ component, but X does not depend on the $\sigma_Y(t)$ component, one can infer projection $X \rightarrow Y$.

An example of such a simple, directed causal relationship between two variables is demonstrated in Fig. 3B: the relationship between age and length in a fish. If fish length is expressed in a function of fish age (upper panel), the residual noise in the dependent variable (length) is uncorrelated with the independent variable (age). Therefore, the noise variance is constant over a large range of fish age. On the contrary, once the variables are flipped and fish age becomes a function of fish length (lower panel), the noise variance becomes dependent on the independent variable (length) as it is small for small values of fish length and large for the large values of fish length. Therefore, the first causal model (fish age influencing fish length) is correct.

In applications to causal research in fMRI, the LiNGAM inference procedure is often accompanied by an Independent Component Analysis (ICA, [66]) as follows. The connectivity matrix $\mathbf{B}$ in Eq. 7 describes how signals in the network mix together. By convention, not $\mathbf{B}$ itself but a transformation of $\mathbf{B}$ into

$$\mathbf{A} = (1 - \mathbf{B})^{-1} \tag{8}$$

is used as a *mixing matrix* in the LiNGAM inference procedure. By using this mixing matrix $\mathbf{A}$, one can look at Eq. 7 in a different way:

$$\vec{X} = \mathbf{A}\vec{\sigma} \tag{9}$$

Now, the BOLD time course in the network $\vec{X}(t)$ can be represented as a mixture of independent *sources of noise* $\vec{\sigma}(t)$. This is the well known *cocktail party problem* and it was originally described in acoustics [14]: in a crowded room, a human ear registers a linear combination of the noises coming from multiple sources. In order to decode the components of this cacophony, the brain needs to perform a *blind source separation* [21]: to decompose the incoming sound into a linear mixture of independent sources of sounds. In the LiNGAM procedure, Independent Component Analysis (ICA, [66]) is used to approach

this issue. ICA assumes that the noise components $\vec{\sigma}$ are independent and have a *non-Gaussian distribution*, and finds these components as well as the mixing matrix $\mathbf{A}$ through dimensionality reduction with Principal Component Analysis [72, 132]. From this mixing matrix, one can in turn estimate the desired adjacency matrix $\mathbf{B}$ with use of Eq. 8.

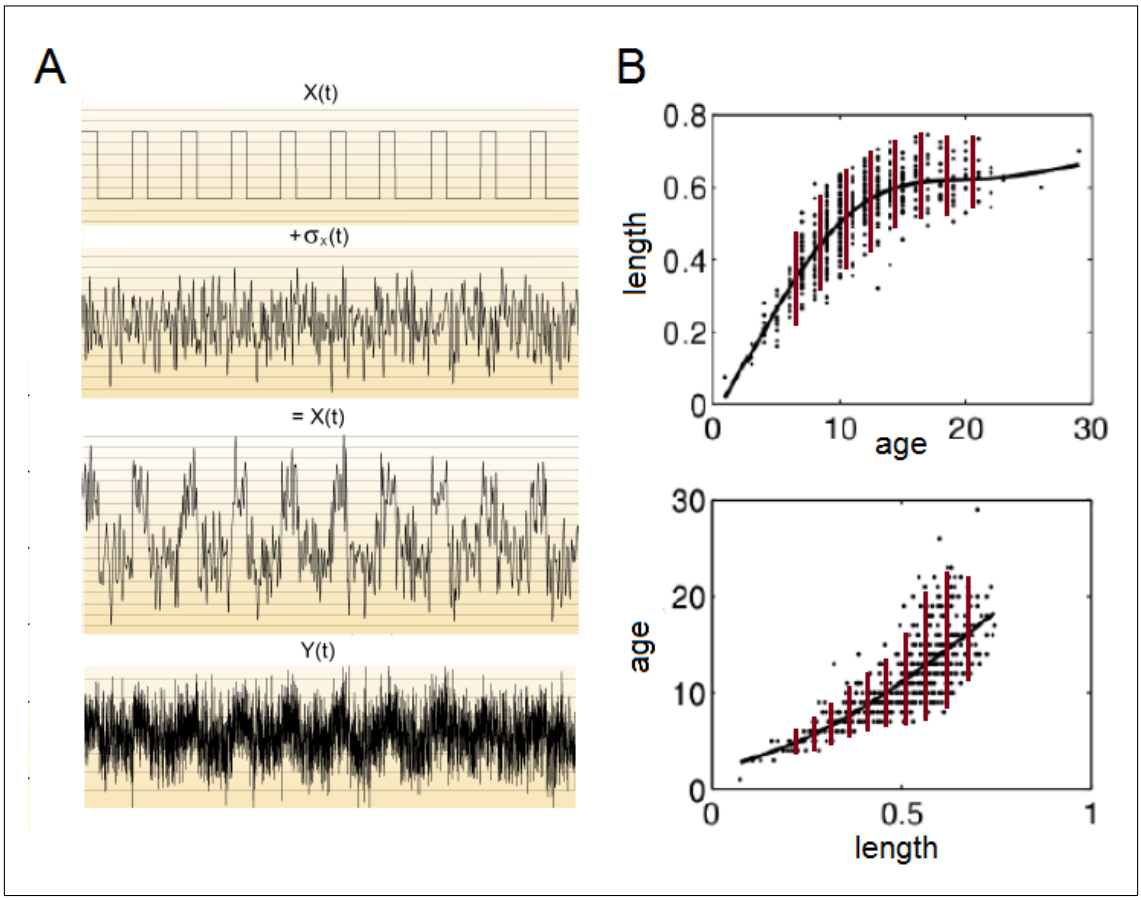

Figure 3: LiNGAM. **A**: The noisy time series $\hat{X}(t)$ consists of signal $X(t)$ and noise $\sigma_X(t)$. $Y(t)$ thus becomes a function of both the signal and the noise in $\hat{X}(t)$. **B**: Causal inference through the analysis of the noise residuals (figure reprinted from http://videolectures.net/bbci2014_grosse_wentrup_causal_inference/). The causal link from age to length in a population of fish can be inferred from the properties of the residual noise in the system. the relationship between age and length in a fish. If fish length is expressed in a function of fish age (upper panel), the residual noise in the dependent variable (length) is uncorrelated with the independent variable (age): the noise variance is constant over a large range of fish age (red bars). On the contrary, once the variables are flipped and fish age becomes a function of fish length (lower panel), the noise variance becomes dependent on the independent variable (length) as it is small for small values of fish length and large for the large values of fish length (red bars).

Since the entries $\mathbf{B}_{ij}$ of the connectivity matrix $\mathbf{B}$ can take any value, LiNGAM can in principle retrieve both excitatory and inhibitory connectivity$_1$ of any strength$_2$. However, LiNGAM makes the assumption of acyclicity, therefore only unidirectional connections can be picked up$_3$. Moreover, the connectivity matrix revealed with the use of LiNGAM is meant to pick up on direct connections$_4$. The original formulation of LiNGAM assumes no latent confounds [131], but the model can be extended to a framework that can capture the causal links even in the presence of (unknown) hidden confounds [63, 19]$_5$. LiNGAM-ICA's causal inference consists of ICA and a simple machine learning algorithm, and, as such,

it is a fully data driven strategy that does not involve model comparison$_6$. Confidence intervals for the connections $\mathbf{B}$ can be found through permutation testing. ICA itself can be computationally costly and its computational stability cannot be guaranteed (the procedure that searches for independent sources of noise can get stuck in a local minimum). Therefore, the computational cost in LiNGAM can vary depending on the dataset$_7$. This also sets a limit on the potential size of the causal network. When the number of connections approaches the number of time points (degrees of freedom), the fitting procedure will become increasingly unstable as it will be overfitting the data$_8$.

