Inferring causal connectivity from pairwise recordings and optogenetics

Mikkel Elle Lepperød[1,6,*], Tristan Stöber[4,5], Torkel Hafting[1], Marianne Fyhn[2], and Konrad Paul Kording[3]

[*]Corresponding author
[1]Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway
[2]Department of Bioscience, University of Oslo, Oslo, Norway
[3]Department of Neuroscience, University of Pennsylvania, Pennsylvania, USA
[4]Department of Informatics, University of Oslo, Oslo, Norway
[5]Department of Computational Physiology, Simula Research Laboratory, Oslo, Norway
[6]Department of Numerical Analysis and Scientific Computing, Simula Research Laboratory, Oslo, Norway

March 4, 2022

Abstract

To study how the brain works mechanistically, neuroscientists want to quantify causal interactions between neurons, typically requiring perturbations. When using optogenetic interventions, multiple neurons are usually perturbed which produces a confound – any of the stimulated neurons can have affected the postsynaptic neuron making it challenging to discern which of the neurons produced the causal effect. Here we show how such confounds produce large biases, and we explain how they can be reduced by combining instrumental variable (IV) and difference in differences (DiD) techniques from econometrics. The interaction between stimulation and the absolute refractory period of the neuron produces a weak, approximately random signal which can be exploited to estimate causal transmission probability. On simulated neural networks, we find that estimates using ideas from IV and DiD outperform naïve techniques suggesting that methods from causal inference can be useful when studying neural interactions.

1 Introduction

A central goal of neuroscience is, arguably, to understand the mechanisms or causal chains that give rise to brain activtiy leading to perception, cognition, and action. The human brain contains about 86 billion neurons[5] with countless connections. A complex system of this scale is incredibly hard to understand because of the numerous ways the contributing elements may interact internally[23]. While observing correlations within such systems are relatively easy, transitioning from observed correlations to a causal or mechanistic understanding is hard. After all, there can be many ways that the same activities emerge from distinct causal chains[12,34].

In today's typical studies in systems neuroscience, we only record from a small subset of all the neurons. The data obtained from such recordings, e.g. from electrophysiology or in vivo imaging, is observational, as opposed to interventional; typical approaches do not involve perturbations. When dealing with observational data, we generally cannot know how much of the observed activity was caused by observed activity, or by unobserved confounding activity. Relationships that are estimated from observational data in the presence of confounding may contain large errors and lead to incorrect conclusions[3]. Unobserved neural activity confounds estimates of causal interactions and makes it difficult to reveal underlying mechanisms.

Confounding is the big threat to causal validity[33] irrespective of the use of simple regressions or advanced functional connectivity techniques[43,21,2,35]. Inference techniques, such as maximum entropy-based inverse Ising inference[40], and generalized linear models (GLMs)[36] may give solutions to confounding from observed variables as they support "explaining away" background activity[43]. However, this is only a meaningful strategy if most neurons are included in the recordings. Even a fully observed system is only identifiable under certain assumptions about nonlinearity or noise[9,41]. Fully observable scenarios are rare in experimental settings, especially in the mammalian brain. Even in a fully observed simple simulated neural network there can be systematic errors[10] due to confounding emergent network states. Brain data will therefore almost never satisfy the criteria (back-door or front-door criterion) needed for causal identifiability[33]. Naïve regressions in partially observed systems will generally not reveal causality. Therefore, we aim to develop methods that can identify causal transmission probability between arbitrary neuron pairs.

We want to make clear which causal effects we want to measure. In this paper we focus on causal transmission probability. This is not just the direct influence of one neuron on a downstream neuron. Instead, it is the average change in activity induced by an extra spike of a neuron. This will have direct effects on all neurons that are immediately affected but will also include bi-synaptic and further effects coming from the network. This measure is, during perturbations, directly measurable. The direct effect will generally require some level of modeling. Estimating the direct causal effect from the causal transmission probability is an interesting and, to our knowledge, so far unsolved problem in computational neuroscience. However, causal transmission probability is what matters when we talk about neuronal activity. After all, a neuron will exert its effect through a multitude of paths. In this paper we focus on causal transmission probability only and this is what we call causal influence.

To estimate causal influence by one neuron on another, stimulating the presynaptic neuron is the gold standard. In fact, a common definition of causality is in terms of the effect of changing one variable in the system, independently of all the other variables – an intervention formalized by the so-called do() operator[33]. If we can experimentally stimulate single neurons, the ability to estimate causal relationships by regression is readily doable - the causal influence is the difference between average firing rates with and without stimulation. However, this approach is experimentally challenging (it is challenging to stimulate just one neuron) and yields low cell counts because it requires intracellular, juxtacellular, or two-photon stimulation[37,27,29,13]. Because gold-standard perturbations are challenging, it is necessary and highly desirable if causality could be obtained from optogenetic[6,49] stimulation in combination with neural recordings of large neuron populations.

Interpreting results from optogenetic stimulation in terms of causal interactions is often difficult. In most experimental settings, optogenetic stimulation will affect multiple neurons simultaneously. Hence, the stimulus will produce a distributed pattern of activity which then percolates through the network of neurons. Postsynaptic activity induced by stimulation could thus in principle come from any of the stimulated neurons, introducing problematic confounding.

For insights into how we may resolve the confounding problem induced by optogenetic stim-

ulation, we may look to other fields that have addressed the problem of confounding or more precisely of endogenous regressors. The inference of causality from observational data has been addressed in the fields of statistics[33], machine learning[34] and econometrics[3]. These fields have extensively worked on methods to estimate causality in the face of unobserved confounding and may provide clues on how to resolve these challenges in neuroscience.

A commonly used approach towards causal inference in economics are instrumental variables (IV), which have a truly long intellectual history[48]. The instrumental variable can be thought of as emulating the treatment assignment in a randomized controlled trial. It is a measurable variable that affects one of the variables of interest directly and the rest only indirectly. For example, imagine each neuron would have a molecular construct that would produce a flash of light and extra current at random times. We could then estimate how that current produces extra spikes and by correlating the flashes of light with both the upstream neuron and downstream neurons we could estimate causal effects.

We want to go through the definition of instrumental variables following an example from econometrics. Say we want to estimate the return β (how much extra money when studying) from education X (did we study) to yearly wages Y with the linear model $Y = \beta X + U$. Here U are the factors other than education that contribute to yearly wages (the error term). One of the factors in U is a person's cognitive ability which may also affect education. The regressor X is thus correlated with the error term U ($p(X) = f(U)$). This implies that the regression estimate of β will not estimate the magnitude of causation from education on wages, but rather, its association. Cognitive ability acts as a confounder. In this case one may use the proximity to a college or university (with the idea that cognitive ability does not affect location) as an instrumental variable (IV)[8] usually denoted by the variable Z (distance to college). This is following the idea that proximity to a college does not affect your cognitive ability but may affect your proclivity to attend college. We expect living in proximity of colleges and universities to give higher probability to attend without affecting other contributing factors to wages such as cognitive ability. Then, in order to attribute the causal average treatment effect of education on wages one may calculate the ratio of covariances $\beta = \mathrm{cov}\,(\text{proximity}, \text{wages})\,/\mathrm{cov}\,(\text{proximity}, \text{education})$. This ratio corrects for the confounding factor. In other words, IVs are a trick to use upstream randomness to identify a causal system.

For an instrument to be good, it needs to be unaffected by other variables (and it may be argued if the distance to the nearest college is a proper IV). In the brain, almost everything is affected by the network state. However, certain variables can be more or less affected. For example, the overall activity of the network is due to slow and strongly non-random dynamics. In contrast, the temporal pattern of when a neuron is in a refractory state may be in good approximation random. If neurons are spiking according to conditional Poisson distributions, their exact timing conditioned on the network state, will be random. While refractoriness may not be perfectly random, the exact times of spiking are notoriously difficult to predict[43] suggesting that refractoriness is quite random. It thus may serve as an instrumental variable.

When dealing with temporal data, we can use an additional method commonly used for causal inference, namely Difference in Differences (DiD). The effect of treatment in a longitudinal study with two time points Y_0, Y_1 can be estimated by taking the differences $(\mathrm{E}\,[Y_1|T=1] - \mathrm{E}\,[Y_0|T=1]) - (\mathrm{E}\,[Y_1|T=0] - \mathrm{E}\,[Y_0|T=0])$ where the two differences in parenthesis denote a treatment group and a control group respectively. The idea here is that we can compare pre and post spikes in otherwise matched brain states producing better estimates of causality.

Here we show that combining the IV and DiD techniques allows estimating the causal transmission probability between neuron pairs. We first show how confounding factors are introduced by conventional optogenetic stimulations of neurons. This confounding effect is then simulated in a simple network of three neurons using a Binomial GLM approach. With this simple model

we show that by using the refractory period as an IV, we are able to distinguish between connected and unconnected neuron pairs. By combining the IV estimate with DiD we correct for biased neuronal activity due to state differences before and after stimuli. We compare these estimates with a naïve, although widely used, cross-correlation histogram (CCH) method that fails to distinguish respective pairs. Furthermore, transmission probability is inferred in a simulated network of randomly connected excitatory and inhibitory neurons with distributed synaptic weights. With this data at hand, we compute errors of the IV and DiD methods and show that they are robust to different simulated network states. Furthermore, we analyze false positive and false negative errors and goodness of fit on synaptic weights with pairwise assessments using ordinary least squares (OLS) and DiD.

The observed differences between the IV with DiD and the CCH estimate, underline the importance of considering potential confounding, when estimating connections based on neural activity measurements.

