The lure of causal statements: Rampant mis-inference of causality in estimated connectivity

David Marc Anton Mehler[1, 2] and Konrad Paul Kording[3, 4]

[1] Cardiff University Brain Research imaging Centre (CUBRIC), School of Psychology, and MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff CF244HQ, UK

[2] Department of Psychiatry, University of Muenster, Germany

[3] Department of Bioengineering and Department of Neuroscience, University of Pennsylvania, USA

[4] Canadian Institute for advanced Research (CIFAR)

Abstract:

As neuroscientists we want to understand how causal interactions or mechanisms within the brain give rise to perception, cognition, and behavior. It is typical to estimate interaction effects from measured activity using statistical techniques such as functional connectivity, Granger Causality, or information flow, whose outcomes are often falsely treated as revealing mechanistic insight. Since these statistical techniques fit models to low-dimensional measurements from brains, they ignore the fact that brain activity is high-dimensional. Here we focus on the obvious confound of common inputs: the countless unobserved variables likely have more influence than the few observed ones. Any given observed correlation can be explained by an infinite set of causal models that take into account the unobserved variables. Therefore, correlations within massively undersampled measurements tell us little about mechanisms. We argue that these mis-inferences of causality from correlation are augmented by an implicit redefinition of words that suggest mechanisms, such as connectivity, causality, and flow.

The state of the estimated connectivity field

A fundamental goal in neuroscience is to understand brain mechanisms that underlie perception, cognition, and behavior, which, arguably, requires understanding the causal interactions between neurons and neuronal populations. Whenever we talk about causal interactions in this opinion paper, we use the counterfactual definition. A variable causally influences another variable if a perturbation of the first variable would induce a change in the activity of another variable (Pearl, 2009a). This nicely approximates what we mean as neuroscientists: if we say a neuron influences another neuron we mean that perturbing the first (say electrically) would affect the second and if we say that a region influences another we mean that perturbing the first region (say magnetically) would affect the second. Causality has a perfectly clean definition (counterfactuals or perturbation) and we should demand our statistical approaches to be measured against it (although see also Gomez-Marin, 2017).

Because we cannot directly measure these interactions, statistical techniques are used that aim to infer interactions from simultaneously recorded brain signals. This often boils down to asking the question of how does neural population A mechanistically affect population B? The approaches that we call estimated connectivity (eC) in this paper convert measured signals into a statistical estimate of "connectivity." The results are typically (implicitly or explicitly) thought of as a measure or at least approximation of the causal strength of interactions. A rich body of literature has described eC techniques: some techniques utilize granger causality (Bressler and Seth, 2011), other techniques are called functional connectivity and look at delayed correlations (Friston et al., 1997). Yet other techniques talk about *information flow* (Babiloni et al., 2005; Honey et al., 2007). Another class more directly talks about Dynamic *Causal* Modeling (Daunizeau et al., 2011). Within the imaging community the term *effective connectivity* (EC) is often used when causality is explicitly claimed but within our definition they fall into our more expansive definition of eC. We will argue that the used observational approaches share the same logical weakness – statistical confounding. Yet, the lure of extracting causality from observational data, is so powerful, that we cannot avoid feeling the pull of it and have effectively referred to correlations in connectivity terms (Stevenson et al., 2008).

The problem of unobserved confounding is the existence of unobserved variables that affect the observed variables in a way that often makes the estimation of connectivity impossible. It is easy to see why it is impossible to use any statistical techniques to estimate causal interactions in the presence of unobserved confounding. Let us say there is a causal influence from A to B. In this case, it is always possible to construct a common input C which will produce the same effect on B that A would have (e.g. by replicating A and feeding into C). Similarly, if

there is no influence from A to B it is always possible to use a confounder to change A and B so that they now look like they do interact, according to any algorithm used. These issues would maybe be a minor problem if we had reasons to believe that confounding is weak, i.e. if there were not many orders of magnitude more confounders than measured variables. It is thus logically impossible that an algorithm could identify the network of interactions.

Techniques used for estimating connectivity typically come from fields that focus on forecasting or describing time-series. For example, Granger Causality (Bressler and Seth, 2011) and Coherence analysis (Bastos and Schoffelen, 2016; Sun et al., 2004) have been developed to describe the relationship between signals. They were then translated to address biological questions such as “How does neural population A interact with neural population B?” (Bressler and Seth, 2011; Friston et al., 2003, 1997). In many cases, scientists simply analyze the correlations and take high correlations to be a sign of neuronal *communication*. Alternatively, delayed or nonlinear correlations are used to estimate the direction of *interactions.* However, in biological questions we usually seek causal mechanisms and not just good predictions. Hence, researchers increasingly use, without much discussion of threats from confounding, techniques that were used to describe the relationships between signals to make claims about causality.

Indeed, many scientists have been developing techniques for estimating the strength of connections between neurons based on simultaneous spike recordings. The underlying idea is that we want to predict each neuron’s spiking probability based on the activity of other observed neurons (Pillow et al., 2008). And indeed, if we record all neurons, they are noisy, and we know that from their transfer function we should be able to estimate the strength and nature of causal interactions (Karbasi et al., 2018). There has been ample speculation about the meaning of the results of such a study, but it is frequently interpreted in causal terms (Pillow et al., 2008; Stevenson et al., 2008).

Another group of scientists has been developing techniques for estimating the strength of interactions between brain areas using neuroimaging techniques such as Electroencephalography (EEG), Magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI). Estimated connectivity (eC) in neuroimaging has been intensively studied with data acquired during tasks, as well as rest (resting-state connectivity). Within the eC community one can delineate two different philosophies. Some authors use eC to describe replicable network properties of functional and anatomical neural data without attributing causal significance to interregional correlations (Raichle et al., 2001; Smith et al., 2009; Yeo et al., 2011). Others typically interpret changes in parameter estimates in more explicitly causal terms such contributing factors of pathophysiological processes (Hacker et al., 2012; Karlsgodt et al., 2008;

Pawela et al., 2010; Wu et al., 2009), or as neuroplasticity in the functional re-organization of the brain in neurological or psychiatric conditions (Hallam et al., 2018; Jin et al., 2011; Lueken et al., 2013).

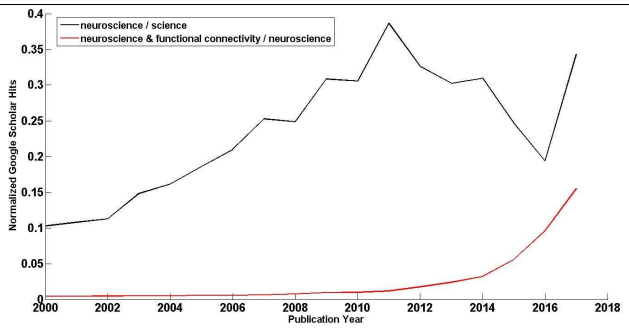

Figure 1: Functional connectivity is becoming extremely popular. The popularity of neuroscience and within it the popularity of talking about functional connectivity.

In search for a better understanding of brain network organization in health and disease, articles increasingly talk about eC. According to our Google Scholar search (May 14th 2018) for literature published in English language between 2000 and 2017 (excluding patents and citations), over 15% of articles that included the term "neuroscience" also mentioned "functional connectivity" (Figure 1; code and data available on: https://osf.io/9cs8p/). FC is a huge area in neuroscience making it important to obtain clarity about its interpretation and usefulness for potential future translational applications.

It is not just semantics?

The eC measures have in common that they are functions of (sometimes generalized) correlations. However, the language used generally does not reflect that. Almost every term in popular use suggests causality. For example, we use terms like Granger *Causality*, functional *connectivity,* information *flow*, *effective connectivity*, dynamic *causal* models, etc. We do this despite the fact that other terms such as *improvement in predictive power*, *correlations*, *conditional correlations*, and *model comparison* can denote more precisely what we actually do. To deal with the disagreement between used correlational techniques and desired mechanistic or causal statements, we are effectively redefining the English language. Connectivity implies a connection between two places. Causality implies cause and effect. Flow implies that something moves from one place to another. Effective implies that something has an effect. This set of re-definitions gives rise to the problem that eC approaches are often misunderstood.

Scientists write about connections within the brain minimizing the wiring length along which signals need to travel (Bullmore and Bassett, 2011; Bullmore and Sporns, 2009), but while the brain may want to minimize the length of its wires, there is no implication that it pays any price for correlations. They write about stimulation to control the network (Taylor et al., 2015), which

requires interactions to be causal. They write about interference to cure diseases (Khambhati et al., 2016), which again requires causality. Or they write about regions that "cause more exchange of causal information" (Bajaj et al., 2015). These examples show how correlations are assumed to indicate causality. Specifically, FC is often used as if it did reveal an approximate understanding of causality, and much of it is due to misleading use of words that imply causality in lay English and merely refer to algorithm properties in statistics.