When tested on synthetic fMRI benchmark datasets [134], LiNGAM-ICA achieved a relatively good performance, but lower than a few other methods discussed in this paper, such as Patel's tau or GC. Interestingly, follow-up methods [108] for group analysis of effective connectivity based on LiNGAM, achieved performance very close to 100% on the same benchmark datasets, which suggests that, under the assumption that causal structures are similar between subjects, for any method for causality group analysis can substantially improve the causal discovery. Despite the promising results obtained for synthetic datasets, LiNGAM is still rarely applied to causal research in fMRI [153], mostly because it fits a SEM model with structural constraint and is therefore limited in terms of the potential applications as compared to SEM.

4.2 Bayesian nets

The use of the LiNGAM inference procedures assumes a linear mixing of signals underlying a causal interaction. Model-free methods do not make this assumption: the bare fact that one is likely to observe Y given the presence of X can indicate that the causal link $X \rightarrow Y$ exists. For instance, if event Y occurs in 80% of the cases when event X occurs (Fig. 4A), but the opposite is not true, the causal link $X \rightarrow Y$ is likely. Note that both model-based and model-free approaches are probabilistic, but in a different sense. In a model-based approach, a model is fitted to the data, and p-values associated with this fit are a measure of confidence that the modelled causal link exists (Fig. 4A, left panel). In contrast, in model-free approaches this confidence is quantified directly by quantifying causal relationships in terms of conditional probabilities (Fig. 4A, right panel). Bayesian Networks (BNs [38]) are based on such a model-free approach (Fig. 4B).

The causal inference in BNs is based on the concept of *conditional dependency*. Suppose that there are two events that could independently cause the grass to get wet: either a sprinkler, or rain. When one only observes the grass being wet, the direct cause for this event is unknown. However, once rain is observed, it becomes less likely that the sprinkler was used. Therefore, one can say that the variables X_1 (sprinkler) and X_2 (rain) are conditionally dependent given variable X_3 (wet grass), because X_1, X_2 become dependent on each other after information about X_3 is provided. In BNs, the assumption of conditional dependency in the network is used to compute the joint probability of a given model - i.e. the model evidence - as detailed in the following paragraph[11].

Implementing a probabilistic BN requires defining a model: choosing a graph of 'parents' who send information to their 'children'. For instance, in Fig. 4B (i), the node X_1 is a parent of nodes X_4 and X_5, and the node X_4 is a child of nodes X_1, X_2 and X_3. The joint probability of the model can then be computed as the product of all marginal probabilities of the parents and conditional probabilities of the children given the parents. *Marginal probability* $P(X_j)$ is the total probability that the variable of interest X_j occurs while disregarding the values of all the other variables in the system. For instance, in Fig. 4B (i), $P(X_1)$ means a marginal probability of X_1 happening in this experiment. *Conditional probability* $P(X_i|X_j)$ is the probability of a given variable (X_i) occurring given

[11] once variables X_i are conditionally dependent on X_j, the joint distribution $P(X_i, X_j)$ factorizes into a product of probabilities $P(X_j)P(X_i|X_j)$

that another variable has occurred (X_j). For instance, in Fig. 4B (i), $P(X_5|X_1, X_3)$ means a conditional probability of X_5 given its parents X_1 and X_3.

Then, once the whole graph is factorized into the chain of marginal and conditional probabilities, the *joint probability* of the model can be computed as the product of all marginal and conditional probabilities. For instance, in Fig. 4B(i), the joint probability of the model M yields

$$P(M) = P(X_1)P(X_2)P(X_3)P(X_4|X_1, X_2, X_3)P(X_5|X_1, X_3) \tag{10}$$

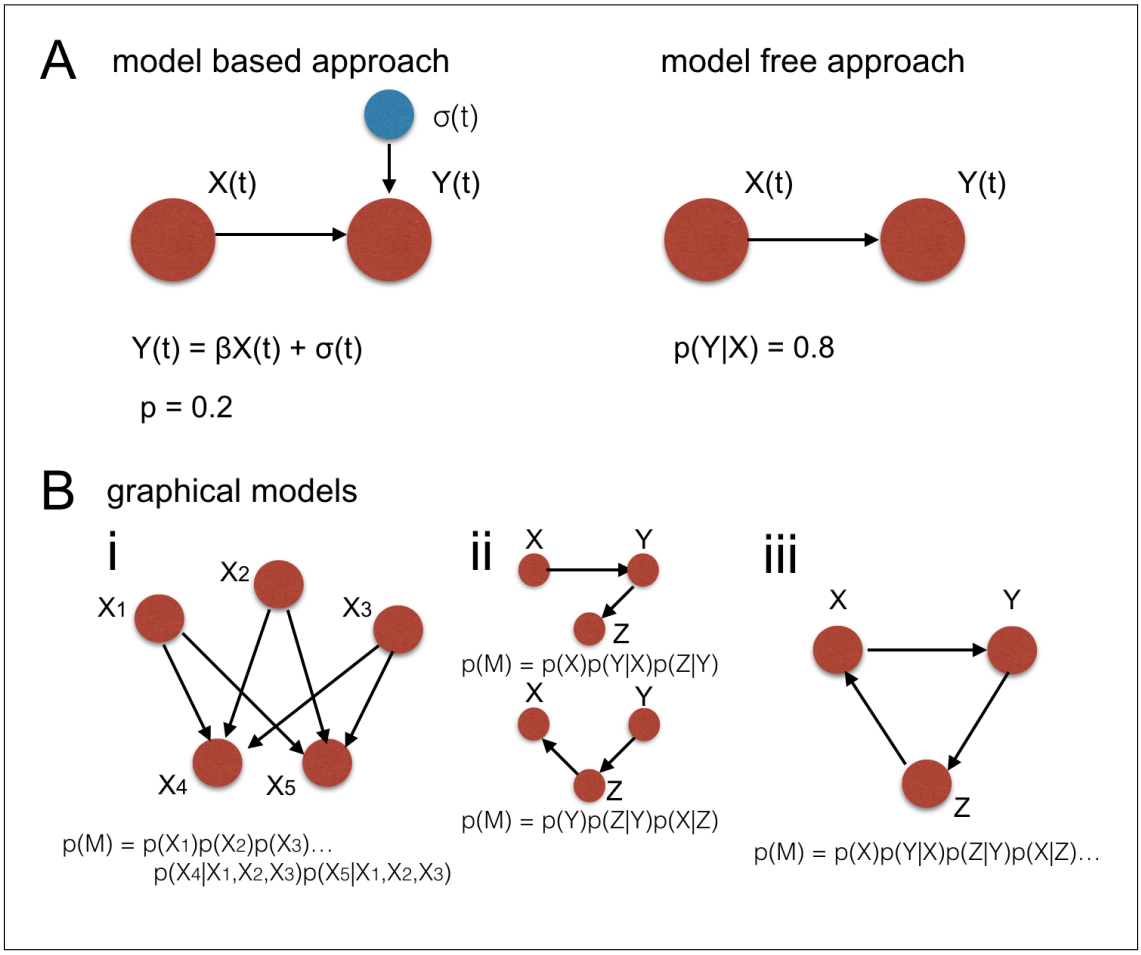

Figure 4: Bayesian nets. **A**: Model-based versus model-free approach. β: a regressor coefficient fitted in the modeling procedure. $\sigma(t)$: additive noise. Both model-based and model-free approach are probabilistic. In a model-based approach, a model is fitted to the data, and p-values associated with this fit are a measure of confidence that the causal link exists (left panel). In a model-free approach, this confidence is quantified directly by expressing causal relationships in terms of conditional probabilities (right panel). **B**: **(i)** an exemplary Bayesian net. X_1, X_2, X_3: parents, X_4, X_5: children. **(ii)** competitive Bayesian nets: one can define competitive models (causal structures) in the network and compare their joint probability derived from the data. **(iii)** cyclic belief propagation: if there was a cycle in the network, the expression for the joint probability would convert into an infinite series of conditional probabilities.