2 Results

2.1 Confounding

When we simultaneously stimulate many neurons and observe the activity of a post-synaptic neuron after stimulation, it is hard to know which subset of the stimulated neurons caused the activity in the putative post-synaptic neuron. To illustrate such confounding effects we simulated a network comprised of three neurons (A,B, and C) Fig. 1(a). The neurons have a baseline activity and are additionally driven by Poisson distributed excitatory and inhibitory inputs of a timescale of 10ms. Together, this simulates slow network effects included in the background noise termed U. The interactions between neurons create a simple system to explore causality: both A and B receive stimulation, but only B drives C. We can then ask how to statistically resolve that only B drives C, a task called causal discovery, and how strongly B drives C, a task called causal inference[3,34].

In order to estimate the actual causal influence of a stimulated neuron on postsynaptic neurons, we need to distinguish the influence of one stimulated neuron from the influence of other stimulated neurons. This is hard, since the two neurons receive exactly the same inputs. Using unsuccessful stimulations, i.e. instances where one neuron is not responding due to membrane potential fluctuations, may be a first suggestion which is fine if neurons are independent outside stimulation periods. However, membrane potentials can be correlated on long timescales such as with brain oscillations[7] or up- down states[47]. During these long timescale states, the effect of stimulations are confounded by the network state. And indeed, a correlation-based measure fails to discover the correct network structure. The CCH and OLS estimates described below show the effect on the stimulated neurons (Fig. 1(c)). Since the stimulation affects both A and B simultaneously, it induces a strong correlation between A and B. This further confounds the system by rendering the correlation between B and C nonzero. When comparing the peak cross correlations with baseline as described in Section 4.4, specifically Eq. (11), $\hat{\beta}_{CCH}$ estimates the connectivity of A and C to be nonzero (Fig. 1(d)). A naïve reading of this result may suggest causal influences of both A and B on C[28]. This raises the question if there is any way of avoiding this problem.

2.2 Refractoriness as an instrumental variable

One way to avoid this kind of confounding, may be a timescale separation - something that is random on a short timescale. Arguably, a neuron's refractoriness is such a variable. If a neuron

is in its refractory period (typically about 1-4ms,[25]), it will be less likely to fire, and during the absolute refractory period, no amount of stimulation could make it spike. This gives us an interesting way of inferring causality, by comparing when the stimulation hits between a time when a neuron is able to spike and when the neuron is muted due to refractoriness. We would effectively not be using the spikes but the missing spikes for causal inference.

Every time we stimulate some neurons, in first approximation a random set, will be refractory. Let $X, Y, Z \in [0,1]^{N_S}$ denote random variables that indicate whether a neuron spiked relative to the stimulus during trials $i = 0, 1, 2, ..., N_S$. As depicted in Fig. 1(b), $Z_0 = 0$ indicates that the putative upstream neuron (here A or B) did not fire within a set time window before the first optogenetic stimulus, indicated by subscript 0. Furthermore, $X_0 = 1$ and $Y_0 = 1$ indicates that the upstream and downstream (here C) neurons respectively did fire in this first trial. We seek to estimate whether a response $Y_0 = 1$ was caused by $X_0 = 1$. The putative upstream neuron is in a refractory state if $Z = 1$, in our simulations this means that the neuron fired 1-2ms before stimulus onset. Now we have the relevant variables defined.

We can be statistically precise about the causal reasoning. Consider the directed acyclic graph (DAG)[33] representing variables and their relations Fig. 1(e). Here, X represents a putative upstream neuron e.g. A or B and Y represent the putative downstream neuron from e.g. C. The set U contains unobserved network activity and the stimulus.

During instrumental variable estimation we use Z (refractoriness of X) which is (at least in first order approximation) assumed to be random and influences spiking in X to infer the causal interaction (β) on spiking of X onto Y. It is known that this approach satisfies Pearl's criteria for causal inference[32] and instrumental variables can reveal real causality.

Let us start with the naïve estimate of β: the OLS estimate. For binary variables it can be written as the difference of conditional expectations

$$\hat{\beta}_{OLS} = \mathrm{E}\left[Y|X=1\right] - \mathrm{E}\left[Y|X=0\right] \tag{1}$$

when X is endogenous, i.e. $\mathrm{E}\left[X|U\right] \neq 0$, the estimate $\hat{\beta}_{OLS}$ will be biased. Importantly, we fail to estimate the background $E[Y|X=0]$ properly. For example, if the network has down states, every stimulus occurring during this time period will fail to elicit spikes in both upstream and downstream neurons which will bias the estimation of $\mathrm{E}\left[Y|X=0\right]$. For this naive estimator we expect strong confounding if an underlying network state is important.

One standard way to utilize the instrument to get an unbiased estimate of the causal effect β is to use IV. In our case the instrument Z is refractoriness. The IV-estimator is:

$$\hat{\beta}_{IV} = \frac{\mathrm{E}\left[Y|Z=0\right] - \mathrm{E}\left[Y|Z=1\right]}{\mathrm{E}\left[X|Z=0\right] - \mathrm{E}\left[X|Z=1\right]} \tag{2}$$

For a derivation see Section 4.1, note that $\mathrm{E}\left[X|Z=1\right] = 0$ by construction. We should expect this formulation of the problem to lower biases and to become correct in the limit where timescale separation is perfect and the instrument is actually random.

We can now investigate if the use of an instrumental variable (Eq. (2)) gives a better estimate of connectivity strength than (Eq. (1)) or simply analyzing the lagged correlations by means of CCH calculated with Eq. (11). We use the IV estimator given by Eq. (2) on the three neuron system Fig. 1(d) red dashed lines. It converges to the correct causal conclusions that the weights $w_{BC} > 0$ and $w_{AC} \sim 0$ as opposed to OLS and CCH methods (Fig. 1(d) where blue dashed lines are computed by Eq. (1), and yellow lines are computed by Eq. (11)) which both falsely conclude that $w_{AC} > 0$. The technique based on refractoriness helps but is not perfect.

The IV method slightly underestimates connection strength (Fig. 1(d) red dashed lines). There are remaining effects of the network dynamics which affect both neural refractoriness and

downstream activities. To correct for these effects we combine the approaches with a Difference in Differences (DiD) correction[1]. For this we introduce superscript $*$ to indicate that the response window (Fig. 1(b)) is shifted one window-size left. The OLS, DiD estimate is given by Eq. (1) with $\beta_{OLS,DiD} = \beta_{OLS}|_{Y \to Y - Y^*}$ and the IV, DiD estimate is given by Eq. (2) $\beta_{OLS,DiD} = \beta_{OLS}|_{Y \to Y - Y^*, X \to X - X^*}$. This corrects for both the ongoing change and the network state.

When there are no upstream spikes preceding stimulus $Z = 0$ the instrument gives no restriction on stimulus response ($X \in 0, 1$) in contrast to $Z = 1$ which leads to stimulus response to be fully restricted by the refractory state ($X = 0$). Unless there is a strong stimulus, one can not know why there is an upstream spike or not for $Z = 0$. For weak stimulations the IV method will not work very well. We therefore propose another estimator – a hybrid version of the IV and OLS method denoted BR (for brew) given by

$$\hat{\beta}_{BR,DiD} = \mathrm{E}\left[Y - Y^* | X = 1, Z = 0\right] - \mathrm{E}\left[Y - Y^* | X = 0, Z = 1\right]. \tag{3}$$

We may expect this hybrid measure to combine some benefits of the underlying techniques.

Both the IV,DiD and BR,DiD methods converge to the correct causal conclusions that the weights $w_{BC} > 0$ and $w_{AC} \sim 0$ Fig. 1(d) red and purple solid lines respectively, with improved accuracy compared to the IV method. However, the BR,DiD method overestimates the transmission probability in BC Fig. 1(d) right panes purple solid line. Finally, in certain experimental settings we know a-priori that connections are positive, we could rectify estimates (not done here), correcting for negative offset. For the simple three neuron system, the IV,DiD and BR,DiD methods, produce meaningful estimates of the causal interactions between neurons.