For example, we ourselves wrote in 2008 " […] [estimated connectivity] methods have become staples of neural data analysis, and have revealed a great deal about the interactions between cortical and subcortical structures." (Stevenson et al., 2008). We could simply have said that models that use other activities as independent variables are good predictors. With LFPs it was argued that "[…] the relative weight of feedforward and lateral inputs in visual cortex is not fixed, but rather depends on stimulus contrast." (Nauhaus et al., 2009). For EEG and MEG power coherence analyses, it was advocated that "[…] amplitude correlation is an informative index of the large-scale cortical interactions that mediate cognition." (Siegel et al., 2012). For fMRI task data, other authors reported "[…] changes in the architecture of functional connectivity patterns that promote learning from initial training through mastery of a simple motor skill." (Bassett et al., 2015). For resting-state fMRI data, others claimed that FC may close the knowledge gap of "[…] the neuronal mechanisms that operate during and early after practice and during sleep to support motor memory consolidation." (Dimyan and Cohen, 2011). However, all these approaches only reach level one in Pearl's hierarchy of causation (Pearl and Mackenzie, 2018). While the impossibility of making causal statements in such situations is well known (Dawid, 2008; Holland, 2015), the field regularly makes causal statements based on statistical analyses that cannot support such statements. Altogether, we conclude that this debate is not merely about semantics and second that "whereof one cannot speak, thereof one must be silent" (Wittgenstein, 1922).

We do not actually learn about causality from estimated connectivity

Here we ask if the eC techniques used should be expected to estimate causal interactions. We ask if the techniques measure causality, connectivity, or flow, in the standard meanings of these words. We will conclude that, due to massive confounding, they merely describe the statistics of the joint neural data without convincing causal insights.

It is important to first observe the way *neurostatisticians* describe their results, themselves. When entering the field, we ourselves stated "unobserved common input is a potential confound" (Stevenson et al., 2008). Daunizeau et al. stated that "the "missing region"" – i.e. underdetermination problem – presents "a potential deep methodological confound" (Daunizeau

et al., 2011). And Anil Seth, a pioneer of Granger Causality approaches to neuroscience explicitly states on Twitter “I am NOT saying that Granger Causality (GC) indicates causality in the colloquial sense of “if I did A then B would happen” (Seth, 2018). These examples illustrate how statistics minded scientists tend to feature a paragraph that discusses how causality cannot be meaningfully estimated.

A growing field in the domain of statistics and econometrics is concerned with causal discovery. They ask how observational data can be used to answer causal questions. There are multiple philosophies. Some base their studies on fitting modern techniques such as directed acyclic Bayesian graphs that represent a set of variables and their conditional dependencies (Pearl, 2009a). Others focus on matching approaches that can account retrospectively for randomization (Rubin, 2010). Importantly though, both have in common that they fail if there are important variables that are unmeasured and affect the variables of interest. The effect induced by other variables on the apparent statistical relation between variables of interest is called confounding. Within observational causal inference there is an understanding that there can be no solution if each pair of measured variables shares an unobserved confounder (Pearl, 2009b).

A lot of the discussion in the field of functional connectivity is about statistics. And indeed, statistically estimating causal interactions is a hard problem (Spirtes and Zhang, 2016). Over time, advocates of functional connectivity research have introduced several statistical innovations, including sophisticated generative models for spiking (Havlicek et al., 2017; Paninski et al., 2008), priors about neuronal interactions (Calabrese et al., 2011; Stevenson et al., 2009), and priors about the network of interactions (Bassett et al., 2006). And, indeed, the statistics of network inference is complex, reflected by a progressively sophisticated field. However, we argue here that it is impossible to overcome the problem of massive confounding by using statistics.

We want to dwell on the idea of confounders to functional connectivity. Let us say we are interested in the interactions between two neurons. Further, let us assume that there is a third neuron (confounder) that activates both neurons. In that case, if we see a correlation between the neurons we cannot know if it is due to the neurons interacting directly or through an induced interaction by the unobserved neuron. More generally, any joint distribution between two neurons could be induced by a single third neuron. Unobserved neurons or information can act as confounders for the inference of functional connectivity, threatening the validity of the results. This confounding problem is central to the statistical field of *causal inference* and it is generally acknowledged that, in typical situations, unobserved confounders render inferences about causality impossible.

We want to use the Simpson's paradox to highlight the threat of confounding (Simpson, 1951), in which unobserved interaction can either cancel out main effects, or artificially induce spurious main effects. Let us say that there are two brain states (e.g. related to two levels of attention), and that one state is associated with high activity of neuron A and low activity of neuron B while another is associated with low activity of neuron A and high activity of neuron B. But let us say that there is a positive instantaneous causal influence from one on the other. If we do not know the brain state (confounded) we may conclude that neuron A has a negative influence on the activity of neuron B. But if we do know the brain state we correctly see that A increases the activity of neuron B (Figure 2). This paradox shows how confounders with relevant structure can arbitrarily influence the resulting functional connectivity.

Indeed, in certain neuroimaging datasets it is possible to directly observe the Simpson's paradox. For instance, the choice of analysis parameters, such as seed regions, and the network they belong to, may mediate whether different FC methods yield similar, or orthogonal results (Roberts et al., 2016). This illustrates that results from FC analyses depend on a series of analysis and design decisions, rather than a *true* underlying, biologically meaningful statistical relationship. It seems that even the underlying brain architecture – interindividual differences in shape and location of neuroanatomical structures – can confound FC modeling results (Bijsterbosch et al., 2018). Datasets from fMRI and MEG are further affected by more obvious confounders including head motion and physiological processes such as the heartbeat and breathing (Driver et al., 2016; Messaritaki et al., 2017; Murphy et al., 2013) that add structured noise to the signal. In response, the field has developed some effective data correction techniques (Ciric et al., 2017; Murphy et al., 2013; Parkes et al., 2017; Power et al., 2014; R et al., 2018). However, for each such obvious external controllable confounder there are countless internal unobservable ones (DiDomenico and Eaton, 1988).

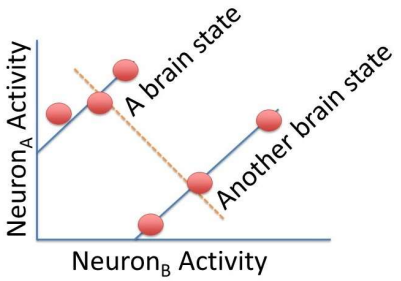

Figure 2: Simpson's paradox. The yellow line indicates a negative correlation between activity of Neuron A and B, which is confounded by two different brain states. But within each state there is a positive correlation. This paradox shows that the sign of a correlation is not indicative of causality.

Massive confounding destroys the causal interpretation of functional connectivity

Our argument about the impossibility to obtain causality from functional connectivity does not come from data – there is too little ground truth about complex brains to make this possible (Jonas and Kording, 2017). It rests on a simple consideration of the factors that are known to make causal

inference theoretically impossible. It rests on the justified assumption that we record only few of the neurons when recording spikes, or a few projections of neural activities in imaging. We have no convincing reasons to assume that the observed dimensions should be more important than the unobserved. Our argument thus boils down to a simple logical statement: if the bulk of causal interactions happen from and within unobserved dimensions, then the correlations between observed variables are simply epiphenomena. Correlation is not causation, regardless the mathematical sophistication we use when calculating it. Causal inference algorithms that work with observational data are generally built on the assumption of causal sufficiency, which boils down to there being no unobserved confounders (although see Ranganath and Perotte, 2018; Wang and Blei, 2018). Without these assumptions we can at best produce families of potential models and if any pair of recorded variables is confounded then this family will contain all models (Spirtes et al., 2001). Recording only few variables in a densely interacting causal system generally renders causal inference impossible (Jonas Peters et al., 2017; Pearl, 2009a).

When analyzing spike data, there are far more unobserved variables than observed variables. When we record a few hundred neurons (Stevenson and Kording, 2011), the number of recorded neurons is a vanishingly small subset of all neurons. We have little reason to assume that the recorded neurons are much more important than the un-recorded neurons. As each neuron receives inputs from so many other un-recorded neurons, we should expect that the parts of neural activity driven by unobserved neurons are arbitrarily larger than the parts coming from observed neurons. In other words, the confounding signal should be many orders of magnitude more important than those coming from observed data. As such, we should not expect that causal inference is possible.

When analyzing imaging data such as fMRI, or LFP, EEG, or MEG, there are also far more unobserved variables than observed variables. Within each signal source, we can, in some abstraction, observe the sum of neural activity. But the same measured activity can be realized by any combination of individual activities rendering a solution of the inverse problem (signals → neuronal spike trains) infeasible. The activity of neurons which are orthogonal to our signal, can span arbitrary dimensions, related to movement, memory, thought or neuronal communication. Importantly, dense physiological recordings in small areas suggest that countless variables are represented (e.g. movement related signals in V1; Musall et al., 2018; Stringer et al., 2018). The signals that we can measure are arbitrarily low-dimensional relative to the full dimensionality of communication in the brain. As such we are still in the situation where we have a number of confounders that is many orders of magnitude larger than the number of measured variables. This again puts us into the domain where causal inference should be impossible.

We might feel that causality may happen at a given scale, rendering the argument about unobserved dimensions invalid and allowing a multi-scale definition of interactions. The argument given is often the analogy to statistical physics: while understanding the interaction between gas molecules is hopeless, a large set of atoms can be well understood in terms of temperature and pressure. However, this analogy quickly breaks down. Every neuron is special, they do not interact with random neurons but with a largely fixed set. The justification of averaging over molecules is often perfectly fine in statistical physics. There is no evidence this logic would work in neuroscience. The separation of scales that sounds so meaningful outside of the brain makes little sense within.