Finally, there are at least three possible approaches to causal inference with BNs:

1. model comparison: choosing the scope of possible models (by defining their structure a priori), and comparing their joint probability.

2. assuming one model structure a priori, and only inferring the weights. This is common practice, related to e.g. Naive Bayes [10] in which the structure is assumed, and the connectivity weights are estimated from conditional probabilities.

3. inferring the structure of the model from the data in an iterative way, by using a variety of approximate inference techniques that attempt to maximize posterior probability of the model by minimizing a cost function called free energy ([38], similar to DCM): expectation maximization (EM, [26, 10]), variational procedures [73], Gibbs sampling [99] or the sum-product algorithm [81] (which gives a broader selection of procedures than in the DCM)

BNs can detect both excitatory and inhibitory connections $X \rightarrow Y$, depending on whether the conditional probability $p(Y|X)$ is higher or lower than the marginal probability $p(X)_1$. Like LiNGAM, BNs cannot pick up on bidirectional connections. The assumption of acyclicity comes from the cyclic belief propagation (Fig. 4B, (iii)): the joint probability of a cyclic graph would be expressed by an infinite chain of conditional probabilities which usually does not converge into a closed form. This restricts the scope of possible models to Directed Acyclic Graphs (DAGs [143]), for instance 'Peter and Clark' [92] or Cyclic Causal Discovery [112], that can be used for causal discovery with BNs $_3$. The value of conditional probability $P(Y|X)$ can be a measure of a connection strength$_2$. In principle, BNs are not resilient to latent confounds, except for particular cases (e.g. Stimulus-based Causal Inference, SCI [54])$_5$. BNs can either work through model comparison or as an exploratory technique$_6$. In the first case, it involves model specification which - like in DCM - requires a priori knowledge about the experimental paradigm. In the latter case, the likelihood is intractable and can only be approximated$_7$ [29]. In principle, networks of any size can be modeled with BNs, either through a model comparison or through exploratory techniques. However, the exploratory techniques typically minimize a cost function during the iterative search for the best model. Since together with the growing network size, the landscape of the cost function becomes multidimensional and complex, the algorithm is more likely to fall into a local minimum$_8$.

BNs cope well with noisy datasets, which makes them an attractive option for causal research in fMRI [98]. They are, however, not widely used up to date, the main reason being the assumption of acyclicity. Constraint-Based Causal Inference (BCCD, [20]) is a promising new variation of the BN approach, designed for modeling causal interactions between any sets of variables. However, to date BCCD has not been tested specifically on (synthetic) fMRI datasets yet.

5 Pairwise inference

The last group of methods reflects the most recent trends in the field of causal inference in fMRI. These methods all involve a two-stage inference procedure. In the first step, functional connectivity is used to find connections, without assessing their directionality. Unlike network-wise methods which eliminate insignificant connections post-hoc, pairwise methods eliminate insignificant connections prior to causal inference. In the second step, each previously found connection is analyzed separately, and the two nodes involved are classified as an upstream or downstream region. These methods do not involve assumptions on the global patterns of connectivity at the network level (recurrent versus feed-forward). However, they involve the assumption that the connections are *non-transitive*: if X projects to Y, and Y projects to Z, it does not imply that X projects to Z. The causal inference is based on the pairs of nodes only, and this has consequences for the interpretation of the network as a whole. As there is uncertainty associated with estimation of every single

causal link, the probability that *all* connections are correctly estimated decreases rapidly with the number of nodes in the network.

5.1 Pairwise Likelihood Ratios

Another example of a model-free methodology is a two step procedure first proposed by Patel (as Patel's tau, PT [102]). The first step involves identifying the (undirected) connections by means of functional connectivity, and is achieved on the basis of correlations between the time series in different regions. This step results in a binary graph of connections, and the edges identified as empty are disregarded from further considerations, because if there is no correlation, there is no causation.

The second step determines the directionality in each one of the previously detected connections. The causal inference boils down to a two-node Bayesian network as the whole concept is based on a simple observation: if there is a causal link $X \to Y$, Y should get a transient boost of activity every time X increases activity. And vice versa: if there is a causal link $Y \to X$, X should react to the activation in Y by increasing activity. Therefore, one can threshold the signals $X(t)$, $Y(t)$, and compute the difference between conditional probabilities $P(Y|X)$ and $P(X|Y)$. Three scenarios are possible:

1. $P(Y|X)$ equals $P(X|Y)$: it is a bidirectional connection $X \leftrightarrow Y$ (since empty connections were sorted out in the previous step)
2. the difference between $P(Y|X)$ and $P(X|Y)$ is positive: the connection $X \to Y$ is likely
3. the difference between $P(Y|X)$ and $P(X|Y)$ is negative: the connection $Y \to X$ is likely

Recently, the Pairwise Likelihood Ratios methodology (PW-LR [67]) was proposed. It builds on the concept of PT. The authors improved on the second step of the inference by analytically deriving a classifier to distinguish between two causal models $X \to Y$ and $Y \to X$, which corresponds to the LiNGAM model for two variables. The authors compared the likelihood of these two competitive models derived under LiNGAM's assumptions [68], and provided with a cumulant based approximation to their ratio. In particular, the authors focused on the approximation of the likelihood ratios with third cumulant for variables X and Y, which is an asymmetry between first (the mean) and second (the variance) moment of the distributions of variables X and Y:

$$C_3 = \frac{1}{N}\sum_{i=1}^{N}(X(i)Y(i)^2 - X(i)^2Y(i)) \tag{11}$$

Then, if the value of this cumulant is positive, it indicates for the connection $X \to Y$, and backwards otherwise. Additionally, the authors proposed a modified version of the third cumulant, which contains a nonlinear transformation of the signal[12] PW-LR methods cannot distinguish between excitation and inhibition$_1$, but provide with a quantitative measure for the strength of the connection$_2$. Following the interpretation from Patel, it is possible to distinguish between uni- and bidirectionality (since scores close to zero might indicate the bidirectionality)$_3$. The authors proposed using partial correlation instead of Pearson's

[12] by pulling down the value of the samples with a high z-score, it is more resilient to the outliers in the signal and reads:

$$\begin{aligned} \mathrm{X}_R &= \log(\cosh(\max(\mathrm{X},0))) \\ C_{3R} &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(X_R(i)Y(i)^2 - X_R(i)Y(i)^2) \end{aligned} \tag{12}$$

correlation in the first step of the causal inference, which aims to find direct connections in the network$_4$. As for the resilience to confounds, PW-LR methods were tested on benchmark data for which common inputs to the nodes of the network were introduced ([134], simulation no 12). PW-LR gave much better performance than the best competitors (LiNGAM-ICA and PT) and reached as much as 84% of correctly classified connections across all the benchmark data$_5$. PW-LR works through the classical hypothesis testing, and the confidence levels for the connections can be obtained with permutation testing$_6$. It has a closed form solution and is therefore computationally cheap$_7$. As the pair-by-pair inferences do not require network fitting procedures, this can easily be applied to larger networks$_8$.

On the benchmark datasets, both versions of PW-LR were performing very well, as contrasted with the best competitors: PT and LiNGAM. In all but one out of 28 simulations PW-LR was performing highly above chance, and in a few cases they even reached 100% accuracy. However, PW-LR has never been validated on the real fMRI datasets.

6 New directions in causal research in fMRI

A number of methods have been discussed, but the search for new ways of extracting causal information from fMRI data is still on, of which we want to highlight two candidates. First of all, one can introduce more prior knowledge into the equation. This is done in laminar analysis, where the layered structure of the cortex is assumed to contain information about the signal. Another new option is a recently presented method based on fractional cumulants of the BOLD distribution [8], in which the statistical properties of the BOLD fMRI signal are used for inferring causal links.