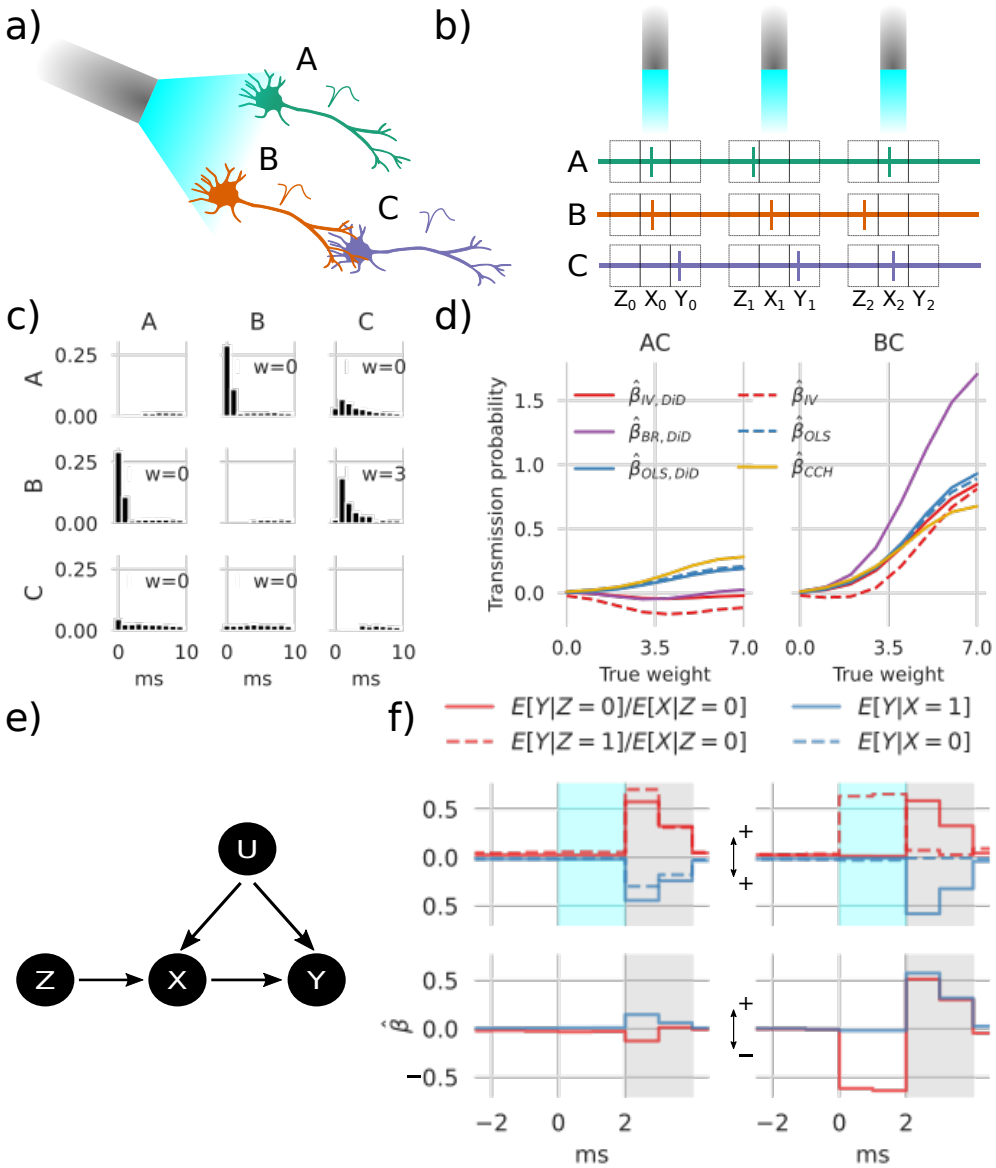

Figure 1: **Instrumental variables corrects spurious correlations induced by optogenetic stimulation**. **a)** Simple network containing three neurons shows stimulation configuration with blue laser light and the connections illustrated by axon touching dendrites. **b)** For each stimulation pulse three spike detection windows are placed relative to stimulus onset time, pre, during and post which are stored as Z, X, Y respectively where Z and X represents either A or B while Y represents C. **c)** Cross correlation histogram (CCH) with the diagonal showing the refractory period, A and B are correlated with each other due to common stimuli which in turn induces correlations between A,C (spurious) and B,C (causal). **d)** Transmission probabilty estimates as a function of true weight between AC, and BC where the AC connection has ground truth at zero. **e)** Directed acyclic graph (DAG) showing the relations between variables. The OLS method only uses X to estimate Y which fails because of the confounding factor U, the refractoriness denoted Z affects X without being affected by U and can therefore be used as an instrumental variable. **f)** Conditional expectancies where the true weight = 7. Response in C is represented by solid lines, or when stimulus is not eliciting spikes due to either refractory states ($Z = 1$), or regardless $X = 0$ represented by dashed lines. Upper panel is mirrored around zero to aid visualization. The estimates in the lower panel represents differences between solid and dashed lines. Shaded gray area depicts estimation window, cyan shading indicates stimulation window, which in this example also depicts DiD reference.

2.3 The strength of neural dynamics characterized by the condition number affects errors

Interacting neurons in a biological network exhibit inhibition and excitation and their interplay produces nontrivial dynamics. If we are estimating the influences between neurons, we are effectively solving an inverse problem. By observing activity and computing the correlation matrix the calculation of coefficients, however performed, will generally contain an estimate of the inverse of this correlation matrix. For example in OLS the Gram matrix $X^T X$ must be inverted which is closely related to the correlation matrix. Such inverses are typically ill-posed problems, only when the matrix has full rank is it even conceptually possible. Moreover, the problem becomes hard, in the sense that noise is amplified along the singular vectors associated with small singular values. The severity of this problem can often be meaningfully described with the condition number, the ratio of the absolute values of the largest to the smallest singular value of the covariance matrix. Even when the problem gets regularized, e.g. in the sense of a pseudoinverse, the condition number is relevant and characterizes how well-posed the inverse problem is. Incidentally, the underlying assumption of the neural manifolds field is often that implicitly the condition number is very large. For many simulated larger dynamical systems similar to the ones we discuss here the condition numbers are large.

Condition numbers are affected by collinearity, which occur when independent variables express a linear relationship in a regression model. When representing neurons as such variables which are driven by a common, unobserved confounder, collinearity may occur. To assess how the proposed estimators performed under such circumstances we simulated a fully connected random neural network of 200 neurons with equal number of excitatory and inhibitory neurons Fig. 2(a). Network dynamics were simulated with a Binomial GLM framework; see Section 4.3. By stimulating 5 neurons we examined the connection strength from these to the rest of the network and produced the results in Fig. 2(b-e). The trial onset times had a temporal Poisson distribution with period 50 ms and were further clipped between 10-200 ms. The connectivity was given by a Gaussian distribution normalized by the square root of its size[17] which forces the largest singular value of the connectivity matrix to be constant as a function of number of neurons.

With higher condition numbers, errors increased in both the IV,DiD and the CCH method for ground truth $w \geq 0$ where the IV method gave smaller errors for lower condition numbers Fig. 2(b). The BR,DiD method gave even smaller errors for lower condition numbers compared to IV,DiD (Fig. 2(c)). The errors for IV,DiD and BR,DiD were significantly lower than all other methods Fig. 2(d). Condition numbers were higher for larger drive strengths but less sensitive to inhibitory drive Fig. 2(e). The network state is the confounder and the condition number is a major way of indexing the level of confounding.

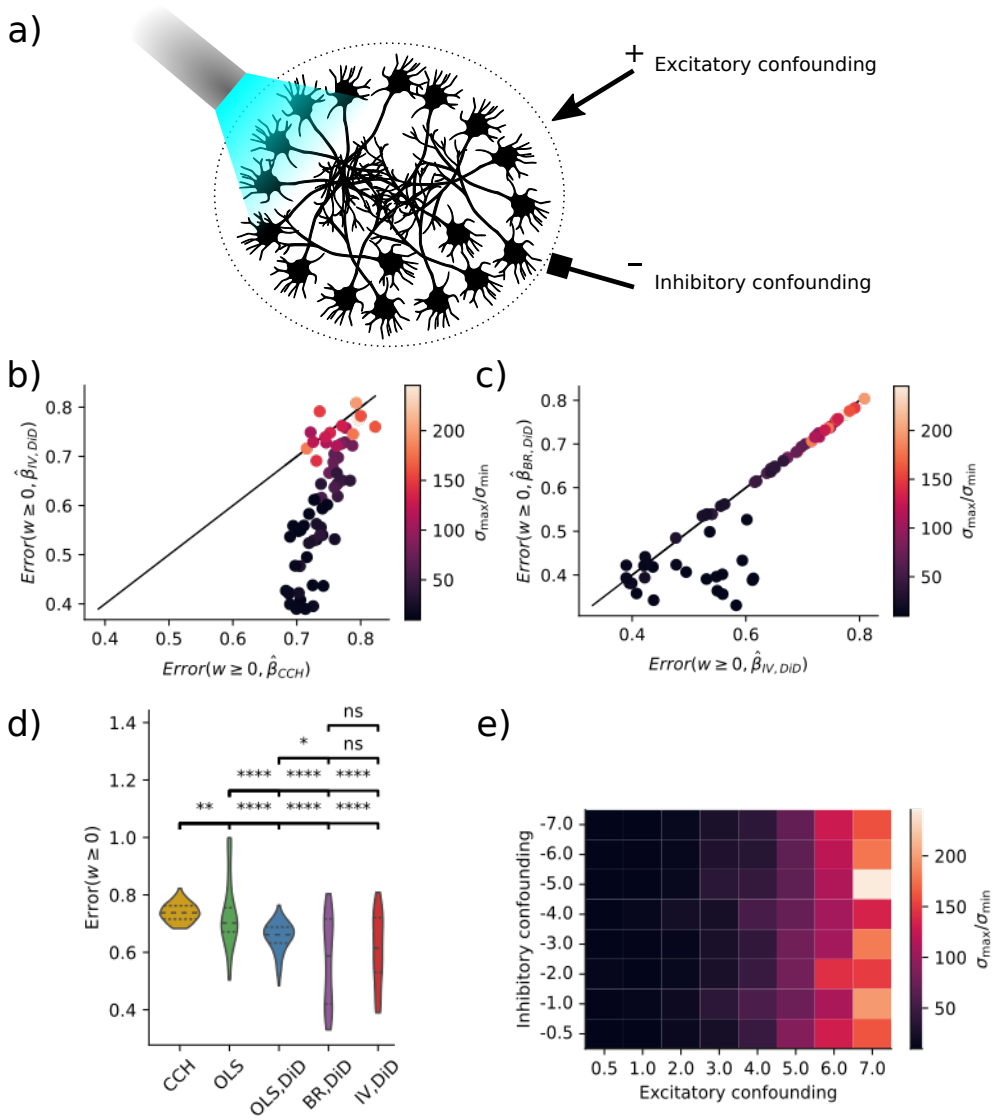

Figure 2: **IV is more robust to collinearity as measured by the condition number compared with CCH. b)** Sketch of the neural network configuration. **b)** Error in IV and CCH for $w \geq 0$ and varying condition numbers as a function of drive strength. **c)** Error in IV and BR for $w \geq 0$ and varying condition numbers as a function of drive strength. **d)** Errors are significantly lower in IV,DiD and BR,DiD than all other methods, each line represents comparison with the estimator corresponding to the left edge.. **e)** Condition number is more affected by excitatory drive than inhibitory drive. Significance vales are encoded as: ns not significant, * $p < 0.05$, ** $p < 0.01$, **** $p < 0.0001$

2.4 Stimulation, connection strength and network size affects condition number and errors

To evaluate the IV method in a more meaningful setting we simulated different stimulation strengths, number of neurons and strength of connectivity while keeping the excitatory and inhibitory drive constant.