As the impossibility of causal inference from subsampled data feels counter-intuitive we want to spell out the problem a bit more. Let us assume we are interested in the connections between two signals, e.g. voxels B and C. But the brain's real dynamics is characterized by the activity of all neurons ($\mathbf{x}_t$). Let us assume, for simplicities sake, that dynamics are linear:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \text{noise} \quad (1)$$

where matrix **A** describes the true dynamic of the system. The activities in our original voxels of interest, B and C (conveniently concatenated into a vector **y** which is actually observable) can be obtained using a projection matrix **M** that projects the activity of all neurons into a two-dimensional space:

$$\mathbf{y}_t = \mathbf{M}\mathbf{x}_t. \quad (2)$$

A neuroscientist may then want to use the time lagged correlation $\mathbf{R} = \langle \mathbf{y}_t \mathbf{y}_{t-1}^t \rangle$ in the measured signal **y** to gain insights into the properties of **A**. What will we measure then? We can now insert the definition of **x** and **y** and obtain

$$\mathbf{R} = \mathbf{MAVM}^t \quad (3)$$

where V is the covariance matrix of x. The question now is what we can learn about **A** from **R**. We can obtain some intuition in special cases. If **V** was the identity matrix and hence all neurons would fire entirely independent of one another then **R** would reveal the average influence of brain area B on C, only that then the connections would have to all be zero. However, in reality, the autocorrelation function in every brain area (the local **V**) is known to have a broad spectrum of singular values. Moreover, **V** between brain areas will be nonzero. Here, confounding becomes obvious, elements of **A** that neither relate to signals A or B can in arbitrary ways affect the correlation **V** between the signals (see code in supplementary material for simulations). Importantly, if N is the number of neurons and K the number of measured signals (2 in the example) there is an $N^2 - K^2$ manifold of **A** matrices that produce an identical correlation matrix.

As the number of neurons is typically much larger than the number of measured signals, the measurements do not really lower the dimensionality of the space of potential models ($N^2 - K^2 \approx N^2$). We note that while in this example we have computed a time-delayed correlation matrix as a metric of eC, also different methods (e.g. in Granger causality) essentially suffer from the same problem.

As the neural dimensionality (or speed of processing) in each measured signal increases, the temporal resolution decreases, the noise (or non-communication related signal) increases, the idea of extracting causality from observation becomes hopeless. Importantly, this is not a problem that can simply be solved by recording data from more subjects. This is a case where a problem can fundamentally not be solved.

For all methods that we have at our disposal to run functional connectivity analyses, the signals are few relative to the many unobserved variables. In these settings, interactions between measured signals should describe much less variance than the interactions between unobserved variables. Just like in Simpson's paradox, the interactions between such unobserved variables can arbitrarily affect the estimates of interactions between measured signals. In neuroscience we are essentially always in the massively confounded situation and then correlations as not informative about causality.

Structure without causality

The statistical fact that we should not expect algorithms to convert sparsely observed neural data to tell us about causality in the brain contrasts with the fact that FC findings often seem meaningful. We, for example, might often find correlations between early visual cortices to higher visual cortices (Buchel and Karl, 1997; Shmuel and Leopold, 2008) or between visual cortices and parietal involved in motor planning and execution (Miller and Vogt, 1984). So how can our functional connectivity results look so meaningful if they do not actually measure causal interactions? Timing is a confounder that readily produces structure and apparent causality. Following the organization of the visuomotor system, the primary visual cortex has the lowest latency, followed by higher level visual areas, premotor areas, and motor areas (Bullier, 2001).

Importantly, most algorithms that utilize coherence or delays will effectively convert the order of activation into apparent causality. However, such timing may be entirely unrelated to causal interactions. It is sufficient, for example, if central planning activates each area at the time where it is relevant. Latency is being converted into an illusion of actual causality. Spatial blurring is also undeniably a factor. Nearby regions will naturally share some variance, both by artifacts of

the measurement techniques. As such, it is not surprising that nearby areas will often have FC. If FC merely reflects distance, then it provides no causal insights. Interestingly, corrections of functional connectivity for distance are rare and have been simple (Hagmann et al., 2008; Honey et al., 2009, 2007; Litvak et al., 2011). In some cases, spatial distance thus may be converted into an illusion of causality.

Ultimately, there are any number of processes which affect correlations. Hence, any aspects of brain function may be reflected in the resulting correlations. Unfortunately, the inversion is impossible – we cannot meaningfully conclude about causal chains in a high-dimensional system by observing the correlations between a (relatively) small number of observed variables.

Testing causality

So far, we have reviewed the logical evidence why FC from a small number of channels should, due to massive confounding, not be a suitable tool to reveal causal structure of brains that contain billions of neurons. However, in principle there may be aspects of brain activity that could rescue the idea. For example, if the brain's activity was very low dimensional (Cunningham and Yu, 2014; Yu et al., 2009), then recording from a small number of voxels or neurons may be equivalent to recording all of them. But in this case, all the neurons that jointly define a dimension will confound the causal inference. Similarly, if we believe that the composition of the activity within a voxel does not matter but just the sum of the activities, and thus subscribe to a strong mean-field view (Cooper and Scofield, 1988; Gerstner et al., 2012), the statistical problems may be resolved. However, within each voxel, we find neurons of countless tuning properties (Hubel and Wiesel, 1962). Alternatively, there may be something informative in the structure of neuronal signals that somehow makes this analysis possible. We may hope to gain some insight into the truth by analyzing the algorithm performance in situations where we know what to expect.

If we should expect FC to meaningfully work then we should, above all, expect that it would work in simulated situations that lack many of the complexities characterizing the real world. And indeed, a recent study has systematically analyzed the quality of FC estimates based on the number of neurons that were not simulated. Removing as many as 20% of the neurons from the recordings let the reconstruction quality drop from 100% to 70% (with a chance level of 50%). Hence, FC measures can tolerate some proportion of missing neurons but only a small proportion (compare to nearly 100% for typical spike analysis; (Karbasi et al., 2018)). This simulation approach casts doubt on the usefulness of functional connectivity for estimating causality.

FC has been shown to be susceptible to the effect of temporal delays across modalities. Electrophysiological recordings show that connectivity estimates are highly susceptible to effects of temporal smoothing (Stevenson and Körding, 2010). fMRI simulations show that temporal down-sampling of neural activity alters connectivity estimates drastically (Barnett and Seth, 2017) such that estimates do not resemble reality when data is down-sampled to a rate typical for fMRI experiments (Handwerker et al., 2004). Moreover, temporal delays of the hemodynamic response function (HRF), which varies between brain locations and subjects, may further confound results (Rangaprakash et al., 2018; Seth et al., 2013). This work shows that combinations of temporal smoothing and delays can massively distort functional connectivity estimates, casting doubt on the use of such techniques for estimating causality (Smith et al., 2011).

A radical philosophical alternative

Parts of the FC community, those using the term *effective connectivity* (EC) introduced a rather interesting philosophical alternative. In this view EC produces, or *causes* observed FC in a mathematical sense (Friston, 2011). As such they are causal given unverified, and arguably unlikely, assumptions, and methodologists thus do not explicitly claim that they are studying real connectivity. They merely compare different models and make statements about the preferable model in a given model class. Interestingly, DCM does derive from ideas of perturbations, it conceptualizes stimuli as perturbations (Friston et al., 2003). However, this strategy equally can not guard against the potential effects of confounding. The statements afforded by this approach are thus not about (biological) causality in the brain but about preference for a model from that model class ("All models are wrong, but some are useful"). Yet, once Dynamic Causal Modeling (DCM) is applied to brain signals, it is claimed that these models "estimate neurobiologically interpretable quantities such as the effective strength of synaptic connections among neuronal populations and their context-dependent modulation" (Stephan et al., 2010). Thus, DCM applications that implicitly aim to unravel neural mechanisms misinterpret causality in the model as biological causality in the same way as other eC techniques. We further note that expanding the model class can arbitrarily change the resulting causal conclusions. In other words, this approach does not afford statements about causal interactions within the brain. Rather, it risks falling for internally consistent statements about network dynamics that may be causal in a mathematical, but not in a biological sense (Etkin, 2018).

The potential for correct "estimated connectivity" in the future

Algorithms claiming eC have been used exhaustively when analyzing brain data with the hope of getting at causality. Using a more accurate terminology could help in the interpretation and, clarify that our results remain foremost descriptive: *Interregional*, or *interneural signal correlations* captures what most techniques measure. After all, despite massive confounding, there seemed to be few alternatives. However, the situation is changing.