6.1 Laminar analysis

Advancements in fMRI acquisition have made it possible to scan at submillimetre resolution, which opens up the possibility of a layer specific examination of the BOLD signal. As the different layers of the cortex receive and process feed forward and feedback information largely in different layers ([34, 4, e.g.], these different processes could be visible in the laminar BOLD response. In rat studies, the BOLD response was indeed shown to have laminar specificity and have its onset in the input layer of rat motor and somatosensory cortex [154]. And also in humans, several studies suggest laminar specificity of feedback processes [80, 97].

These results suggest that human laminar BOLD signal may contain directional and causal information. Hitherto, only single region laminar fMRI has been employed, but it may well be worthwhile to investigate how output layers of one region influence the input layer of the other.

6.2 Fractional cumulants

Certain newly developed methods take a more statistical approach to neuroimaging data. For instance, characterizing the shape of BOLD distributions by means of fractional moments of the BOLD distribution combined into cumulants [8] can improve classification of the two nodes in a connection into an upstream and a downstream node. These fractional moments of a distribution have limited practical interpretation, but could still contain valuable (causal) information.

The authors suggest to classify these momentum curves derived from BOLD distribution using predictions derived from the DCM generative model. The initial results show that the causal classification scores similarly or better than competitive methods when applied to the benchmark synthetic datasets [134], and better once the data is noisy or

has uneven Signal-to-noise ratios in different nodes of the network. However, validation on real fMRI data sets is still pending.

7 Summary

We sum up the characteristics of all the discussed methods in the following table:

Feature | Method	GC	SEM	DCM	LN	BN	TE	PL
group of methods	net	net	net	dag	dag	net	pw
sign of connections	+	+	+	+	-	+	-
directionality	+	+	+	-	-	+	+
connection strength	+	+	+	+	+	+	+
immediacy	+/-	+/-	-	+	+	+/-	+
resilience to confounds	+/-	+/-	-	+/-	+/-	+/-	+
causality through...	c	mc/c	mc	ml+c	mc/ml	c	c
computational cost	l	l/h	h	h	l/h	l	l
model-free?	-	-	-	-	+	+	+
prespecify the graph?	-	-	+	-	+/-	-	-
regression in time	+	-	-	-	-	+	-
promising due to [134]?	-	n/a	n/a	+	-	-	n/a

Table 1: Summary for all the methods discussed in this paper. *GC*: Granger causality, *SEM*: Structural Equation Modeling, *DC*: Dynamic Causal Modeling, *LN*: LINGaM, *BN*: Bayesian nets, *TE*: Transfer Entropy, *PL*: Pairwise Likelihoods, *net*: network-wise, *dag*: Directed Acyclic Graphs only, *pw*: pairwise, *+/-*: depends on implementation, *mc*: model comparison, *c*: classical hypothesis testing, *ml*: machine learning, *l*: low, *h*: high, *n/a*: non-applicable.

8 Discussion

In this review, we discussed the state-of-the-art in causal research in fMRI. In general, one can look at the methods from different angles, and classify them into different categories. For instance, some of the reviewed methods are based on the temporal sequence of the signals (e.g. GC or TE), others ignore the sequence in time, and solely focus on the statistical properties of the time series (e.g. BNs).

Based on the causal structure imposed on the brain on the other hand, the methods fall into three categories. Network-wise methods, such as GC or SEM, do not restrict the connectivity patterns whereas Directed Acyclic Graphs (DAGs), such as BNs, assume a hierarchical structure and unidirectional connections. In the latter category, a primary node receives input from outside the network and distributes information downstream throughout the network. This may be a good approximation for many processes, (see for instance recent work on the visual cortex [93]). However, the feed-forward structure assumes a strictly hierarchical organization, which limits its capacity to model communication between different brain networks. Under what circumstances DAGs can be an accurate representation for causal structures in the brain, remains an open question.

Next to network-wise methods and DAGs, we also discussed a third group of methods, referred to as 'pairwise'. In this approach, the causal inference is done by splitting the inference into many pairwise inferences. Prior to this, the dimensionality is reduced based

on functional connectivity, based on the idea that (partial) correlation is a good indicator for the existence of causal links [134] and therefore allows for simplifying the problem, both computationally and conceptually. Since the inference in this class of methods is split into a set of pairwise inferences, it is important to be aware of the fact that the confidence levels are also obtained connection by connection. Therefore, for a network represented by a set of connections with p-values p_i, the joint probability of the model is roughly $\Pi_i(1-p_i)$ [13]. This also means that there is a trade-off between the joint probability of the graph and its density: the joint probability of the whole network pattern can be increased by decreasing the threshold for connectivity at more conservative p-values. Furthermore, one can look at the pairwise inference methods as a sort of model comparison, because in the second step of the inference, for every connection only three options are possible to choose from. The difference with DCM procedure lies in the fact that pairwise inference methods are based on the simple statistical properties emerging from causation in linear systems, and do not involve minimizing the cost function — such as negative free energy — as is done in DCM.

In the fMRI community, the DCM family [42] is currently the most popular approach to causal inference. This is partially because DCM was tailor-made for fMRI, and includes a generative model based on the biological underpinnings of the BOLD dynamics [16]. Some of the GC studies also involve estimation of the haemodynamic response function, and deconvolving the data before applying the estimation procedure [24, 118, 117, 65, 151, 119, 50]. This notion of the haemodynamics is both a strength and a weakness: the generative model fits the data well, but only as long as the current state of knowledge is accurate. New studies suggest that human haemodynamics are very dynamic and driven by state-dependent processes [94, 55]. The influence of this complex behavior on the performance of DCM is hard to estimate.

The DCM procedure performs causal inference through model comparison, and as such, it is restricted to causal research in small networks containing a few nodes - since the computational costs increase like a factorial with the number of nodes. With the rise of research into resting state networks that contain up to 200 nodes, this may prove to be a limiting characteristic [133]. This issue can be addressed with new methods for pairwise inference such as PT and PW-LR, which do not impose any upper bound on the size of the network.

It is important to remember that there are always two aspects to a method for causal inference. First, the method should have assumptions grounded in a biologically plausible framework, well-suited for the given dataset. For instance, a method for causal inference in fMRI should respect: (1) the confounding, region- and subject-specific BOLD dynamics [56]; and (2) co-occurance of cause and effect (since the time resolution of the data is low compared to the underlying neuronal dynamics, the causes and their effects most likely happen within the same frame in the fMRI data). The new methods for pairwise inference address this issue by (1) breaking the time order, and performing causal inference on the basis of statistical properties of the distribution of the BOLD samples, and not from the timing of events; (2) using correlation in order to detect connections. A good counterexample here is GC. GC has been proven useful in multiple disciplines, and its estimation procedure is impeccable: nonparametric, computationally straightforward, and it gives a unique, unbiased solution. However, there is an ongoing discussion on whether or not GC is suited for causal interpretations of fMRI data. On the one hand, theoretical work by Seth et al. [128] and Roebroeck et al. [113] suggest that despite the slow haemodynamics, GC can still be informative about the directionality of causal links in the brain. On the other hand, the work by Webb [150] demonstrates that the spatial distribution of GC corresponds to the Circle of Willis, the major blood vessels in the brain.

[13] In practice, confidence values for the existence of single connections are not independent, therefore this is only a rough approximation of the joint probability

Secondly, an estimation procedure needs to be computationally stable. Even if the generative model faithfully describes the data, it still depends on the estimation algorithm whether the method will return correct results. DCM is a good example here: the generative model was based on detailed knowledge of the physiology of the brain, and as such, it is currently the best representation of BOLD fMRI as a function of network dynamics. However, the estimation procedure is stochastic, and the two available implementations of DCM - Variational Bayes and MCMC - both involve a trade-off between computational costs and the chance of convergence to a global minimum, and in both implementations, obtaining the global solution is not guaranteed.