By stimulating 5 neurons and examining the connection strength from these to the rest of the network we produced the results in Fig. 3. With higher condition numbers, the IV,DiD method outperformed the CCH method for ground truth $w \geq 0$ Fig. 3(a). The BR,DiD method was more robust for lower condition numbers than IV,DiD (Fig. 3(a-b)), and was generally better than CCH Fig. 3(b). The errors for BR,DiD were significantly lower than all other methods Fig. 3(c). Condition numbers were higher for stronger stimulation and connection strengths and were more sensitive to these variables in larger networks Fig. 3(d).

The hybrid (BR,DiD) method given in Eq. (3) yields lower error for lower stimulation strengths Fig. 3(e) as compared with IV,DiD. Intuitively, we need fairly strong stimuli in order for the downstream neuron to be affected by the instrument, this might explain the smaller errors with stronger stimuli seen in Fig. 3(e) for IV,DiD.

To assert how the estimators performed as binary classifiers, classifying whether there is a connection or not, we computed the area under the receiver operating curve (AUROC). This measures the quality of classification at various levels of thresholds and indicates how well the estimator will function in a classification task where AUROC= 1 indicates a perfect classifier, and AUROC= 0.5 indicates chance level. The BR,DiD method gave in general better classification results as measured with the AUROC, both for various stimulus and connection strengths Fig. 3(g,h).

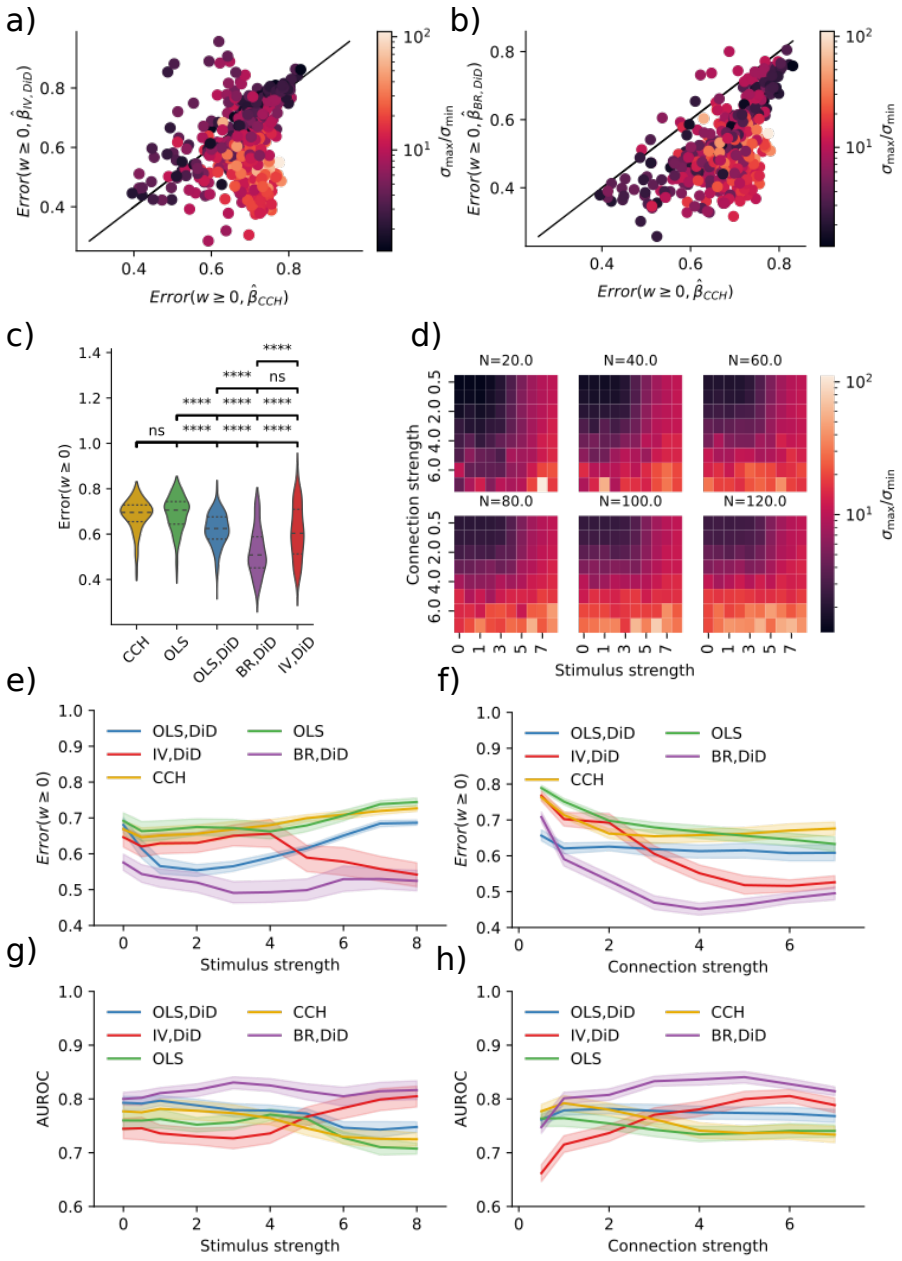

Figure 3: **Errors and corresponding condition numbers are affected by network size, stimulation and connection strengths. a)** Error in IV,DiD and CCH for $w \geq 0$ and varying condition numbers determined by stimulus strength, connection strength and number of neurons. **b)** Error in BR,DiD and CCH for $w \geq 0$ and varying condition numbers determined by stimulus strength, connection strength and number of neurons. **c)** Errors are significantly lower in BR,DiD than all other estimators, each line represents comparison with the estimator corresponding to the left edge. **d)** Condition number as a function of stimulus strength, connection strength and number of neurons, color coded according to colorbar given in (b). **e)** Errors from BR,DiD are in general lower, IV,DiD gets lower with stronger stimuli as opposed to those in OLS, and CCH. **f)** Errors from BR,DiD are in general lower, IV,DiD gets lower with stronger connections whereas those in OLS, and CCH varies less. **g)** AUROC from BR,DiD are in general higher, IV,DiD gets higher with stronger stimuli as opposed to those in OLS, and CCH. **h)** AUROC from BR,DiD are in general higher, IV,DiD gets higher with stronger connections whereas those in OLS, and CCH varies less.

2.5 Connectivity affects statistical power, stimulus percentage affects error

Causal Inference techniques obtain better causal validity at the cost of statistical power. When the number of synapses per neuron increases, estimating the weights should become harder even if we could record from all neurons. To quantify this we simulated 10 networks with 200 neurons without additional confounding factors where we reduced connectivity or equivalently increased sparsity. We found that the error decrease as a function of sparsity in IV,DiD, BR,DiD and OLS (Fig. 4(a)). To further quantify robustness we also simulated 10 networks with 200 neurons where we increased the percentage of stimulated neurons. The error increased with increased percentage of stimuli Fig. 4(b) as would be expected when considering that the condition number increases with increasing stimulus percentage (data not shown).

By plotting the estimates as a function of the true weight and sparsity it is evident that the error is due to a larger spread which for the IV,DiD method decrease with sparsity Fig. 4(c) while not so much for CCH Fig. 4(e), data for BR,DiD was similar to IV,DiD (not shown). When increasing the stimulus percentage the spread increases for IV,DiD but also the centre of the estimates at $w = 0$ is shifted upwards, however the slope seemingly remains steady Fig. 4(d), data was similar for BR,DiD (not shown). For the CCH method the spread increased with increased stimulus percentage, the slope went towards zero and the centre at $w = 0$ is shifted upwards Fig. 4(f).

To reach asymptotic convergence for IV,DiD and BR,DiD a large number of trials is needed ($\sim 10^6$) as seen in Fig. 4(g), for larger sparsity more trials are needed, but with the result of reaching lower error. When increasing the stimulus percentage the convergence curve flattens out while being shifted towards larger errors Fig. 4(h).

The IV,DiD and BR,DiD method introduced here for the estimation of causal effects require considerable amounts of data. The applicability of the proposed method may be limited to cases where weights are sufficiently sparse and stimulation effect is sparsely distributed among neurons with strong stimuli amplitude. It seems likely that the right domain of application for this method would thus not be densely stimulated in-vivo mammalian cortex but systems that are smaller or sparser or, alternatively, scenarios where only a small number of neurons are stimulated.