It is possible to perturb the brain in many ways and thus to go beyond purely correlational experiments. To confidently get at causality requires perturbations to evaluate how a given input to the system modulates activities. Invasive brain stimulation techniques (e.g. optogenetics, intracortical electrical stimulation, and deep brain stimulation) and non-invasive brain stimulation techniques (e.g. transcranial magnetic stimulation and transcranial electrical stimulation) allow brain perturbations. Combined with classification of inter-regional correlations these approaches may yield biomarkers for treatment response (Drysdale et al., 2016; but see also Dinga et al., 2018). Perturbation techniques are currently, despite their limitations (Bestmann et al., 2015; Häusser, 2014; Sack and Linden, 2003; Siebner et al., 2009), as close as we can get to causality in neuroscience (Chen and Rothwell, 2012; Muldoon et al., 2016). However, we also note that inference about the inverse is more complex: lack of behavioral response after perturbing a certain area does not a imply that it is not causally but may be merely due to compensatory recruitment (Krakauer et al., 2017; O'Shea et al., 2007; Sack and Linden, 2003).

Perturbation techniques such as bio-feedback training and brain computer interfaces (BCIs) may be seen as an approximation to direct (self-) control in humans (Arns et al., 2017; Chaudhary et al., 2016; Sitaram et al., 2017; Watanabe et al., 2017). These allow to entrain correlations between brain regions and test for desired behavioral changes (Ramot et al., 2017; Yamashita et al., 2017). Similarly, BCIs allow coupling neurons causally to outcomes and thus they may provide new tools for trying to understand causality within the brain (Golub et al., 2018; Sadtler et al., 2014). When working with randomized stimuli, BCI perturbation studies may allow to directly measure causal interactions between brain regions (Grosse-Wentrup et al., 2016). While these examples show that real-time entrainment of statistical dependencies in in neuroimaging shows promise, correction for confounds that are in principle correctable (e.g. spurious correlations due to head motion and physiological noise) still pose challenges (Maclaren et al., 2013; Misaki et al., 2015). Closed-loop optogenetic stimulation, where stimulation is triggered by learned neural activity patterns, has provided further means of studying causality invasively in animal models (Athalye et al., 2018). Optogenetic fMRI yields new possibilities to test for causal *effective connectivity* (Bernal-Casas et al., 2017), although temporal differences

for responses due to physiological and optogenetic stimulation (Albers et al., 2018) may impose limitations.

Simulated perturbation experiments allow us to check if our assumptions are correct and what happens if they aren't. Computational studies are an extension of thought experiments and have a long tradition in neuroscience (Gerstner et al., 2012; Lapicque, 1907). For example, we can evaluate the susceptibility to pitfalls of various common FC (Bastos and Schoffelen, 2016) and test how low signal-to-noise ratios affect false positive rates. Another pitfall of non-invasive electrophysiological studies (EEG/MEG) is volume conductance, which can lead to spurious FC estimates, and thus require sanity checks (Haufe et al., 2012). Simulation will help us to validate approaches, test the sensitivity of sanity checks and the effectiveness of potential remedies.

For the analysis of some spiking systems there may be ways of sidestepping the confounding problem. If we have fast recordings and sparse connectivity patterns so that there is no time for the signal to travel sufficiently quickly from one neuron to another through any path but the most direct one, then the system effectively gets to be unconfounded. While this is not the typical setting, there may thus be ways of sidestepping it for certain subproblems (Bartho, 2004; English et al., 2017; Swadlow and Giisev, 2001; Swadlow and Gusev, 2002; Usrey et al., 2000). The intuition here is that if we record sufficiently fast, we see the direct and immediate effect of one neuron on another such that we can effectively use the timescale to *de-confound* estimates.

It may also be possible that there is something about brain signals that, given the right analysis methods, makes it possible to approximately estimate causal interactions, say between brain regions. Maybe there are signals that are low dimensional and localized, maybe there are sparse localized events that result in signals that approximate external perturbations. Statistical techniques may be possible. For example, quasi-experimental techniques may allow meaningful causal estimates about brain connectivity (Lansdell and Kording, 2018; Lepperød et al., 2018). However, we cannot meaningfully use such techniques to learn about causality without having established that the strong extra assumptions we would need to make about the brain are justified and before having used simulations to check that under those assumptions the methods would actually work.

For the obvious confounders the field is starting to use simulations to test basic assumptions of their techniques (Bastos and Schoffelen, 2016; Bright and Murphy, 2015; Haufe et al., 2013; Ramsey et al., 2010; Rangaprakash et al., 2018; Seth et al., 2013; Stokes and Purdon, 2018; Thompson et al., 2018). We propose that *Interregional*, or *interneural signal correlations* should be routinely checked for obvious signs of the Simpson's paradox as has been suggested for other fields (Kievit et al., 2013; Rousselet and Pernet, 2012; Tu et al., 2008).

Because network analyses come with many degrees of freedom in data processing (Carp, 2012a, 2012b), researchers should publish robustness checks, for instance by showing that results within the possible analysis space largely converge (Eippert et al., 2017; Karbasi et al., 2018), and ideally pre-register these pipelines (Allen and Mehler, 2018). Further, estimates of functional network can be validated against known structural network architecture or simulations (Lennartz et al., 2018; Schiefer et al., 2018). Taken together, thorough simulation checks within confounds that are observable can already flag problems. They thus help us to deal with rather trivial confounding problems that occur in FC, but also other areas of (neuro) science.

The extensive confounding from only observing a small subset of neurons, however, seems hard to overcome without massive scale recording technologies or small brains. Many algorithms used for the inference of functional connectivity are quite meaningful when applied to systems that are exhaustively recorded. As such, it is an interesting question how well they work on small animals or engineered systems. For example, in the worm c. elegans, aplysia, or microprocessors (Jonas and Kording, 2017) recording all “neurons” at high temporal precision should be possible and in the larval zebrafish that may be possible soon (Ahrens et al., 2013). In systems where the complete circuitry is described, perturbation (e.g. pharmacological) studies can be used to test whether algorithms can reconstruct expected patterns (Gerhard et al., 2013). Such systems are much closer to the implied assumptions of the various causal inference techniques, mainly because they make the confounders observable. There would obviously still be confounding, e.g. from limited temporal resolution, making the causal inference problem quite hard. However, there are interesting approaches from econometrics, e.g. regression discontinuity designs and other pseudo experiments (Angrist and Pischke, 2008) that may be helpful (Marinescu et al., 2018).

We have argued here, that eC approaches cannot meaningfully get at causality due to massive confounding, but it is important to point out that other branches of neuroscience and biology more generally have the same problem (Jonas and Kording, 2017). For instance, tuning curves, which describe how neurons are affected by a stimulus dimension such as color, do not reveal how they are computed. Lesions usually induce compensation making it hard to interpret their causality. Pharmacological interventions or stimulation studies typically perturb many neurons making it hard to assign the undeniable causal link to any specific neuronal path. However, the logical problems in other areas of neuroscience do not render a lack of logical precision in the FC field (more) acceptable. It is time for the field to take the reality of causal inference seriously (Angrist and Pischke, 2008; Jonas Peters et al., 2017; Pearl, 2009a).

Understanding the joint statistics of neurons is still interesting

We have reviewed why FC approaches cannot meaningfully get at causality due to massive confounding from unobserved variables. However, this circumstance does not imply that trying to understand the joint dynamics of many neurons or brain areas is not interesting. For example, looking at the brain in lower dimensional projections may allow us to see its invariances (Bruno et al., 2017; Gallego et al., 2017; Gordon et al., 2017). *Interregional correlations* may have no causal meaning, but they may allow us to derive biomarkers (Drysdale et al., 2016; Faiman et al., 2018; Medaglia et al., 2017; Muldoon et al., 2018). Informative markers may even be derived from brain-body correlations (Rebollo et al., 2018; Valenza et al., 2016). In fact, there are many biologically meaningful questions (Gomez-Marin, 2017; Krakauer et al., 2017) – and insights about brains can come from answering any of them. It is just important to be clear about the statements permitted by any one approach – statements about joint statistics are not meaningful statements about causality.

Along with these considerations comes an important set of insights into the sociology of neuroscience. The current publication system almost forces authors to make causal statements using filler verbs (e.g. to drive, alter, promote) as a form of storytelling (Gomez-Marin, 2017); without such a statement they are often accused of just collecting meaningless facts. In the light of our discussion this is a major mistake, which incites the field to misrepresent its findings. Understanding the structure of brain data is interesting in its own right. Scientists who actually measure causality using carefully designed perturbations should be lauded for the hard work. At the same time, scientists who describe joint statistics should be rewarded for careful characterizations. We do learn about the brain by analyzing joint distributions. We simply should not claim causality.

References

Ahrens, M.B., Orger, M.B., Robson, D.N., Li, J.M., Keller, P.J., 2013. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nat. Methods 10, 413–420. https://doi.org/10.1038/nmeth.2434

Albers, F., Schmid, F., Wachsmuth, L., Faber, C., 2018. Line scanning fMRI reveals earlier onset of optogenetically evoked BOLD response in rat somatosensory cortex as compared to sensory stimulation. Neuroimage 164, 144–154. https://doi.org/10.1016/j.neuroimage.2016.12.059

Allen, C.P.G., Mehler, D.M.A., 2018. Open Science challenges, benefits and tips in early career and beyond. PsyArXiv Prepr. https://doi.org/10.31234/OSF.IO/3CZYT

Angrist, J.D., Pischke, J.-S., 2008. Mostly Harmless Econometrics: An Empiricist's Companion, 1st ed. Princeton University Press.