For the two aforementioned reasons, it depends on the research question, which method should be used for the research purpose at hand.

Coming back to the main question posed in this review, can we hope to uncover causal relations in the brain using fMRI? Although there are new concepts in the field, which propose to consider causal interactions in the brain in probabilistic terms [87, 52], the 'traditional', deterministic models of causality are prevalent in neuroimaging. Within these deterministic models, in the light of the existing literature, the new research directions based on breaking the time order as the axiom of causal inference (such as PWLR, PT, and LiNGAM), prove more successful than the more 'traditional' approaches which take regression in time into account (such as GC or TE, [134, 67]). Also, Patel's two-step design to achieve a causal map of connections is very promising, especially once the Pearson correlation is replaced with partial correlation as is done in PW-LR. One note to add is that 'success' of any method for causal inference in fMRI depends on the forward model used for generating the synthetic dataset. In the seminal paper by Smith et al. we are referring to, [134], multiple methods were evaluated and critically discussed on the basis of simulations of the DCM generative model. However, there are alternatives, e.g., generative model by Seth et al. [128], which might potentially yield other hierarchy of methods in terms of success rate.

With the current rapid growth of translational research and increase in use of invasive and acute stimulation techniques such as optogenetics [25, 117] or TMS [78], a rigid validation of methodology for causal inference becomes feasible through interventional studies. Recently, multiple methods for inferring causality from fMRI data were validated using a joint fMRI and MEG experiment [95], with promising results for GC and BNs. This gives hope for establishing causal relations in neural networks, using fMRI.

Disclosure/Conflict-of-Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. JCG has acted as a consultant to Boehringer Ingelheim in the last 4 years, but is not an employee or shareholder of this company.

Author Contributions

NZB drafted the manuscript. NZB, SU, and TvM restructured the manuscript. NZB, SU, TvM, MNH and PA revised the work as a team. JKB and JCG critically revised the final manuscript.

Acknowledgements

We would like to thank to **Lionel Barnett, Christian Beckmann, Daniel Borek, Patrick Ebel, Daniel Gomez, Moritz Grosse-Wentrup, Max Hinne, Maciej Je-**

dynak, Christopher Keown, Sándor Kolumbán, Vinod Kumar, Randy McIntosh, Nils Müller, Hanneke den Ouden, Payam Piray, Thomas Rhys-Marshall, Gido Schoenmacker, Ghaith Tarawneh, Fabian Walocha and **Johannes Wilbertz** for sharing knowledge about causal inference in fMRI, and for providing a valuable content. We would like to further thank **Peter Vavra** for his contribution to the conceptual work. In addition, we would like to cordially thank **Thomas Wolfers** for encouragement and help at an early stage.

Funding

NB, MNH, JG and JB are supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 305697 (OPTIMISTIC), the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 278948 (TACTICS) and European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 603016 (MATRICS). SU was supported by grant 657605 of the Marie Sklodowska-Curie Horizon 2020 framework of the European Union.

References

[1] N. Altman and M. Krzywiński. Association, correlation and causation. *Nature Methods*, 12(10):899–900, 2015.

[2] T. Arichi, G. Fagiolo, M. Varela, A. Melendez-Calderon, A. Allievi, N. Merchant, N. Tusor, S. J. Counsell, E. Burdet, C. F. Beckmann, and A. D. Edwards. Development of BOLD signal hemodynamic responses in the human brain. *NeuroImage*, 63(2):663–73, 2012.

[3] L. Barnett, A. B. Barrett, and A. K. Seth. Granger causality and transfer entropy are equivalent for gaussian variables. 2009.

[4] A. M. Bastos, J. Vezoli, C. A. Bosman J.-M. Schoffelen, R. Oostenveld, J. R. Dowdall, P. De Weerd, H. Kennedy, and P. Fries. Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, 85(2):390–401, 2015.

[5] P. Bellec, V. Perlbarg, S. Jbabdi, M. Pélégrini-Issac, J. L. Anton, J. Doyon, and H. Benali. Identification of large-scale networks in the brain using fMRI. *NeuroImage*, 29(4):1231–43, 2006.

[6] P. Bellec, P. Rosa-Neto, O. C. Lyttelton, H. Benali, and A. C. Evans. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *NeuroImage*, 51(3):1126–39, 2010.

[7] D. Bernal-Casas, E. Balaguer-Ballester, M. F. Gerchen, S. Iglesias, H. Walter, A. Heinz, A. Meyer-Lindenberg, K. E. Stephan, and P. Kirsch. Multi-site reproducibility of prefrontal-hippocampal connectivity estimates by stochastic DCM. *NeuroImage*, 82:555–63, 2013.

[8] N. Z. Bielczyk, A. Llera, J. K. Buitelaar, J. C. Glennon, and C. F. Beckmann. Increasing robustness of pairwise methods for effective connectivity in Magnetic Resonance Imaging by using fractional moment series of BOLD signal distributions. *arXiV preprint*, 2016.

[9] N. Z. Bielczyk, A. Llera, Jan K. Buitelaar, J. C. Glennon, and C. F. Beckmann. The impact of haemodynamic variability and signal mixing on the identifiability of effective connectivity structures in BOLD fMRI. *Brain and Behavior*, 2017.

[10] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] T. Blumensath, S. Jbabdi, M. F. Glasser, D. C. Van Essen, K. Ugurbil, T. E. Behrens, and S. M. Smith. Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. *NeuroImage*, 76:313–24, 2013.

[12] J. L. Boxerman, P. A. Bandettini, K. K. Kwong, J. R. Baker, T. L. Davis, B. R. Rosen, and R. M. Weisskoff. The intravascular contribution to fMRI signal change: Monte Carlo modeling and diffusion-weighted studies in vivo. *Magnetic Resonance in Medicine*, 34(1):4–10, 1995.

[13] S. L. Bressler and A. K. Seth. Wiener-Granger causality: a well established methodology. *NeuroImage*, 58(2):323–9, 2011.

[14] A. W. Bronkhorst. The cocktail party phenomenon: A review on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86:117–28, 2000.

[15] A. W. G. Buijink, A. M. M. van der Stouwe, M. Broersma, S. Sharifi, P. F. C. Groot, J. D. Speelman, N. M. Maurits, and A.-F. van Rootselaar. Motor network disruption in essential tremor: a functional and effective connectivity study. *Brain*, 138(10):2934–47, 2015.

[16] R. B. Buxton, E. C. Wong, and L. R. Frank. Dynamics of blood flow and oxygenation changes during brain activation: the Balloon model. *Magnetic Resonance in Medicine*, 39(6):855–64, 1998.

[17] A. Carballedo, J. Scheuerecker, E. Meisenzahl, V. Schoepf, A. Bokde, H. J. Möller, M. Doyle, M. Wiesmann, and T. Frodl. Functional connectivity of emotional processing in depression. *Journal of Affective Disorders*, 134(1-3):272–9, 2011.

[18] B. Chai, D. Walther, D. Beck, and L. Fei-fei. Exploring functional connectivities of the human brain using multivariate information analysis. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 270–278. Curran Associates, Inc., 2009.

[19] Z. Chen and L. Chan. Causality in linear nongaussian acyclic models in the presence of latent gaussian confounders. *Neural Computation*, 25(6):1605–41, 2013.

[20] T. Claassen and T. Heskes. A Bayesian approach to constraint based causal inference. In *UAI, Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 2012.

[21] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

[22] J. Daunizeau, K. J. Friston, and S. J. Kiebel. Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D: Nonlinear Phenomena*, 238(21):2089–118, 2009.

[23] J. Daunizeau, K. E. Stephan, and K.J. Friston. Stochastic dynamic causal modelling of fMRI data: Should we care about neural noise? *NeuroImage*, 62(1):464–81, 2012.