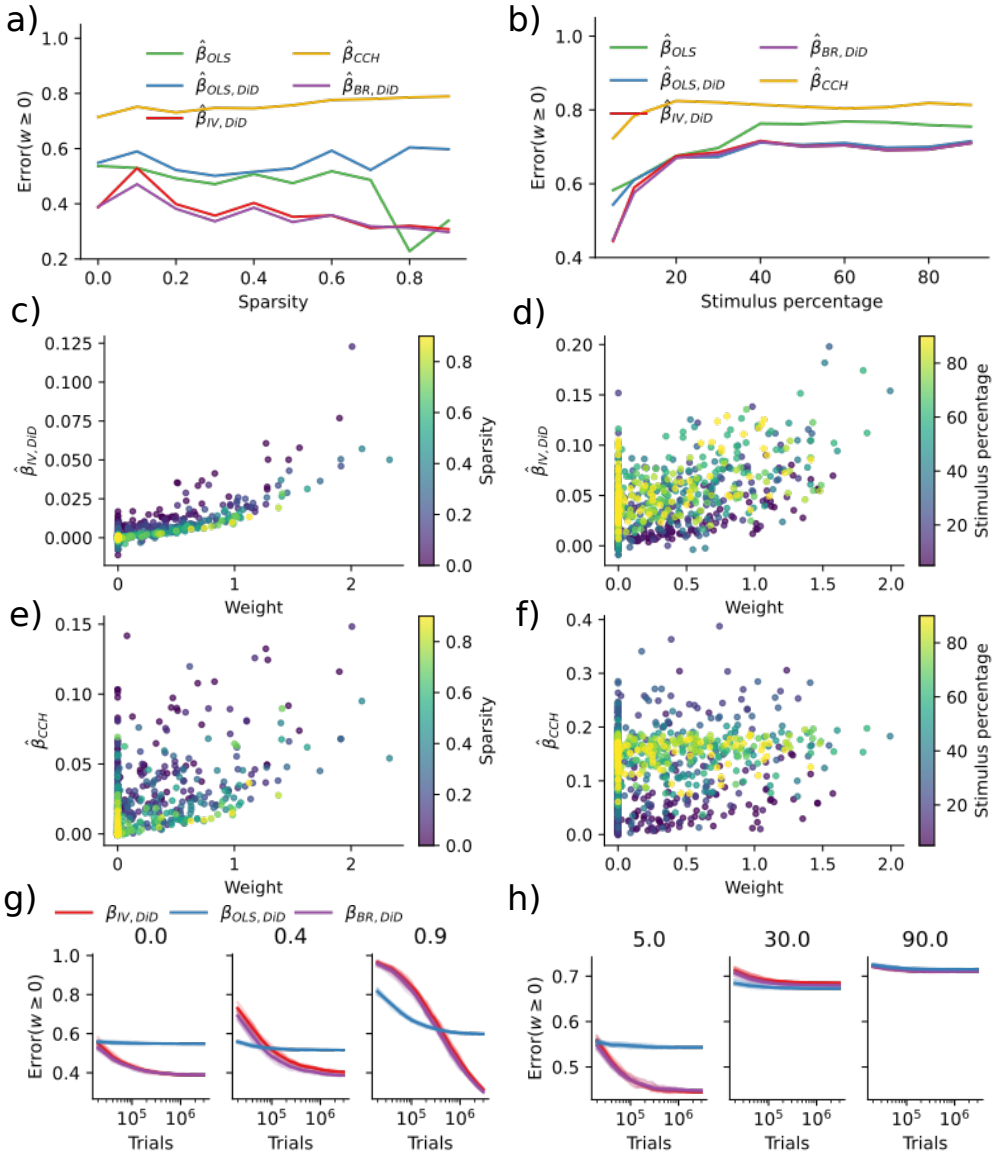

Figure 4: **Error diminishes with sparsity while is increased by increasing stimulus percentage. a)** Error as a function of sparsity. **b)** Error as a function of stimulus percentage. **c)** IV,DiD estimates as a function of true weight and sparsity. **d)** IV,DiD estimates as a function of true weight and stimulus percentage. **e)** CCH estimates as a function of true weight and sparsity. **f)** CCH estimates as a function of true weight and stimulus percentage. **g)** Error as a function of number of stimulus trials, sparsity is denoted in the panel titles. **h)** Error as a function of number of stimulus trials, stimulus percentage is denoted in the panel titles.

2.6 Distribution of optogenetic stimulation effect in realistic scenarios

Results presented in Fig. 4(b) show that a uniform distribution of strong stimuli across the excitatory population is detrimental for the estimation of connectivity, although better with BR,DiD and IV,DiD compared with OLS,(DiD) and CCH. However, a uniform distribution of stimuli strength does not reflect the distribution of light intensity obtained with optogenetic stimulation. Optogenetic stimulation is generally seen as a perturbation that by-and-large affects neurons in proximity ($<$ 1mm) of the light source with the effect decreasing with distance. However, this is a misleading way of conceptualizing the spatial effect of stimulation as it depends on multiple factors. Light intensity and opsin density are important as more light and ion channels will cause a stronger effect on each cell. Moreover, the number of potentially stimulated neurons is critical as more neurons will have a larger impact on the overall population activity. Finally, physiological properties of the cells are important as light may have a stronger effect on spiking activity when the membrane potential of the cell is sufficiently close to the firing threshold. The induced effect of optogenetic stimulation as a function of distance should be given by a relation between four parameters: light intensity, spatial distribution of neurons, distributions of membrane potential across neurons, and the distribution of induced photo currents.

To estimate light intensity, we calculated the spatial extent of laser light delivered by fiber-optics under plausible experimental conditions according to Aravanis et al. [4]; see Section 4.5. While their experiment uses blue light, their estimates assume no absorption making the equations somewhat general. This modeling of light intensity yield an approximately $1/r^2$ reduction with distance r from the stimulation site Fig. 5(a), cyan line. This is explained by the surface of a 3D shell growing with $4\pi r^2$ and photons will be roughly isotropic (but see [45]) beyond the scattering length as depicted in the inset of Fig. 5(a). The same number of photons has to cross each of the spheres around the stimulation location unless they are absorbed or scattered inwards. As a result, the density of photons decreases with distance.

The number of illuminated neurons at a given distance will, however, increase with distance to the stimulation site given that neurons are roughly uniformly distributed in brain tissue (Fig. 5(a), black line). In fact, it will increase by approximately r^2 with distance. This derives from the same surface scaling as for the 3D shell as for the photon flow. Thus, the number of neurons that can be activated increases rapidly with distance.

To estimate the effect of stimulation, the mechanism with which light affects spiking activity needs to be considered. This can largely be characterized by the distribution of membrane potentials across neurons. Surprisingly, this distribution has been observed to be symmetrically distributed and relatively flat [31,11,39]. The expected response from a pulse of light that induces a charge Q should be proportional to the number of affected neurons whose membrane potential sit within a Q/C range of the threshold (C is the capacitance). Given that the distribution of membrane potentials is relatively flat (the density close to the threshold is generally within an order of magnitude of the density of its mode) suggests that the spiking response to a perturbation for any neuron is roughly proportional to the induced photo current. The peak amplitude of the photo current relates approximately logarithmically to the light intensity [46]; see Section 4.5. Assuming that opsins are evenly distributed across neurons, the induced photo current will not be proportional to light intensity - it will fall slower. Based on this, we calculate the overall stimulation effect to be the product of the number of neurons in a spherical slice and the peak amplitude photo current. This product actually increases with distance (up to the distance where absorption becomes important) Fig. 5(a), blue line. In other words, there is more activation at 500um than at 100um – when neurons densely express opsins optogenetics is not a local technique.

To verify this intuition with experimental data we extracted data with WebPlotDigitizer[1]

[1] `apps.automeris.io/`

from[18] (Supplementary figure 2,B1) where an optic fibre was retracted while measuring average spiking activity (Fig. 5(b), blue line). By multiplying the distance squared with this activity we get an experimentally valid estimate of the scaled population spiking activity (Fig. 5(b), orange line), which increase with distance. Thus, optogenetic stimulation utilizing single photon activation experimentally does not produce a localized effect.

With this knowledge at hand we proceeded to simulate more realistic distributions of stimuli strength to evaluate its effect on the proposed estimators. We simulated 5 networks of sizes ranging from 100-500 neurons with additional confounding factors and balanced excitation and inhibition. For stimulation we gave each neuron a random spatial distance from the simulated optogenetic stimulus. The stimulus intensity was then set according to Eq. (15), and was constant throughout the trials. The trial onset had a temporal Poisson distribution with period 50 ms and was further clipped between 10-200 ms. To evaluate the effect of stimulation we measured the hit rate as percentage of successful stimulation trials ($\mathrm{E}[X]$). The hit rate was larger for larger network size and stimulus strength (Fig. 5(c)). The effect of larger hit rate with network size is likely due to more neurons being stimulated with high stimulus strength creating a cascade effect throughout the network. By increasing the maximum stimulus strength we further observed a non-linear increase in hit rate across the population where at some point almost all neurons had hit rate at $\sim 90\%$ (data not shown). Condition numbers increased to large numbers with larger network size (Fig. 5(d)). Errors as measured by the combined error across all hit rates increased with larger network size (Fig. 5(e)). Classifications were less accurate with larger network size (Fig. 5(f)).