Arns, M., Batail, J.-M., Bioulac, S., Congedo, M., Daudet, C., Drapier, D., Fovet, T., Jardri, R., Le-Van-Quyen, M., Lotte, F., Mehler, D., Micoulaud-Franchi, J.-A., Purper-Ouakil, D., Vialatte, F., 2017. Neurofeedback: One of today's techniques in psychiatry? Encephale. 43, 135–145. https://doi.org/10.1016/j.encep.2016.11.003

Athalye, V.R., Santos, F.J., Carmena, J.M., Costa, R.M., 2018. Evidence for a Neural Law of Effect 1029, 1024–1029. https://doi.org/10.1126/science.aao6058

Babiloni, F., Cincotti, F., Babiloni, C., Carducci, F., Mattia, D., Astolfi, L., Basilisco, A., Rossini, P.M., Ding, L., Ni, Y., Cheng, J., Christine, K., Sweeney, J., He, B., 2005. Estimation of the cortical functional connectivity with the multimodal integration of high-resolution EEG and fMRI data by directed transfer function. Neuroimage 24, 118–131. https://doi.org/10.1016/j.neuroimage.2004.09.036

Bajaj, S., Butler, A.J., Drake, D., Dhamala, M., 2015. Brain effective connectivity during motor-imagery and execution following stroke and rehabilitation. NeuroImage Clin. 8, 572–582. https://doi.org/10.1016/j.nicl.2015.06.006

Barnett, L., Seth, A.K., 2017. Detectability of Granger causality for subsampled continuous-time neurophysiological processes. J. Neurosci. Methods 275, 93–121. https://doi.org/10.1016/j.jneumeth.2016.10.016

Bartho, P., 2004. Characterization of Neocortical Principal Cells and Interneurons by Network Interactions and Extracellular Features. J. Neurophysiol. 92, 600–608.

https://doi.org/10.1152/jn.01170.2003

Bassett, D.S., Meyer-Lindenberg, A., Achard, S., Duke, T., Bullmore, E., 2006. Adaptive reconfiguration of fractal small-world human brain functional networks. Proc. Natl. Acad. Sci. 103, 19518–19523. https://doi.org/10.1073/pnas.0606005103

Bassett, D.S., Yang, M., Wymbs, N.F., Grafton, S.T., 2015. Learning-induced autonomy of sensorimotor systems. Nat. Neurosci. 18, 744–751. https://doi.org/10.1038/nn.3993

Bastos, A.M., Schoffelen, J.-M., 2016. A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls. Front. Syst. Neurosci. 9, 1–23. https://doi.org/10.3389/fnsys.2015.00175

Bernal-Casas, D., Lee, H.J., Weitz, A.J., Lee, J.H., 2017. Studying Brain Circuit Function with Dynamic Causal Modeling for Optogenetic fMRI. Neuron 93, 522–532.e5. https://doi.org/10.1016/j.neuron.2016.12.035

Bestmann, S., de Berker, A.O., Bonaiuto, J., Berker, A.O. De, Bonaiuto, J., 2015. Understanding the behavioural consequences of noninvasive brain stimulation. Trends Cogn. Sci. 19, 13–20. https://doi.org/10.1016/j.tics.2014.10.003

Bijsterbosch, J.D., Woolrich, M.W., Glaser, M.F., Robinson, E.C., Beckmann, C.F., van Essen, D.C., Harrison, S.J., Smith, S.M., 2018. The relationship between spatial configuration and functional connectivity of brain regions in press. https://doi.org/10.7554/eLife.32992

Bressler, S.L., Seth, A.K., 2011. Wiener-Granger Causality: A well established methodology. Neuroimage 58, 323–329. https://doi.org/10.1016/j.neuroimage.2010.02.059

Bright, M.G., Murphy, K., 2015. Is fMRI “noise” really noise? Resting state nuisance regressors remove variance with network structure. Neuroimage 114, 158–169. https://doi.org/10.1016/j.neuroimage.2015.03.070

Bruno, A.M., Frost, W.N., Humphries, M.D., 2017. A spiral attractor network drives rhythmic locomotion. Elife 6, 1–24. https://doi.org/10.7554/eLife.27342

Buchel, C., Karl, F., 1997. Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. Cereb. Cortex 7, 768–778. https://doi.org/10.1093/cercor/7.8.768

Bullier, J., 2001. Integrated model of visual processing. Brain Res. 36, 96–107.

Bullmore, E.T., Bassett, D.S., 2011. Brain Graphs: Graphical Models of the Human Brain Connectome. Annu. Rev. Clin. Psychol. 7, 113–140. https://doi.org/10.1146/annurev-clinpsy-040510-143934

Bullmore, E.T., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Neurosci. 10, 186–198. https://doi.org/doi: 10.1038/nrn2575

Calabrese, A., Schumacher, J.W., Schneider, D.M., Paninski, L., Woolley, S.M.N., 2011. A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. PLoS One 6. https://doi.org/10.1371/journal.pone.0016104

Carp, J., 2012a. On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. Front. Neurosci. 6, 149. https://doi.org/10.3389/fnins.2012.00149

Carp, J., 2012b. The secret lives of experiments: Methods reporting in the fMRI literature. Neuroimage 63, 289–300. https://doi.org/10.1016/j.neuroimage.2012.07.004

Chaudhary, U., Birbaumer, N., Ramos-Murguialday, A., 2016. Brain-computer interfaces for communication and rehabilitation. Nat. Rev. Neurol. 12, 513–525. https://doi.org/10.1038/nrneurol.2016.113

Chen, R., Rothwell, J. (Eds.), 2012. Cortical Connectivity. Brain Stimulation for Assessing and Modulating Cortical Connectivity and Function, 1st ed. Springer Berlin Heidelberg.

Ciric, R., Wolf, D.H., Power, J.D., Roalf, D.R., Baum, G.L., Ruparel, K., Shinohara, R.T., Elliott, M.A., Eickhoff, S.B., Davatzikos, C., Gur, R.C., Gur, R.E., Bassett, D.S., Satterthwaite, T.D., 2017. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. Neuroimage 154, 174–187. https://doi.org/10.1016/j.neuroimage.2017.03.020

Cooper, L., Scofield, C., 1988. Mean-field theory of a neural network. Proc. Natl. … 85, 1973–1977. https://doi.org/10.1142/9789812795885_0007

Cunningham, J.P., Yu, B.M., 2014. Dimensionality reduction for large-scale neural

recordings. Nat. Neurosci. 17, 1500–1509. https://doi.org/10.1038/nn.3776

Daunizeau, J., David, O., Stephan, K.E., 2011. Dynamic causal modelling: A critical review of the biophysical and statistical foundations. Neuroimage 58, 312–322. https://doi.org/10.1016/j.neuroimage.2009.11.062

Dawid, A.P., 2008. Beware of the DAG! NIPS 2008Workshop Causality 59–86. https://doi.org/10.2459/JCM.0b013e32833dadd5

DiDomenico, R., Eaton, R., 1988. Seven Principles for Command and the Neural Causation of Behavior. Brain. Behav. Evol. 31, 125–140.

Dimyan, M.A., Cohen, L.G., 2011. Neuroplasticity in the context of motor rehabilitation after stroke. Nat. Rev. Neurol. 7, 76–85. https://doi.org/10.1038/nrneurol.2010.200

Dinga, R., Schmaal, L., Penninx, B., Tol, M.J. Van, Veltman, D.J., Velzen, L. Van, Wee, N. Van Der, Marquand, A., 2018. Evaluating the evidence for biotypes of depression: attempted replication of Drysdale et.al. 2017. BioRxiV.

Driver, I.D., Whittaker, J.R., Bright, M.G., Muthukumaraswamy, S.D., Murphy, K., 2016. Arterial CO2 Fluctuations Modulate Neuronal Rhythmicity: Implications for MEG and fMRI Studies of Resting-State Networks. J. Neurosci. 36, 8541–50. https://doi.org/10.1523/JNEUROSCI.4263-15.2016

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R.N., Zebley, B., Oathes, D.J., Etkin, A., Schatzberg, A.F., Sudheimer, K., Keller, J., Mayberg, H.S., Gunning, F.M., Alexopoulos, G.S., Fox, M.D., Pascual-Leone, A., Voss, H.U., Casey, B., Dubin, M.J., Liston, C., 2016. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat. Med. 23. https://doi.org/10.1038/nm.4246

Eippert, F., Kong, Y., Winkler, A.M., Andersson, J.L., Finsterbusch, J., Büchel, C., Brooks, J.C.W., Tracey, I., 2017. Investigating resting-state functional connectivity in the cervical spinal cord at 3 T. Neuroimage 147, 589–601. https://doi.org/10.1016/j.neuroimage.2016.12.072

English, D.F., McKenzie, S., Evans, T., Kim, K., Yoon, E., Buzsáki, G., 2017. Pyramidal Cell-Interneuron Circuit Architecture and Dynamics in Hippocampal Networks. Neuron 96, 505–520.e7. https://doi.org/10.1016/j.neuron.2017.09.033

Etkin, A., 2018. Addressing the causality gap in human psychiatric neuroscience. JAMA