[24] O. David, I. Guillemain, S. Saillet, S. Reyt, C. Deransart, C. Segebarth, and A. Depaulis. Identifying neural drivers with functional mri: An electrophysiological validation. *PLoS Biology*, 6(12):e315, 2008.

[25] K. Deisseroth. Optogenetics. *Nature Methods*, 8:26–9, 2011.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[27] I. M. Devonshire, N. G. Papadakis, M. Port, J. Berwick, A. J. Kennerley, J. E. Mayhew, and P. G. Overton. Neurovascular coupling is brain region-dependent. *NeuroImage*, 59(3):1997–2006, 2012.

[28] Francis X Diebold. *Elements of Forecasting (2nd ed.)*. Cincinnati: South Western, 2001.

[29] P. J. Diggle. Monte Carlo Methods of Inference for Implicit Statistical Models. *Journal of the Royal Statistical Society, Series B*, 46:193–227, 1984.

[30] A. M. DSouza, A. Z. Abidin, L. Leistritz, and A. Wismüller. Exploring connectivity with large-scale Granger causality in resting-state functional MRI. *Journal of Neuroscience Methods*, 287:68–79, 2017.

[31] D. C. Van Essen, S. M. Smith, D. M. Barch, T.E.J. Behrens, E. Yacoub, K. Ugurbil, and WU-Minn HCP Consortium. The Human Connectome Project: a data acquisition perspective. *NeuroImage*, 62(4):2222–31, 2013.

[32] E. Fedorenko, P.-J. Hsieh, A. Nieto-Castañón, S. Whitfield-Gabrieli, and N. Kanwisher. New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–94, 2010.

[33] D. A. Feinberg and K. Setsompop. Ultra-fast MRI of the human brain with simultaneous multi-slice imaging. *Journal of Magnetic Resonance*, 229:90–100, 2013.

[34] D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991.

[35] A. Fornito, A. Zalesky, and M. Breakspear. Graph analysis of the human connectome: Promise, progress, and pitfalls. *NeuroImage*, 80:426–44, 2013.

[36] S. Frässle, E. I. Lomakina ad A. Razi, K. J. Friston, J. M. Buhmann, and K. E. Stephan. Regression DCM for fMRI. *NeuroImage*, 155:406–21, 2017.

[37] S. Frässle, E. Lomakina-Rumyantseva, A. Razi, J. M. Buhmann, and K. J. Friston. Whole-brain Dynamic Causal Modeling of fMRI data. 2016.

[38] B. J. Frey and N. Jojic. A comparison of algorithms for inference and learning in probabilistic Graphical Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1392–416, 2005.

[39] K. Friston, R. Moran, and A. K. Seth. Analysing connectivity with Granger causality and dynamic causal modelling. *Current Opinion in Neurobiology*, 23(2):172–8, 2013.

[40] K. J. Friston, J. Ashburner, S. J. Kiebel, T. E. Nichols, and W. D. Penny. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, 2007.

[41] K. J. Friston, C. Buchel, G. R. Fink, J. Morris, E. Rolls, and R. Dolan. Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage*, 6(3):218–29, 1997.

[42] K. J. Friston, L. Harrison, and W. Penny. Dynamic Causal Modeling. *NeuroImage*, 19(4):1273–302, 2003.

[43] K. J. Friston, J. Kahan, B. Biswal, and A. Razi. A DCM for resting state fMRI. *NeuroImage*, 94:396–407, 2011.

[44] K. J. Friston, K. H. Preller, C. Mathys, H. Cagnan, J. Heinzle, A. Razi, and P. Zeidman. Dynamic causal modelling revisited. *NeuroImage*, S1053-8119(17):30156–8, 2017.

[45] K. J. Friston and K. E. Stephan. Free-energy and the brain. *Synthese*, 159(3):417–58, 2007.

[46] J. F. Geweke. Measurement of Linear Dependence and Feedback between Multiple Time Series. *Journal of the American Statistical Association*, 77(378):304–13, 1982.

[47] J. F. Geweke. Measures of Linear Dependence and Feedback between Multiple Time Series. *Journal of the American Statistical Association*, 79(388):907–15, 1984.

[48] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil K, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–8, 2016.

[49] G. H. Glover, T. Q. Li, and D. Ress. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magn Reson Med*, 44(1):162–167, Jul 2000.

[50] K. Goodyear, R. Parasuraman, S. Chernyak, P. Madhavan, G. Deshpande, and F. Krueger. Advice taking from humans and machines: An fMRI and effective connectivity study. *Frontiers in Human Neuroscience*, 4(10):542, 2016.

[51] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrika*, 37(3):424–38, 1969.

[52] J. D. Griffiths. Causal influence in neural systems: Reconciling mechanistic-reductionist and statistical perspectives. comment on "foundational perspectives on causality in large-scale brain networks" by m. mannino & s.l. bressler. *Physics of Life Reviews*, 15:130–2, 2015.

[53] M. Grosse-Wentrup. Lecture: An introduction to causal inference in neuroimaging. *Max Planck Institute for Intelligent Systems*, 2014.

[54] M. Grosse-Wentrup, D. Janzing, M. Siegel, and B. Schölkopf. Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. *NeuroImage*, 125:825–33, 2016.

[55] D. A. Handwerker, J. Gonzalez-Castillo, M. D'Esposito, and P. A. Bandettini. The continuing challenge of understanding and modeling hemodynamic variation in fMRI. *NeuroImage*, 62(2):1017–23, 2012.

[56] D. A. Handwerker, J. M. Ollinger, and M. D'Esposito. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, 21(4):1639–51, 2004.

[57] M. Havlicek, A. Roebroeck, K. Friston, A. Gardumi, D. Ivanov, and K. Uludag. Physiologically informed Dynamic Causal Modeling of fMRI data. *NeuroImage*, 122:355–72, 2015.

[58] F. Hayashi. *Econometrics*. Princeton University Press, 2000.

[59] B. Y. He. Scale-free brain activity: past, present, and future. *Trends in Cognitive Neurosciences*, 18(9):480–87, 2014.

[60] J. Heinzle, M. A. Wenzel, and J.-D. Haynes. Visuomotor functional network topology predicts upcoming tasks. *Journal of Neuroscience*, 32(29):9960–8, 2012.

[61] W. Hesse, E. Möller, M. Arnold, and B. Schack. The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies. *Journal of Neuroscience Methods*, 124(1):27–44, 2003.

[62] H. Hillebrandt, K. J. Friston, and S. J. Blakemore. Effective connectivity during animacy perception–dynamic causal modelling of human connectome project data. *Scientific Reports*, 4:6240, 2014.

[63] P. O. Hoyer, S. Shimizu, A. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–78, 2008.

[64] D. Hume. *Cause and Effect.* 1772.

[65] N. L. Hutcheson, K. R. Sreenivasan, G. Deshpande, M. A. Reid, J. Hadley, D. M. White, L. Ver Hoef, and A. C. Lahti. Effective connectivity during episodic memory retrieval in schizophrenia participants before and after antipsychotic medication. *Human Brain Mapping*, 36(4):1442–57, 2015.

[66] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

[67] A. Hyvärinen and S. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, 14(1):111–52, 2013.

[68] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity. *Journal of Machine Learning Research*, 11:1709–31, 2010.

[69] R. Jacobucci, K. J. Grimm, and J. J. McArdle. Regularized Structural Equation Modeling. *Structural Equation Modeling*, 23(4):555–66, 2016.

[70] G. James, M. Kelley, R. Craddock, P. Holtzheimer, B. Dunlop, C. Nemeroff, H. Mayberg, and X. Hu. Exploratory structural equation modeling of resting-state fMRI: applicability of group models to individual subjects. *NeuroImage*, 45(3):778–87, 2009.

[71] R. J. Janssen, P. Jylänki, R. P. Kessels, and M. A. van Gerven. Probabilistic model-based functional parcellation reveals a robust, fine-grained subdivision of the striatum. *NeuroImage*, S1053-8119(15):00589–3, 2015.