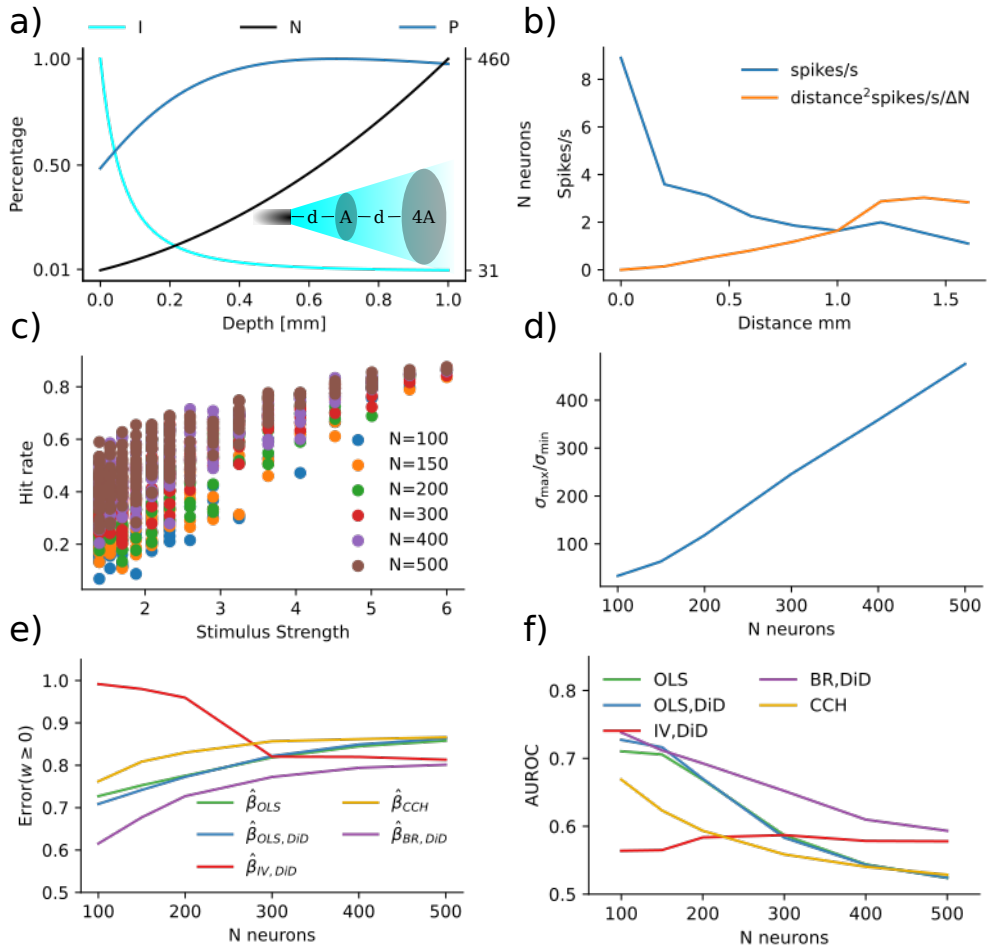

Figure 5: **Spatial extent of optogenetic stimulus.** **a)** Due to scattering and geometric loss the light intensity (I, cyan line) with an intensity of $10mW/mm^2$ exiting the optogenetic fiber follows approximately an inverse square law r^{-2} where r is the distance from the fiber. If neurons are uniformly distributed, the number of affected neurons in a spherical slice increases by r^2 (N, black line). The total photo current (P, blue line) summed over neurons in a spherical slice increased with distance due to the nonlinear relation between light intensity and photo current, depicted as percentage of maximum. **b)** Experimental data for verification[18]. **c)** Hit rate increases relative to stimulus strength with network size. **d)** Condition numbers get severely large with increasing network size. **e)** Errors increase with network size. **f)** AUROC decrease with network size.

2.7 Inhibitory networks

In previous sections we have only explored excitatory connections, however the estimators are also possible to be used with inhibitory connections. To assess effectiveness on inhibitory estimates, we simulated a network of 120 neurons with balanced excitation and inhibition while stimulating 5 inhibitory neurons. By examining the connectivity of these stimulated neurons we generated the results presented in Fig. 6. The IV,DiD method outperformes CCH Fig. 6(a), but is again outperformed by the OLS,DiD method Fig. 6(b), for most variations of condition numbers. Comparing overall results, the OLS,DiD method was significantly better than the other methods while notably BR,DiD performed worse than IV,DiD. Condition numbers were mainly affected by connection strength Fig. 6(c). Errors were smaller for OLS,DiD for all connection and stimulation strengths Fig. 6(e-f). Despite giving poor transmission probability estimates, classification were comparable across methods except for IV,DiD which performed comparably worse Fig. 6(g-h). Estimating the existence of inhibitory connections is relatively easy but estimating their strengths appears harder.

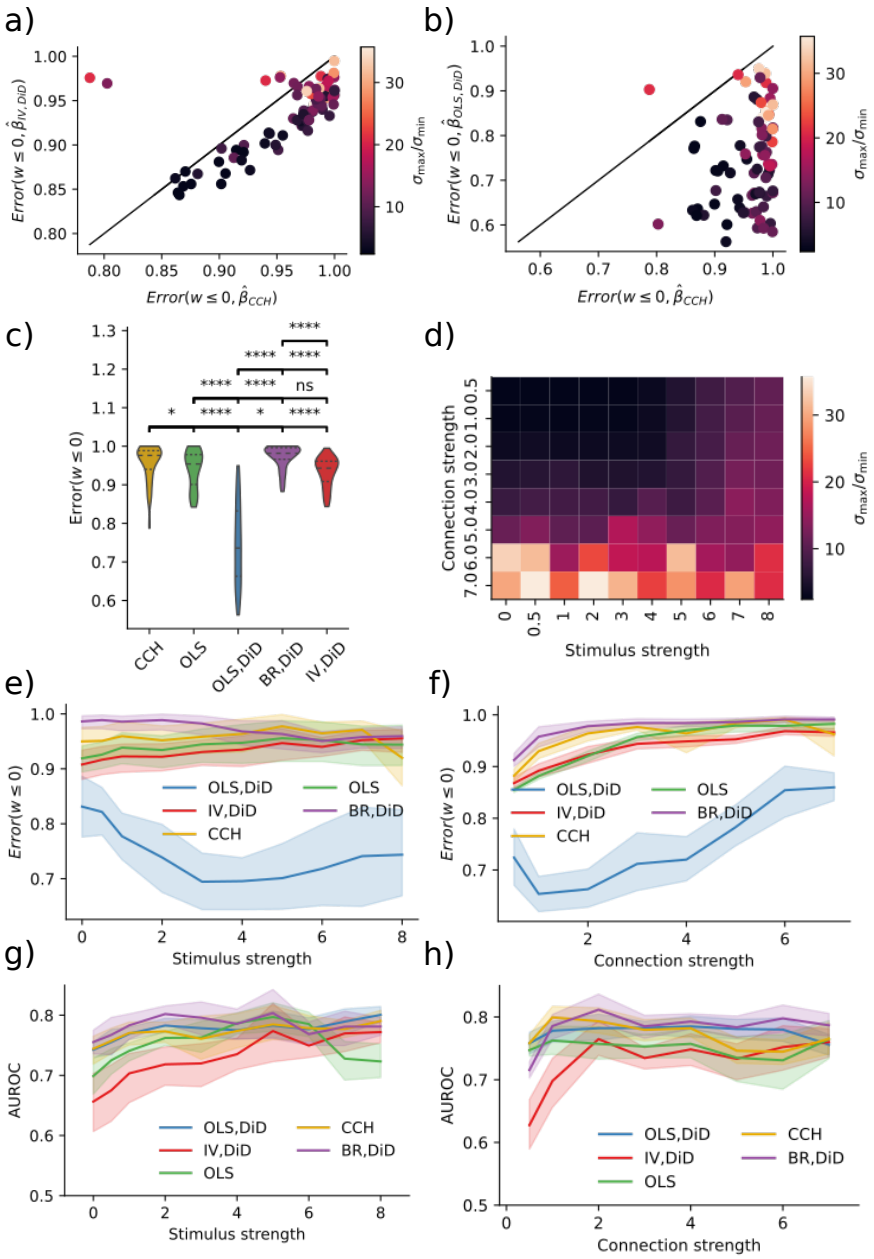

Figure 6: **IV is better than naive, but not OLS,DiD. a)** Error in IV,DiD and CCH for $w \leq 0$ and varying condition numbers as a function of stimulus strength, connection strength and number of neurons. **b)** Error in OLS,DiD and CCH for $w \leq 0$ and varying condition numbers as a function of stimulus strength, connection strength and number of neurons. **c)** Errors are significantly lower in OLS,DiD than all other estimators, each line represents comparison with the estimator corresponding to the left edge. **d)** Condition number as a function of stimulus strength and connection strength, color coded according to colorbar given in (a,b). **e)** Errors from OLS,DiD are in general lower with varying stimulation strength. **f)** Errors from OLS,DiD are in general lower with varying connection strength. **g)** AUROC from are comparable across methods with varying stimulation strength. **h)** AUROC from are comparable across methods with varying connection strength.

3 Discussion

Here we have asked if the refractory period of neurons can be used as an instrumental variable to reverse engineer the causal flow of activity in a network of simulated neurons. We have found that this approach performs considerably better than the naïve approaches. Moreover, we have found that neither a naïve linear regression model nor a naïve cross correlation method produce reliable estimates of connectivity between neuron pairs. The IV approach improves the inference process by looking at the responses that is missing because of refractoriness which effectively allows better estimates of causal effects.

At the moment, we have no ground-truth data set at hand to test our technique and compare with other approaches on real neural activities. Ideally, we would have known causal effects from single-cell stimulation (e.g. from two-photon optogenetics) to establish causal effects. Such data should contain many randomly distributed, short and intensive stimulation trials combined with traditional optogenetics, designed in a way where refractoriness matter. Since most optogenetic protocols use set stimulation frequency and light intensity, such a dataset, to the best of our knowledge, is currently not available and prevents us from testing how good our estimator would work on experimental data. Future experiments are needed to obtain reliable insights.

For the refractory period to be a good instrument, it is necessary that it is not overly affected by the network activity. This will clearly be problematic in many cases. After all, network activity affects neuron activity and hence refractoriness. However, there are multiple scenarios where refractoriness will be a good instrument. For example, if we have balanced excitation and inhibition, we may expect largely independent refractory states of individual neurons. If a neuron biophysically implements something like conditional Poisson spiking, its refractory states will be random conditioned on the network state. Importantly, we may expect the phase of a neuron to be far more random than the activity of the network as a whole.

The randomness of refractory times is the one factor which makes or breaks the IV approach we introduced here. Even if neurons' refractory states are strongly correlated during normal network operation, there may be ways of randomizing refractoriness. First, it would help to use a task and situation where neurons are as uncorrelated as possible. Second, we may use a set of conditioning pulses of stimulation to increase independence of refractory states. Giving one burst of stimulation which is strong enough to elicit multiple spikes from each neuron may effectively randomize their phases[15]. Third, we may utilize chemical, behavioral, or molecular perturbations to produce a good instrumental variable. For example, we may be able to construct intracellular oscillators that are unaffected by neural activities or constructs that force a neuron into quiescence at random times. In neuroscience there has been no effort yet to produce good instrumental variables, so there may be many possibilities for improvements.