Psychiatry 75, 3–4. https://doi.org/10.1001/jamapsychiatry.2017.3610

Faiman, I., Pizzamiglio, S., Turner, D.L., 2018. Resting-state functional connectivity predicts the ability to adapt arm reaching in a robot-mediated force field. Neuroimage 174, 494–503. https://doi.org/10.1016/j.neuroimage.2018.03.054

Friston, K.J., 2011. Functional and Effective Connectivity: A Review. Brain Connect. 1, 13–36. https://doi.org/10.1089/brain.2011.0008

Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J., 1997. Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6, 218–229. https://doi.org/10.1006/nimg.1997.0291

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. Neuroimage 19, 1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7

Gallego, J.A., Perich, M.G., Naufel, S.N., Ethier, C., Solla, S.A., Miller, L.E., 2017. Multiple tasks viewed from the neural manifold: Stable control of varied behavior. bioRxiv 1–27. https://doi.org/10.1101/176081

Gerhard, F., Kispersky, T., Gutierrez, G.J., Marder, E., Kramer, M., Eden, U., 2013. Successful Reconstruction of a Physiological Circuit with Known Connectivity from Spiking Activity Alone. PLoS Comput. Biol. 9, 32–34. https://doi.org/10.1371/journal.pcbi.1003138

Gerstner, W., Sprekeler, H., Deco, G., 2012. Theory and simulation in neuroscience. Science (80-.). 338, 60–65. https://doi.org/10.1126/science.1227356

Golub, M.D., Sadtler, P.T., Oby, E.R., Quick, K.M., Ryu, S.I., Tyler-Kabara, E.C., Batista, A.P., Chase, S.M., Yu, B.M., 2018. Learning by neural reassociation. Nat. Neurosci. 21, 607–616. https://doi.org/10.1038/s41593-018-0095-3

Gomez-Marin, A., 2017. Causal Circuit Explanations of Behavior: Are Necessity and Sufficiency Necessary and Sufficient? Decod. Neural Circuit Struct. Funct. 283–306. https://doi.org/10.1007/978-3-319-57363-2_11

Gordon, E.M., Laumann, T.O., Gilmore, A.W., Newbold, D.J., Greene, D.J., Berg, J.J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J.M., Coalson, R.S., Nguyen, A.L., McDermott, K.B., Shimony, J.S., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., Nelson, S.M., Dosenbach, N.U.F., 2017. Precision Functional Mapping of Individual Human Brains. Neuron 95, 791–807.e7.

https://doi.org/10.1016/j.neuron.2017.07.011

Grosse-Wentrup, M., Janzing, D., Siegel, M., Schölkopf, B., 2016. Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. Neuroimage 125, 825–833. https://doi.org/10.1016/j.neuroimage.2015.10.062

Hacker, C.D., Perlmutter, J.S., Criswell, S.R., Ances, B.M., Snyder, A.Z., 2012. Resting state functional connectivity of the striatum in Parkinson's disease. Brain 135, 3699–3711. https://doi.org/10.1093/brain/aws281

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C.J., Van Wedeen, J., Sporns, O., 2008. Mapping the structural core of human cerebral cortex. PLoS Biol. 6, 1479–1493. https://doi.org/10.1371/journal.pbio.0060159

Hallam, G.P., Thompson, H.E., Hymers, M., Millman, R.E., Rodd, J.M., Lambon Ralph, M.A., Smallwood, J., Jefferies, E., 2018. Task-based and resting-state fMRI reveal compensatory network changes following damage to left inferior frontal gyrus. Cortex 99, 150–165. https://doi.org/10.1016/j.cortex.2017.10.004

Handwerker, D.A., Ollinger, J.M., D'Esposito, M., 2004. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. Neuroimage 21, 1639–1651. https://doi.org/10.1016/j.neuroimage.2003.11.029

Haufe, S., Nikulin, V. V., Müller, K.R., Nolte, G., 2013. A critical assessment of connectivity measures for EEG data: A simulation study. Neuroimage 64, 120–133. https://doi.org/10.1016/j.neuroimage.2012.09.036

Haufe, S., Nikulin, V. V., Nolte, G., 2012. Alleviating the Influence of Weak Data Asymmetries on Granger-Causal Analyses. Springer, Berlin, Heidelberg, pp. 25–33. https://doi.org/10.1007/978-3-642-28551-6_4

Häusser, M., 2014. Optogenetics: The age of light. Nat. Methods 11, 1012–1014. https://doi.org/10.1038/nmeth.3111

Havlicek, M., Ivanov, D., Roebroeck, A., Uludağ, K., 2017. Determining excitatory and inhibitory neuronal activity from multimodal fMRI data using a generative hemodynamic model. Front. Neurosci. 11, 1–20. https://doi.org/10.3389/fnins.2017.00616

Holland, P.W., 2015. Causal Inference , Path Analysis , and Recursive Structural Equations Models 18, 449–484.

Honey, C.J., Honey, C.J., Sporns, O., Sporns, O., Cammoun, L., Cammoun, L., Gigandet, X., Gigandet, X., Thiran, J.P., Thiran, J.P., Meuli, R., Meuli, R., Hagmann, P., Hagmann, P., 2009. Predicting human resting-state functional connectivity from structural connectivity. Proc. Natl. Acad. Sci. U. S. A. 106, 2035–40. https://doi.org/10.1073/pnas.0811168106

Honey, C.J., Kötter, R., Breakspear, M., Sporns, O., 2007. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. Proc. Natl. Acad. Sci. 104, 10240–10245. https://doi.org/10.1073/pnas.0701519104

Hubel, D., Wiesel, T., 1962. RECEPTIVE FIELDS,BINOCULAR INTERACTION AND FUNCTIONAL ARCHITECTURE IN THE CAT ' S VISUAL CORTEX From the Neurophysiolojy Laboratory , Department of Pharmacology central nervous system is the great diversity of its cell types and inter- receptive fields of. J. Physiol. 160, 106–154. https://doi.org/10.1523/JNEUROSCI.1991-09.2009

Jin, S.H., Lin, P., Auh, S., Hallett, M., 2011. Abnormal functional connectivity in focal hand dystonia: Mutual information analysis in EEG. Mov. Disord. 26, 1274–1281. https://doi.org/10.1002/mds.23675

Jonas, E., Kording, K.P., 2017. Could a Neuroscientist Understand a Microprocessor? PLoS Comput. Biol. 13, 1–24. https://doi.org/10.1371/journal.pcbi.1005268

Jonas Peters, Dominik Janzing, Bernhard Scholkopf, 2017. Elements of causal inference: foundations and learning algorithms. MIT Press.

Karbasi, A., Salavati, A.H., Vetterli, M., 2018. Learning neural connectivity from firing activity: efficient algorithms with provable guarantees on topology. J. Comput. Neurosci. https://doi.org/10.1007/s10827-018-0678-8

Karlsgodt, K.H., Sun, D., Jimenez, A.M., Lutkenhoff, E.S., Willhite, R., Van erp, T.G.M., Cannon, T.D., 2008. Developmental disruptions in neural connectivity in the pathophysiology of schizophrenia. Dev. Psychopathol. 20, 1297–1327. https://doi.org/10.1017/S095457940800062X

Khambhati, A.N., Davis, K.A., Lucas, T.H., Litt, B., Bassett, D.S., 2016. Virtual Cortical Resection Reveals Push-Pull Network Control Preceding Seizure Evolution.

Neuron 91, 1170–1182. https://doi.org/10.1016/j.neuron.2016.07.039

Kievit, R.A., Frankenhuis, W.E., Waldorp, L.J., Borsboom, D., 2013. Simpson's paradox in psychological science: A practical guide. Front. Psychol. 4, 1–14. https://doi.org/10.3389/fpsyg.2013.00513

Krakauer, J.W., Ghazanfar, A.A., Gomez-Marin, A., MacIver, M.A., Poeppel, D., 2017. Neuroscience Needs Behavior: Correcting a Reductionist Bias. Neuron 93, 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Lansdell, B., Kording, K.P., 2018. Spiking allows neurons to estimate their causal effect. bioRxiv 253351. https://doi.org/10.1101/253351

Lapicque, L., 1907. Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarisation. J. Physiol. Pathol. Gen. 9, 620–635.