[72] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[73] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. *An Introduction to Variational Methods for Graphical Models*. Kluwer Academic, 1998.

[74] K. G. Joresköog and M. Van Thillo. Educational Testing Service, 1972.

[75] J. Kahan and T. Foltynie. Understanding DCM: ten simple rules for the clinician. *NeuroImage*, 83:542–9, 2013.

[76] S. J. Kiebel, M. I. Garrido, R. J. Moran, and K. J. Friston. Dynamic causal modelling for EEG and meg. *Cognitive Neurodynamics*, 2(2):121–36, 2008.

[77] S. J. Kiebel, S. Kloppel, N. Weiskopf, and K. J. Friston. Dynamic Causal Modeling: a generative model of slice timing in fMRI. *NeuroImage*, 34(4):1487–96, 2007.

[78] D. R. Kim, A. Pesiridou, and J. P. OReardon. Transcranial magnetic stimulation in the treatment of psychiatric disorders. *Current Psychiatry Reports*, 11(6):447–52, 2009.

[79] S. Kiyama, M. Kunimi, T. Iidaka, and T. Nakai. Distant functional connectivity for bimanual finger coordination declines with aging: an fMRI and SEM exploration. *Frontiers in Human Neuroscience*, 8:251, 2014.

[80] Peter Kok, LaurenJ. Bains, Tim vanMourik, DavidG. Norris, and FlorisP. de Lange. Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Current Biology*, 26(3):371–376, 2016.

[81] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[82] B. Li, J. Piriz, M. Mirrione, C. Chung, C. D. Proulx, D. Schulz, F. Henn, and R. Malinow. Synaptic potentiation onto habenula neurons in the learned helplessness model of depression. *Nature*, 470(7335):535–9, 2011.

[83] J. Lizier, M. Prokopenko, and A. Zomaya. Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 77(2):026110, 2008.

[84] J. T. Lizier, J. Heinzle, A. Horstmann, J. D. Haynes, and M. Prokopenko. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. *Journal of Computational Neuroscience*, 30(1):85–107, 2011.

[85] G. Lohmann, K. Erfurth, K. Muller, and R. Turner. Critical comments on dynamic causal modelling. *NeuroImage*, 59(3):2322–29, 2012.

[86] L. Ma, J. L. Steinberg, K. A. Cunningham ans S. D. Lane, J. M. Bjork, H. Neelakantan, A. E. Price, P. A. Narayana, T. R. Kosten, A. Bechara, and F. G. Moeller. Inhibitory behavioral control: A stochastic Dynamic Causal Modeling tudy comparing cocaine dependent subjects and controls. *NeuroImage Clinical*, 24:837–47, 2015.

[87] M. Mannino and S. L. Bressler. Foundational perspectives on causality in large-scale brain networks. *Physics of Life Reviews*, 15:107–23, 2015.

[88] A. C. Marreiros, S. J. Kiebel, and K. J. Friston. Dynamic causal modelling for fMRI: a two-state model. *NeuroImage*, 39(1):269–78, 2008.

[89] G. Marrelec and P. Fransson. Assessing the influence of different ROI selection strategies on functional connectivity analyses of fMRI data acquired during steady-state conditions. *PLoS One*, 6(4):e14788, 2011.

[90] G. Marrelec, A. Krainik, H. Duffau, M. Pélégrini-Issac, S. Lehéricy, J. Doyon, and H. Benali. Partial correlation for functional brain interactivity investigation in functional mri. *NeuroImage*, 32(1):228–37, 2006.

[91] A.R. Mclntosh and F. Gonzalez-Lima. Structural equation modeling and its application to network analysis in functional brain imaging. *Human Brain Mapping*, 2:2–22, 1994.

[92] C. Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence*, pages 403–10, 1995.

[93] G. Michalareas, J. Vezoli, S. van Pelt, J.-M. Schoffelen, H. Kennedy, and P. Fries. Alpha-beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas. *Neuron*, 89(2):384–97, 2016.

[94] F. M. Miezin, L. Maccotta, J. M. Ollinger, S. E. Petersen, and R. L. Buckner. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage*, 11(6):735–59, 2000.

[95] R. D. Mill, A. Bagic, A. Bostan, W. Schneider, and M. W. Cole. Empirical validation of directed functional connectivity. *NeuroImage*, 146:275–87, 2017.

[96] A. Montalto, L. Faes, and D. Marinazzo. Mute: A matlab toolbox to compare established and novel estimators of the multivariate transfer entropy. *PLoS One*, 9(10):e109462, 2014.

[97] Lars Muckli, Federico DeÂMartino, Luca Vizioli, LucyÂS. Petro, FraserÂW. Smith, Kamil Ugurbil, Rainer Goebel, and Essa Yacoub. Contextual feedback to superficial layers of v1. *Current Biology*, 25(20):2690–2695, 2015.

[98] J. A. Mumford and J. D. Ramsey. Bayesian networks for fMRI: a primer. *NeuroImage*, 86:573–82, 2014.

[99] R. M. Neal. Probabilistic inference using Markov Chain Monte Carlo methods. Technical report, Dept. of Computer Science, University of Toronto, 1993.

[100] S. Ogawa, R. S. Menon, D. W. Tank, S. G. Kim, H. Merkle, J. M. Ellermann, and K. Ugurbil. Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging. a comparison of signal characteristics with a biophysical model. *Biophysics Journal*, 64(3):803–12, 1993.

[101] D. Ostwald and A. P. Bagshaw. Information theoretic approaches to functional neuroimaging. *Magn Reson Imaging*, 29(10):1417–28, 2011.

[102] R.S. Patel, F. Dubois Bowman, and J.K. Rilling. A Bayesian approach to determining connectivity of the human brain. *Human Brain Mapping*, 27(3):267–76, 2006.

[103] W. D. Penny, K. E. Stephan, J. Daunizeau, M. J. Rosa, K. J. Friston, and T. M. Schofield et al. Comparing families of dynamic causal models. *PLoS Computational Biology*, 6(3):e1000709, 2010.

[104] W.D. Penny, K.E. Stephan, A. Mechelli, and K.J. Friston. Modelling functional integration: a comparison of structural equation and dynamic causal models. *NeuroImage*, 23(S1):264–74, 2004.

[105] R. A. Poldrack. Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience*, 2(1):67–70, 2007.

[106] G. Prando, M. Zorzi, A. Bertoldo, and A. Chiuso. Estimating effective connectivity in linear brain network models. *arXiV preprint*, 2017.

[107] A. B. Protzner and A. R. McIntosh. Testing effective connectivity changes with structural equation modeling: what does a bad model tell us? *Human Brain Mapping*, 27(12):935–47, 2006.

[108] J. D. Ramsey, S. J. Hanson, and C. Glymour. Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the smith et al. simulation study. *NeuroImage*, 58(3):838–48, 2011.

[109] J. D. Ramsey, S. J. Hanson, C. Hanson, Y. O. Halchenko, R.A. Poldrack, and C. Glymour. Six problems for causal inference from fMRI. *NeuroImage*, 49(2):1545–58, 2010.

[110] A. Razi and K. J. Friston. The connected brain: Causality, models, and intrinsic dynamics. *IEEE Signal Processing Magazine*, 33(3):14–35, 2016.

[111] A. Razi, M. L. Seghier, Y. Zhou, P. McColgan, P. Zeidman, H.-J. Park, O. Sporns, G. Rees, and K. J. Friston. Large-scale DCMs for resting state fMRI. *Network Neuroscience*, pages 1–41, 2017.

[112] T. Richardson and P. Spirtes. *Automated discovery of linear feedback models.* C MIT Press, 2001.

[113] A. Roebroeck, E. Formisano, and R. Goebel. Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage*, 25(1):230–42, 2005.