One popular way of estimating causal effects is fitting generalized linear models (GLMs) to simultaneously recorded neuron activities[36,38]. GLMs are multiple nonlinear regressions and require multiple neurons to perform well. However, even if activity from all neurons were recorded, GLMs might fail to estimate causal connections[10]. However, complete recordings are not possible in the mammalian brain, especially not in primates, where recordings include only a very small subset of the neurons involved in the actual computation. When using GLMs one may accurately estimate latency distributions and sequences of spikes from individual neurons. These ideas should, arguably, be merged with IV approaches. One of the strengths of the IV estimator presented here is that it only requires one pair to be recorded because we can utilize the randomness of the refractory periods along with random stimulations. Under those assumptions, the IV estimator can produce meaningful causal estimates.

The main problem with optogenetic stimulation, when used to infer connectivity, is its nonlocal property. This is due to the inverse relation between changes in light intensity and af-

fected number of neurons combined with a logarithmic relation between light intensity and photocurrent[46]. In addition, the distribution of membrane potentials across neurons is relatively flat[11,39,31] making neurons highly sensitive to perturbations. One could however, imagine situations where optogenetic activation was more local. If for example, the membrane potential distributions were skewed with the mode far from threshold, a very strong stimulus would be required for a neuron to elicit spikes. There could also be other ways of making optogenetic stimulation more local. For example, if one engineered opsins or brain tissue that are more light absorbent (e.g. by ubiquitously producing melanin) one could stimulate more locally. Having melanin under a ubiquitously expressed promotor in the brain would dramatically make optogenetics more local and should probably be a target for the construction of transgenic animals. How to engineer more localized stimulation is an important problem when causally interrogating a system.

Very weak laser pulses in noisy networks might mainly elicit spikes in very few close-by neurons in each trial[14]. However, the stimulus will still affect the membrane potential of many neurons further away, some of which will spike. Therefore, weak stimulation does not remove the principal problem of correlation based techniques. After all, the network still acts as a confounder and, if anything, the weak stimulation will reduce the statistical power of the approach. Lowering stimulation amplitudes does not appear to be a way of obtaining meaningful causal estimates.

There are many techniques for causal inference, most of which are largely unknown to the field of neuroscience, and are based on approximating randomness in a seemingly regular world. In many cases, one could use regression discontinuity designs in a spiking system[26,22]. Moreover, one could use a difference in difference approach[1]. Matching approaches[44,24], can be used when comparing similar network states and their evolution over time. In general, neuroscience is in a quest for causal interpretations, we should therefore be able to benefit considerably by utilizing techniques that are popular in the field of causal inference.

The IV technique has been used extensively in econometrics and can provide provable causal estimates given three main assumptions[19]. (i) The instrument (Z) must be associated with the regressor (X). (ii) There must be no direct influence of the instrument (Z) on the outcome variable (Y), but only through the regressor variable (X). (iii) The instrument (Z) and the outcome variable (Y) must not share causes. The validity of these assumptions is central when using the IV approach. For a variable Z (e.g. refractoriness) to be a valid instrumental variable it is required that it affects one variable of interest X (spiking in the presynaptic neuron) (Assumption i Section 4.1) which is trivially fulfilled. Moreover, it must only affect the outcome variable Y (the putative postsynaptic neuron) through the variable X (Assumption ii Section 4.1). Finally, Z and Y should not share causal drivers (Assumption iii Section 4.1). In biology this can at best be approximately right but for many scenarios this assumption should be good.

We only care about activity within the stimulation windows, and thus the chain of events $S \to X \to Y$.

Activity in Y is thus assumed to be caused by the controlled stimulation mediated through stimulated neurons, X and U. Moreover, we assume the stimulus duration to be on the same order of magnitude as the refractory period and stimulation onset time t_i random. Finally, we assume the Markov property that trial i is independent of trial $i-1$. As a consequence, Z, the refractory state, cannot share causes with Y, fulfilling Assumption iii. Under the assumptions i-iii and iv described below we can provably correctly quantify the influence of X on the rest of the neurons (Y) in the network[19].

Importantly, there is one realistic short term use of the techniques we proposed here. It is possible to optogenetically stimulate just a local cluster of neurons, say 10 of them. This puts us into the situation of Fig. 4(b,h), where errors can be quite low, at least in the limit of relatively long recordings. As such, applying the methods we introduced here to such experimental data,

ideally with measured ground truth connectivity, is a promising and realistic next step.

4 Methods

4.1 Instrumental Variables

The central idea in causal inference is the counterfactual statement also called potential outcomes which compares what happened with what could have happened under different circumstances. Using this idea the causal effect is then the difference $Y_i|_{do(T=1)} - Y_i|_{do(T=0)}$ where $Y_i|_{do(T=1)}$, typically shortened to $Y_i(1)$, denotes the value Y of subject i given that we force some treatment $T = 1$, denoted by the *do* operator. For example, the treatment T can indicate that a medication is given to patient i and Y_i indicate survival. Obviously, we can not observe the same individual with and without a given treatment, and therefor the causal effect is typically estimated by the average treatment effect (ATE) given by the expected difference $\mathrm{E}\left[Y(1) - Y(0)\right]$.

The IV technique has been used extensively in econometrics and can provide provable causal estimates given three main assumptions[19]. (i) The instrument (Z) must be associated with the regressor (X). (ii) There must be no direct influence of the instrument (Z) on the outcome variable (Y), but only through the regressor variable (X). (iii) The instrument (Z) and the outcome variable (Y) must not share causes. The validity of these assumptions is central when using the IV approach.

Moreover, a final assumption (iv) is required in order to prove causal validity of the estimator Eq. (2), either a linear parametric assumption or a non-parametrically assumption with effect homogeneity or monotonicity[19]. In the following non-parametric derivation using monotonicity we continue to use $Z = 1$ to denote if an upstream spike preceded stimulus, and thus $Z = 1$ indicates that $X = 0$ or in other words X did comply to treatment. This is opposite to conventional notation where $Z = 1 \rightarrow X = 1$, and might confuse readers with previous knowledge from causal inference.

We can stratify stimulus response trials X given Z in four groups, "compliers": $(X(Z = 0) = 1, X(Z = 1) = 0)$, "deniers": $X(Z = 0) = 0, X(Z = 1) = 1$, "always takers": $X(Z = 0) = 1, X(Z = 1) = 1$ and "never takers": $X(Z = 0) = 0, X(Z = 1) = 0$. By making the assumption that there are no "deniers" we get the monotonicity $X_i(Z = 0) \geq T_i(Z = 1) \forall i$. Moreover, it is assumed that the net effect of "always takers" and "never takers" will be zero in Y i.e. for "never takers": $\mathrm{E}\left[Y(Z = 0) - Y(Z = 1)|X(0) = 0, X(1) = 0\right] = 0$. This applies for both groups, but for "always takers" we can also claim that this group is negligible in size because the refractory period forces $X(1) = 0$. Given these assumptions we can derive the compliers average causal effect (CACE) estimator Eq. (2) by

$$\mathrm{E}\left[Y(Z = 0) - Y(Z = 1)\right] = \mathrm{E}\left[Y(Z = 0) - Y(Z = 1)|X(0) = 1, X(1) = 0\right] P(X(0) = 1, X(1) = 0)$$

which further indicates that

$$\mathrm{E}\left[Y(Z = 0) - Y(Z = 1)|X(0) = 1, X(1) = 0\right] = \frac{\mathrm{E}\left[Y(Z = 0) - Y(Z = 1)\right]}{P(X(0) = 1, X(1) = 0)}$$

where by assumption (iii) $\mathrm{E}\left[Y(Z = 0) - Y(Z = 1)\right] = \mathrm{E}\left[Y|Z = 0\right] - \mathrm{E}\left[Y|Z = 1\right]$. The probability for being a "complier" $(P(X(0) = 1, X(1) = 0))$ is given by the probability of not being an "always taker" or "never taker" i.e. $P(X(0) = 1, X(1) = 0) = 1 - P(X = 1|Z = 0) - P(X = 0|Z = 1)$ which can be given by $P(X(0) = 1, X(1) = 0) = \mathrm{E}\left[X|Z = 0\right] - \mathrm{E}\left[X|Z = 1\right]$, QED.

4.2 Difference in Differences

When dealing with temporal data we can use an additional method commonly used for causal inference, namely Difference in Differences (DiD). The effect of treatment in a longitudinal study with two time points Y^*, Y is given by the average treatment effect of the treated (ATT) $\mathrm{E}\left[Y(1) - Y(0) | X = 1\right]$. The ATT can be estimated by taking the differences $(\mathrm{E}\left[Y|X = 1\right] - \mathrm{E}\left[Y^*|X = 1\right]) - (\mathrm{E}\left[Y|X = 0\right] - \mathrm{E}\left[Y^*|X = 0\right])$ where the two differences in parenthesis denotes a treatment group and a control group respectively. The DiD estimate is valid given two assumptions: (i) consistency and (ii) parallel trends; see[3].

To sample the stimulus trials let t_i denote stimulus onset time for stimulus trials $i = 0, 1, 2, ..., N_S$, and spike response windows $[\tau_0^V, \tau_1^V]$ for variables $V \in \{X, Y, Z\}$ which are associated with spikes t_k^V for upstream neurons X, Z and downstream neurons Y. The variables are given by

$$V_i = \begin{cases} 1 & \tau_0^V \leq t_k^V - t_i \leq \tau_1^V, \ \forall t_k^V \\ 0 & \text{else.} \end{cases}$$

The spike windows used for DiD $[\tau_0^{V^*}, \tau_1^{V^*}]$ are given by $[\tau_0^V - \tau^V, \tau_1^V - \tau^V]$ for window size $\tau^V = \tau_1^V - \tau_0^V$ where $\tau_0^V < \tau_1^V$.