Lennartz, C., Schiefer, J., Rotter, S., Hennig, J., LeVan, P., 2018. Sparse estimation of resting-state effective connectivity from fMRI cross-spectra. Front. Neurosci. 12, 1–19. https://doi.org/10.3389/fnins.2018.00287

Lepperød, M.E., Tristan, S., Torkel, H., Marianne, F., Kording, K.P., 2018. Inferring causal connectivity from pairwise recordings enabled by optogenetics. BioRxiv. https://doi.org/10.1101/463760

Litvak, V., Mattout, J., Kiebel, S., Phillips, C., Henson, R., Kilner, J., Barnes, G., Oostenveld, R., Daunizeau, J., Flandin, G., Penny, W., Friston, K., 2011. EEG and MEG data analysis in SPM8. Comput. Intell. Neurosci. 2011. https://doi.org/10.1155/2011/852961

Lueken, U., Straube, B., Konrad, C., Wittchen, H.-U., Ströhle, A., Wittmann, A., Bettina Pfleiderer, D.-P., Uhlmann, C., Arolt, V., Jansen, A., Kircher, T., Pfleiderer, B., Uhlmann, C., Arolt, V., Jansen, A., Kircher, T., 2013. Neural Substrates of Treatment Response to Cognitive-Behavioral Therapy in Panic Disorder With Agoraphobia. Am. J. Psychiatry 170, 1345–55. https://doi.org/10.1176/appi.ajp.2013.12111484

Maclaren, J., Herbst, M., Speck, O., Zaitsev, M., 2013. Prospective motion correction in brain imaging: A review. Magn. Reson. Med. 69, 621–636. https://doi.org/10.1002/mrm.24314

Marinescu, I.E., Lawlor, P.N., Kording, K.P., 2018. Towards causality in neuroscience

and behavioral research. Nat. Hum. Behav. https://doi.org/10.1038/s41562-018-0466-5

Medaglia, J.D., Huang, W., Segarra, S., Olm, C., Gee, J., Grossman, M., Ribeiro, A., McMillan, C.T., Bassett, D.S., 2017. Brain network efficiency is influenced by the pathologic source of corticobasal syndrome. Neurology 89, 1373–1381. https://doi.org/10.1212/WNL.0000000000004324

Messaritaki, E., Koelewijn, L., Dima, D.C., Williams, G.M., Perry, G., Singh, K.D., 2017. Assessment and elimination of the effects of head movement on MEG resting-state measures of oscillatory brain activity. Neuroimage 159, 302–324. https://doi.org/10.1016/j.neuroimage.2017.07.038

Miller, M.W., Vogt, B.A., 1984. Direct connections of rat visual cortex with sensory, motor, and association cortices. J. Comp. Neurol. 226, 184–202. https://doi.org/10.1002/cne.902260204

Misaki, M., Barzigar, N., Zotev, V., Phillips, R., Cheng, S., Bodurka, J., 2015. Real-time fMRI processing with physiological noise correction – Comparison with off-line analysis. J. Neurosci. Methods 256, 117–121. https://doi.org/10.1016/j.jneumeth.2015.08.033

Muldoon, S.F., Costantini, J., Webber, W.R.S., Lesser, R., Bassett, D.S., 2018. Locally stable brain states predict suppression of epileptic activity by enhanced cognitive effort. NeuroImage Clin. #pagerange#. https://doi.org/10.1016/j.nicl.2018.02.027

Muldoon, S.F., Pasqualetti, F., Gu, S., Cieslak, M., Grafton, S.T., Vettel, J.M., Bassett, D.S., 2016. Stimulation-Based Control of Dynamic Brain Networks. PLoS Comput. Biol. 12. https://doi.org/10.1371/journal.pcbi.1005076

Murphy, K., Birn, R.M., Bandettini, P.A., 2013. Resting-state fMRI confounds and cleanup. Neuroimage 80, 349–59. https://doi.org/10.1016/j.neuroimage.2013.04.001

Musall, S., Kaufman, M.T., Gluf, S., Churchland, A., 2018. Movement-related activity dominates cortex during sensory-guided decision making. bioRxiv. https://doi.org/10.1101/308288

Nauhaus, I., Busse, L., Carandini, M., Ringach, D.L., 2009. Stimulus contrast modulates functional connectivity in visual cortex. Nat. Neurosci. 12, 70–76.

https://doi.org/10.1038/nn.2232

O'Shea, J., Johansen-Berg, H., Trief, D., Göbel, S., Rushworth, M.F.S., 2007. Functionally Specific Reorganization in Human Premotor Cortex. Neuron 54, 479–490. https://doi.org/10.1016/j.neuron.2007.04.021

Paninski, L., Brown, E., Iyengar, S., Kass, R., 2008. Statistical models of spike trains. Stoch. Methods Neurosci. 278–303.

Parkes, L., Fulcher, B., Yu¨cel, M., Fornitod, A., 2017. An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. Neuroimage. https://doi.org/10.1016/j.neuroimage.2017.12.073

Pawela, C.P., Biswal, B.B., Hudetz, A.G., Li, R., Jones, S.R., Cho, Y.R., Matloub, H.S., Hyde, J.S., 2010. Interhemispheric neuroplasticity following limb deafferentation detected by resting-state functional connectivity magnetic resonance imaging (fcMRI) and functional magnetic resonance imaging (fMRI). Neuroimage 49, 2467–2478. https://doi.org/10.1016/j.neuroimage.2009.09.054

Pearl, J., 2009a. Causality: Models, Reasoning and Inference, 2nd ed. Cambridge University Press.

Pearl, J., 2009b. Remarks on the method of stationary phases. Stat. Med. 127, 1415–1416.

Pearl, J., Mackenzie, D., 2018. The book of why. The new science of cause and effect. Basic Books.

Pillow, J.W., Shlens, J., Paninski, L., Sher, A., Litke, A.M., Chichilnisky, E.J., Simoncelli, E.P., 2008. Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature 454, 995–999. https://doi.org/10.1038/nature07140

Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage 84, 320–41. https://doi.org/10.1016/j.neuroimage.2013.08.048

R, C., AFG, R., G, E., M, C., A, A., PA, C., DS, B., C, D., DH, W., TD, S., 2018. Mitigating head motion artefact in functional connectivity MRI. Nat. Protoc. Press 13, 2801–2826. https://doi.org/10.1038/s41596-018-0065-y

Raichle, M.E., Macleod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L., 2001. A default mode of brain function. Proc. Natl. Acad. Sci. 98, 676–682.

https://doi.org/10.1073/pnas.98.2.676

Ramot, M., Kimmich, S., Gonzalez-Castillo, J., Roopchansingh, V., Popal, H., White, E., Gotts, S.J., Martin, A., 2017. Direct modulation of aberrant brain network connectivity through real - time NeuroFeedback. Elife 6:e28974, 1–23. https://doi.org/10.1101/139824

Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., Glymour, C., 2010. Six problems for causal inference from fMRI. Neuroimage 49, 1545–1558. https://doi.org/10.1016/j.neuroimage.2009.08.065

Ranganath, R., Perotte, A., 2018. Multiple Causal Inference with Latent Confounding. arXiv.

Rangaprakash, D., Wu, G.-R., Marinazzo, D., Hu, X., Deshpande, G., 2018. Hemodynamic response function (HRF) variability confounds resting-state fMRI functional connectivity. Magn. Reson. Med. 1–17. https://doi.org/10.1002/mrm.27146

Rebollo, I., Devauchelle, A.-D., Béranger, B., Tallon-Baudry, C., 2018. Stomach-brain synchrony reveals a novel, delayed-connectivity resting-state network in humans. Elife 7, e33321. https://doi.org/10.7554/eLife.33321

Roberts, R.P., Hach, S., Tippett, L.J., Addis, D.R., 2016. The Simpson's paradox and fMRI: Similarities and differences between functional connectivity measures derived from within-subject and across-subject correlations. Neuroimage 135, 1–15. https://doi.org/10.1016/j.neuroimage.2016.04.028

Rousselet, G.A., Pernet, C.R., 2012. Improving standards in brain-behavior correlation analyses. Front. Hum. Neurosci. 6. https://doi.org/10.3389/fnhum.2012.00119

Rubin, D.B., 2010. Matched Sampling for Causal Effects, 1st ed. Cambridge University Press.

Sack, A.T., Linden, D.E.J., 2003. Combining transcranial magnetic stimulation and functional imaging in cognitive brain research: possibilities and limitations. Brain Res. Brain Res. Rev. 43, 41–56. https://doi.org/S0165017303001917 [pii]

Sadtler, P.T., Quick, K.M., Golub, M.D., Chase, S.M., Ryu, S.I., Tyler-Kabara, E.C., Yu, B.M., Batista, A.P., 2014. Neural constraints on learning. Nature 512, 423–426. https://doi.org/10.1038/nature13665

Schiefer, J., Niederbühl, A., Pernice, V., Lennartz, C., Hennig, J., LeVan, P., Rotter, S., 2018. From correlation to causation: Estimating effective connectivity from zero-lag covariances of brain signals. PLoS Comput. Biol. 14, 1–18. https://doi.org/10.1371/journal.pcbi.1006056

Seth, A.K., 2018. https://twitter.com/anilkseth/status/1026958179425570817. [WWW Document].

Seth, A.K., Chorley, P., Barnett, L.C., 2013. Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling. Neuroimage 65, 540–555. https://doi.org/10.1016/j.neuroimage.2012.09.049

Shmuel, A., Leopold, D.A., 2008. Neuronal correlates of spontaneous fluctuations in fMRI signals in monkey visual cortex: Implications for functional connectivity at rest. Hum. Brain Mapp. 29, 751–761. https://doi.org/10.1002/hbm.20580

Siebner, H.R., Hartwigsen, G., Kassuba, T., Rothwell, J.C., 2009. How does transcranial magnetic stimulation modify neuronal activity in the brain? Implications for studies of cognition. Cortex 45, 1035–1042. https://doi.org/10.1016/j.cortex.2009.02.007

Siegel, M., Donner, T.H., Engel, A.K., 2012. Spectral fingerprints of large-scale neuronal interactions. Nat. Rev. Neurosci. 13, 121–134. https://doi.org/10.1038/nrn3137

Simpson, E.H., 1951. The interpretation of interaction in contigency tables. J. R. Stat. Soc. Ser. B 13, 238–241.