[114] A. Roebroeck, A. K. Seth, and P. Valdes-Sosa. Causal time series analysis of functional magnetic resonance imaging data. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 12:65–94, 2011.

[115] J. M. Rohrer. Clarifying the Confusion Surrounding Correlations, Statistical Control and Causation. *PsyArXiv preprint*, 2017.

[116] J.B. Rowe, L.E. Hughes, R.A. Barker, and A.M. Owen. Dynamic causal modelling of effective connectivity from fMRI: Are results reproducible and sensitive to parkinson's disease and its treatment? *NeuroImage*, 52(3):1015–26, 2010.

[117] S. Ryali, Y. Y. Shih, T. Chen, J. Kochalka, D. Albaugh, Z. Fang, K. Supekar, J. H. Lee, and V. Menon. Combining optogenetic stimulation and fMRI to validate a multivariate dynamical systems model for estimating causal brain interactions. *NeuroImage*, 132:398–405, 2016.

[118] S. Ryali, K. Supekar, T. Chen, and V. Menon. Multivariate dynamical systems models for estimating causal interactions in fMRI. *NeuroImage*, 54(2):807–23, 2011.

[119] K. Sathian, G. Deshpande, and R. Stilla. Neural changes with tactile learning reflect decision-level reweighting of perceptual readout. *Journal of Neuroscience*, 33(12):5387–98, 2013.

[120] R. Schlösser, T. Gesierich, B. Kaufmann, G. Vucurevic, S. Hunsche, J. Gawehn, and P. Stoeter. Altered effective connectivity during working memory performance in schizophrenia: a study with fMRI and structural equation modeling. *NeuroImage*, 19(3):751–63, 2003.

[121] T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–4, 2000.

[122] A. Schurger and S. Uithol. Nowhere and everywhere: the causal origin of voluntary action. *Review of Philosophy and Psychology*, 6(4):761–78, 2015.

[123] B. Schuyler, J. M. Ollinger, T. R. Oakes, T. Johnstone, and R. J. Davidson. Dynamic Causal Modeling applied to fMRI data shows high reliability. *NeuroImage*, 49(1):603–11, 2010.

[124] G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–4, 1978.

[125] M. L. Seghier and K. J. Friston. Network discovery with large DCMs. *NeuroImage*, 68:181–91, 2013.

[126] B. Sengupta, K. J. Friston, and W. D. Penny. Gradient-free mcmc methods for dynamic causal modelling. *NeuroImage*, 112:375–81, 2015.

[127] A. K. Seth, A. B. Barrett, and L. Barnett. Granger Causality Analysis in Neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–7, 2015.

[128] A. K. Seth, P. Chorley, and L. C. Barnett. Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling. *NeuroImage*, 65:540–55, 2013.

[129] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(4):623–56, 1948.

[130] M. Sharaev, V. Ushakov, and B. Velichkovsky. pages 213–18. Springer International Publishing, 2016.

[131] S. Shimizu, P. O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–30, 2006.

[132] J. Shlens. A tutorial on principal component analysis. 2014.

[133] S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. A. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, and C. F. Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106(31):13040–5, 2009.

[134] S.M. Smith, K.L. Miller, G. Salimi-Khorshidi, M. Webster, C.F. Beckmann, T.E. Nichols, J.D. Ramsey, and M.W. Woolrich. Network modelling methods for fMRI. *NeuroImage*, 54(2):875–91, 2011.

[135] V. Solo. State-space analysis of Granger-geweke causality measures with application to fMRI. *Neural Computation*, 28(5):914–49, 2016.

[136] M. L. Stanley, M. N. Moussa, B. M. Paolini, R. G. Lyday, J. H. Burdette, and P. J. Laurienti. Defining nodes in complex brain networks. *Frontiers in Computational Neuroscience*, 7:169, 2013.

[137] K. E. Stephan, L. Kasper, L.M. Harrison, J. Deaunizeau, H. E. M. van den Ouden, M. Breakspear, and K. J. Friston. Nonlinear dynamic causal models for fMRI. *NeuroImage*, 42(2):649–62, 2008.

[138] K. E. Stephan, W. D. Penny, R. J. Moran, H. E. den Ouden, J. Daunizeau, and K. J. Friston. Ten simple rules for Dynamic Causal Modeling. *NeuroImage*, 49(4):3099–109, 2010.

[139] K. E. Stephan and A. Roebroeck. A short history of causal modeling of fMRI data. *NeuroImage*, 62(2):856–63, 2012.

[140] K. E. Stephan, N. Weiskopf, P. M. Drysdale, P. A. Robinson, and K. J. Friston. Comparing hemodynamic models with DCM. *NeuroImage*, 38(3):387–401, 2007.

[141] P. A. Stokes and P. L. Purdon. A study of problems encountered in Granger causality analysis from a neuroscience perspective. *Proceedings of the National Academy of Sciences*, 2017.

[142] B. Thirion, G. Varoquaux, E. Dohmatob, and J. B. Poline. Which fMRI clustering gives good brain parcellations? *Front Neurosci*, 8:167, 2014.

[143] K. Thulasiraman and M. N. S. Swamy. *Directed Acyclic Graphs*. John Wiley and Son, 1992.

[144] P. A. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston. Effective connectivity: influence, causality and biophysical modeling. *NeuroImage*, 58(2):339–61, 2011.

[145] M. van den Heuvel, R. Mandl, and R. Hulshoff Pol. Normalized cut group clustering of resting-state fMRI data. *PLoS One*, 3(4):e2001, 2008.

[146] E. S. B. van Oort, M. Mennes, T. Navarro Schröder, V. J. Kumar, N. I. Zaragoza Jimenez, W. Grodd, C. F. Doeller, and C. F. Beckmann. Functional parcellation using time courses of instantaneous connectivity. *NeuroImage*, 2017.

[147] A. E. Vaudano, P. Avanzini, L. Tassi, A. Ruggieri, G. Cantalupo, F. Benuzzi, P. Nichelli, L. Lemieux, and S. Meletti. Causality within the epileptic network: an EEG-fMRI study validated by intracranial EEG. *Frontiers in Neurology*, 14(4):185, 2013.

[148] R. Vicente, M. Wibral, M. Lindner, and G. Pipa. Transfer entropy–a model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, 30(1):45–67, 2011.

[149] Y. Wang, S. Katwal, B. Rogers, J. Gore, and G. Deshpande. Experimental validation of dynamic Granger causality for inferring stimulus-evoked sub-100ms timing differences from fMRI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, PP(99), 2016.

[150] J. T. Webb, M. A. Ferguson, J. A. Nielsen, and J. S. Anderson. BOLD Granger causality reflects vascular anatomy. *PLoS One*, page 8:e84279, 2013.

[151] M. D. Wheelock, K. R. Sreenivasan, K. H. Wood, L. W. Ver Hoef, G. Deshpande, and D. C. Knight. Threat-related learning relies on distinct dorsal prefrontal cortex network connectivity. *NeuroImage*, 102(2):904–12, 2014.

[152] S. Wright. The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences*, 6(6):320–32, 1920.

[153] L. Xu, T. Fan, X. Wu, K. Chen, X. Guo, J. Zhang, and L. Yao. A pooling-LiNGAM algorithm for effective connectivity analysis of fMRI data. *Front Comput Neurosci*, 8:125, 2014.

[154] Xin Yu, Chunqi Qian, Der-yow Chen, Stephen J Dodd, and Alan P Koretsky. Deciphering laminar-specific neural inputs with line-scanning fMRI. *Nature methods*, 11(1):55–58, 2014.

[155] J. Zhuang, S. LaConte, S. Peltier, K. Zhang, and X. Hu. Connectivity exploration with structural equation modeling: an fMRI study of bimanual motor coordination. *NeuroImage*, 25(2):462–70, 2005.