Sitaram, R., Ros, T., Stoeckel, L., Haller, S., Scharnowski, F., Lewis-Peacock, J., Weiskopf, N., Blefari, M.L., Rana, M., Oblak, E., Birbaumer, N., Sulzer, J., 2017. Closed-loop brain training: The science of neurofeedback. Nat. Rev. Neurosci. 18, 86–100. https://doi.org/10.1038/nrn.2016.164

Smith, S.M., Fox, P.T., Miller, K.L., Glahn, D.C., Fox, P.M., Mackaya, C.E., Filippinia, N., Watkinsa, K.E., Torod, R., Laird, A.R., Beckmann, C.F., 2009. Correspondence of the brain's functional architecture during activation and rest. Proc. Natl. Acad. Sci. 106, 13040–13045. https://doi.org/10.1051/0004-6361:20052893

Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., Ramsey, J.D., Woolrich, M.W., 2011. Network modelling methods for FMRI. Neuroimage 54, 875–891. https://doi.org/10.1016/j.neuroimage.2010.08.063

Spirtes, P., Glymour, C., Scheines, R., 2001. Causation, Prediction and Search (Adaptive Computation and Machine Learning Series), 2nd ed. MIT University Press.

Spirtes, P., Zhang, K., 2016. Causal discovery and inference: concepts and recent methodological advances. Appl. Informatics 3, 3. https://doi.org/10.1186/s40535-016-0018-x

Stephan, K.E., Penny, W.D., Moran, R.J., den Ouden, H.E.M., Daunizeau, J., Friston, K.J., 2010. Ten simple rules for dynamic causal modeling. Neuroimage 49, 3099–3109. https://doi.org/10.1016/j.neuroimage.2009.11.015

Stevenson, I.H., Kording, K.P., 2011. How advances in neural recording affect data analysis. Nat. Neurosci. 14, 139–142. https://doi.org/10.1038/nn.2731

Stevenson, I.H., Körding, K.P., 2010. On the similarity of functional connectivity between neurons estimated across timescales. PLoS One 5. https://doi.org/10.1371/journal.pone.0009206

Stevenson, I.H., Rebesco, J.M., Hatsopoulos, N.G., Haga, Z., Member, L.E.M., Konrad, P.K., 2009. Bayesian inference of functional connectivity and\nnetwork structure from spikes. IEEE Trans. Neural Syst. Rehab. 17, 1–10.

Stevenson, I.H., Rebesco, J.M., Miller, L.E., Körding, K.P., 2008. Inferring functional connections between neurons. Curr. Opin. Neurobiol. 18, 582–588. https://doi.org/10.1016/j.conb.2008.11.005

Stokes, P.A., Purdon, P.L., 2018. A study of problems encountered in Granger causality analysis from a neuroscience perspective. Proc. Natl. Acad. Sci. 115, E7063–E7072. https://doi.org/10.1073/pnas.1809324115

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C., Carandini, M., Harris, K.D., 2018. Spontaneous behaviors drive multidimensional , brain-wide neural activity. bioRxiv 1–26. https://doi.org/10.1101/306019

Sun, F.T., Miller, L.M., D'Esposito, M., 2004. Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data. Neuroimage 21, 647–658. https://doi.org/10.1016/j.neuroimage.2003.09.056

Swadlow, H.A., Giisev, A.G., 2001. The impact of'bursting'thalamic impulses at a neocortical synapse. Nat. Neurosci. 4, 402–408.

Swadlow, H.A., Gusev, A.G., 2002. Receptive-field construction in cortical inhibitory interneurons. Nat. Neurosci. 5, 403–404. https://doi.org/10.1038/nn847

Taylor, P.N., Thomas, J., Sinha, N., Dauwels, J., Kaiser, M., Thesen, T., Ruths, J., 2015. Optimal control based seizure abatement using patient derived connectivity. Front. Neurosci. 9, 1–10. https://doi.org/10.3389/fnins.2015.00202

Thompson, W.H., Richter, C.G., Plavén-Sigray, P., Fransson, P., 2018. Simulations to benchmark time-varying connectivity methods for fMRI. PLOS Comput. Biol. 14, e1006196. https://doi.org/10.1371/journal.pcbi.1006196

Tu, Y.-K., Gunnell, D., Gilthorpe, M.S., 2008. Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon – the reversal paradox. Neuphilol. Mitt. 5. https://doi.org/10.1186/1742-7622-5-2

Usrey, W.M., Alonso, J.M., Reid, R.C., 2000. Synaptic interactions between thalamic inputs to simple cells in cat visual cortex. J. Neurosci. 20, 5461–5467. https://doi.org/20/14/5461 [pii]

Valenza, G., Toschi, N., Barbieri, R., 2016. Uncovering brain-heart information through advanced signal and image processing. Philos. Trans. A. Math. Phys. Eng. Sci. 374, 387–390. https://doi.org/10.1098/rsta.2016.0020

Wang, Y., Blei, D.M., 2018. The Blessings of Multiple Causes. arXiv 1–59. https://doi.org/10.1245/s10434-012-2513-1

Watanabe, T., Sasaki, Y., Shibata, K., Kawato, M., 2017. Advances in fMRI Real-Time Neurofeedback. Trends Cogn. Sci. 21, 997–1010. https://doi.org/10.1016/J.TICS.2017.09.010

Wittgenstein, L., 1922. Tractatus logico-philosophicus, 2nd ed. Kegan Paul, Trench, Truber & Co., Ltd., London.

Wu, T., Wang, L., Chen, Y., Zhao, C., Li, K., Chan, P., 2009. Changes of functional connectivity of the motor network in the resting state in Parkinson's disease. Neurosci. Lett. 460, 6–10. https://doi.org/10.1016/j.neulet.2009.05.046

Yamashita, A., Hayasaka, S., Kawato, M., Imamizu, H., 2017. Connectivity Neurofeedback Training Can Differentially Change Functional Connectivity and Cognitive Performance. Cereb. Cortex 27, 4960–4970. https://doi.org/10.1093/cercor/bhx177

Yeo, B.T.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R.L., 2011. The organization of the human cerebral cortex estimates by intrinsic functional connectivity. J. Neurophysiol. 1125–1165. https://doi.org/10.1152/jn.00338.2011.

Yu, B.M., Cunningham, J.P., Santhanam, G., Ryu, S.I., Shenoy, K. V, Sahani, M., 2009. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. J. Neurophysiol. 102, 614–35. https://doi.org/10.1152/jn.90941.2008

Acknowledgements

We would like to thank our colleagues for insightful comments and discussions that have substantially contributed to the content of this manuscript: Alex Gomez Martin, Anil Seth, Brian Collopy, Daniele Marinazzo, David Linden, Dimitris Pinotsis, Ian Stevenson, Jean Daunizeau, Johannes Algermissen, Jonathan Pillow, Juan Alvaro Galego, Karl Friston, Lionel Barnett, Manjari Narayan, Michael Haufe, Robert Kass, Roscoe Brady, Russ Poldrack and Stefan Rotter (in alphabetical order). Through this process, we believe it has gotten considerably deeper, and, hopefully, less wrong.

Appendix

We provide some simple code here that may be the basis of exploring the various effects in which large partially observable systems are hard to reverse engineer. We explicitly do not use any specific estimated connectivity technique as any such choice would seem arbitrary. All code and data used in Figure 1 also available on: https://osf.io/9cs8p/

```
%% Define a brain
nNeurons=100;
pConnect=.1;
A=rand(nNeurons)<pConnect;
A=double(A);
%A=A+diag(ones(1,nNeurons));
%consider neurons with self history.
%A=A+100*diag(ones(nNeurons,1));
A=sparse(A);
s=svds(A,2);
A=A/s(1)/1.01; % set the longest time constant to be 100.
imagesc(A)

%% Simulate it
nTimes=2000;
noiseLevel=1;
x=zeros(nNeurons,nTimes);
x(:,1)=noiseLevel*randn(nNeurons,1);
for i=2:nTimes
    x(:,i)=A*x(:,i-1)+noiseLevel*randn(nNeurons,1);
end
imagesc(x);

%% relate correlations to generator
subplot(2,1,1)
R=corrcoef([x(:,1:end-1);x(:,2:end)]');
Rdelayed=R(1:nNeurons,nNeurons+1:2*nNeurons);
imagesc(Rdelayed)
subplot(2,1,2)
Adisp=A';
imagesc(Adisp)
j=corrcoef(Rdelayed(:),Adisp(:))

%% relate correlations between areas to generator
nRegions=10; %must evenly divide nNeurons, e.g. 10 of 300
B=zeros(nNeurons,nRegions);
for i=1:nRegions
    B((1:nNeurons/nRegions)+(i-1)*nNeurons/nRegions,i)=1;
end

% Now get the activity
y=B'*x;
```

```
subplot(2,1,1)
Ry=corrcoef([y(:,1:end-1);y(:,2:end)]');
RyDelayed=Ry(1:nRegions,nRegions+1:2*nRegions);
imagesc(RyDelayed)
subplot(2,1,2)
Adisp=(B'*A*B)';
imagesc(Adisp)
j=corrcoef(RyDelayed(:),Adisp(:))
```