4.3 Binomial GLM simulation

To model neuronal populations we use a GLM framework similar to[30]. However, instead of using the Poisson GLM we employed the binomial GLM and instead of fitting the filters to neural data we hand tune them. As discussed in a recent blogpost[2] the Poisson GLM comes with an exponential (unbounded) inverse link function which can produce nonphysical high spike rates. The binomial GLM comes with a (bounded) logistic inverse link function which is ideal for spike simulations given as

$$f(x) = \frac{\exp(x)}{\exp(x) + 1} \tag{4}$$

Spikes were spike-history dependent through the following equations

$$A_k^{(i,j,s)} = \sum_l W^{(l,j,i)} H_k^{(l,s)} \tag{5}$$

$$x_{k+1}^{(i)} \sim \text{Bernoulli}(f(\sum_{s=1}^{S} A_k^{(s,i,S-s)} + b)) \tag{6}$$

Where $\boldsymbol{x}_k = \left[x_k^{(1)}, ..., x_k^{(N)}\right]$ and $\boldsymbol{x}_k \in \{0, 1\}^N$ indicated spikes at time $t = \Delta tk$ where time bins were of size $\Delta t = 1ms$. Further, $H_k \in \mathrm{R}^{N,S}$ denoted spike-history with $S = 10$ number of time steps backwards in time such that $H_k = [\boldsymbol{x}_{k-S}, \boldsymbol{x}_{k+1-S}, ..., \boldsymbol{x}_k]$, initialized as zero for $k - S < 0$. The activation A_k is a tensor contraction between the coupling tensor $W \in \mathrm{R}^{N \times N \times S}$ and the spike-history H. The bias $b = 5$ was hand tuned such that average spike rates were on the order of $\sim 10Hz$.

The spike-indicator vector $\boldsymbol{x}_k$ was given by sampled number of successes l from the Bernoulli distribution with probability p.

[2] https://xcorr.net/2011/10/03/using-the-binomial-glm-instead-of-the-poisson-for-spike-data/ aug6

The coupling tensor W contained cross coupling and refractory filters with the diagonal elements representing the absolute and relative refractory period given by

$$W_{i=j,s} = \begin{cases} -100 \text{ for } s = 1,2,3 \text{ (absolute)} \\ -30\exp(-\frac{1}{2}(s+4)) \text{ for } 4 \leq s \leq S \text{ (relative)} \end{cases} \tag{7}$$

The cross coupling filter was given by

$$W_{i,j,s} = \begin{cases} \hat{W}_{ij}\exp(-\alpha s) \text{ for } 1 \leq s \leq 5 \\ 0 \text{ for } 6 \leq s \leq S \end{cases} \tag{8}$$

Where $\hat{W} \in \mathrm{R}^{N\times N}$ denotes the connectivity matrix, and $\alpha = 0.2$.

The connection strength was given by a normal distribution $W^* \sim \mathcal{N}(0, \sigma/\sqrt{N})$ with zero mean and standard deviation scaled by the square root of the number of neurons. This scaling had two purposes, one to keep the spike variance constant through time [17] and to keep the largest singular value constant as a function of network size [16].

To impose perturbations upon the system additional dimensions to the two first axes of W was added. The simulated optogenetic stimulation strength was given by γ_s and imposed by

$$W_{i=N+1,j,s} = \begin{cases} \gamma_s \text{ for } 1 \leq s < 2 \\ 0 \text{ for } 2 < s \leq S. \end{cases} \tag{9}$$

Stimulation onset times were drawn from a clipped Poisson distribution.

4.4 Cross correlation histogram

The statistical tests giving the probabilities p_{diff} and p_{fast} were done according to [42,14]. Briefly, to test if the cross correlation histogram (CCH) peak was significant we employed two tests. By using the Poisson distribution with a continuity correction [42] given by Eq. (10) we calculated p_{diff} by comparing the peak in positive time lag with the maximum peak in negative time lag, called p_{causal} in English et al. [14]. The probability p_{fast} represents the difference between CCH and it's convolution with a hollow Gaussian kernel [42]. These two measures of significance were required to be < 0.01 and given by

$$p(N|\lambda(m)) = 1 - \sum_{k=0}^{N-1} \frac{e^{-\lambda(m)}\lambda(m)^k}{k!} - \frac{e^{-\lambda(m)}\lambda(m)^N}{2N!}. \tag{10}$$

Here λ represents the counts at bin m and N is the number of bins considered. To naively estimate the connection weight between pairs we used the spike transmission probability defined in English et al. [14] as

$$p_{trans} = \frac{1}{n}\sum_{m=3ms}^{6ms} CCH(m) - \lambda_{Gauss}(m), \tag{11}$$

where n is the number of spikes detected in the presynaptic neuron and $\lambda_{Gauss}(m)$ is the CCH count convolved with a hollow Gaussian kernel at bin m.

name	model 1	model 2	model 3	units
density	~ 1960			Nmm^{-3}
S	10.3			mm^{-1}
NA	0.37			
r	0.1			μm
n	1.36			
n_{Hill}	0.76			
K	0.84			
depth	0.7			mm

Table 1: Stimulation parameters of three different models.

4.5 Realistic optogenetic perturbation intensity

In order to replicate an optogenetic experiment we modeled transmission of light through brain tissue with the Kubelka-Munk model for diffuse scattering in planar, homogeneous media [20] given by

$$T = \frac{1}{Sr + 1}. \tag{12}$$

Here T denotes a transmission fraction, S is the scattering coefficient for mice [4] and r is the distance from a light source. Further we combined diffusion with geometric loss assuming that absorption is negligible as in [4] and computed the intensity as presented in Fig. 1(a) by

$$\frac{I(r)}{I(r=0)} = \frac{\rho^2}{(Sr+1)(r+\rho)^2} \tag{13}$$

where r is the distance from the optical fiber and

$$\rho = \frac{d}{2}\sqrt{\left(\frac{n}{NA}\right)^2 - 1}. \tag{14}$$

Here d is the diameter of the optical fiber, NA is the numerical aperture of the optical fiber and n is the refraction index for gray matter [20]; see numerical values for parameters in Table 1.

To estimate the distribution of light intensity on stimulated neurons we distributed neurons uniformly in 15 spherical slices in the range $[0, 1mm]$ which had a radius given by the cone shaped light; see 1(a) inset.

To further estimate the peak amplitude photo current we used the Hill equation fitted by parameters found in Wang et al. [46] given by

$$P = I_{max}\frac{I^n}{K^n + I^n} \tag{15}$$

Here, $I_{max} = 642pA$ is the maximum current, $n = 0.76$ is the Hill coefficient and $K = 0.84mW/mm^2$ represents the half-maximal light sensitivity of the ChR2. We further used the light intensity I given by Eq. (13) multiplied by an initial intensity of $10mW/mm^2$.

Since the model neurons do not have membrane potentials we scaled the maximum value of input current relative to hit rate close to 100%.

4.6 Computing errors

We computed errors according to

$$\text{Error} = \min_a \frac{||W - a\hat{W}||}{||W||} \tag{16}$$

with $||\cdot||$ denoting the l_2 norm. This error measure is equivalent to the sine of the angle between W and $\hat{W}$.

5 Data availability

All data and code used to produce respective results can be found at `https://github.com/lepmik/causal-optoconnectics`

6 Acknowledgments

This research was partly funded by the National Institutes of Health (NIH) grant R01MH103910 and the University of Pennsylvania. In addition, the Research Council of Norway Grant 231248, 300504, the University of Oslo and Simula Research Laboratory.

References

1. Alberto Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72 (1):1–19, 2005.

2. Laurence Aitchison and Máté Lengyel. With or without you: predictive coding and bayesian inference in the brain. *Current opinion in neurobiology*, 46:219–227, 2017.

3. Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

4. Alexander M Aravanis, Li Ping Wang, Feng Zhang, Leslie A Meltzer, Murtaza Z Mogri, M Bret Schneider, and Karl Deisseroth. An optical neural interface: in vivo control of rodent motor cortex with integrated fiberoptic and optogenetic technology. *J. Neural Eng.*, 4(3):S143–S156, 2007. ISSN 17412560. doi: 10.1088/1741-2560/4/3/S02.

5. Frederico AC Azevedo, Ludmila RB Carvalho, Lea T Grinberg, José Marcelo Farfel, Renata EL Ferretti, Renata EP Leite, Roberto Lent, Suzana Herculano-Houzel, et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513 (5):532–541, 2009.

6. Edward S Boyden, Feng Zhang, Ernst Bamberg, Georg Nagel, and Karl Deisseroth. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature neuroscience*, 8(9):1263, 2005.

7. Gyorgy Buzsaki. *Rhythms of the Brain*. Oxford university press, 2006.

8. David Card. Using geographic variation in college proximity to estimate the return to schooling. Technical report, National Bureau of Economic Research, 1993.

9. Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. *arXiv preprint arXiv:1203.3475*, 2012.

10. Abhranil Das and Ila R Fiete. Systematic errors in connectivity inferred from activity in strongly coupled recurrent circuits. January 2019. doi: 10.1101/512053. URL `https://doi.org/10.1101/512053`.

11. Alain Destexhe and Denis Paré. Impact of network activity on the integrative properties of neocortical pyramidal neurons in vivo. *Journal of neurophysiology*, 81(4):1531–1547, 1999.

12. Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, pages 865–886, 2011.

13. Valentina Emiliani, Adam E Cohen, Karl Deisseroth, and Michael Häusser. All-optical interrogation of neural circuits. *Journal of Neuroscience*, 35(41):13917–13926, 2015.

14. Daniel Fine English, Sam McKenzie, Talfan Evans, Kanghwan Kim, Euisik Yoon, and György Buzsáki. Pyramidal Cell-Interneuron Circuit Architecture and Dynamics in Hippocampal Networks. *Neuron*, 96(2): 505–520, 2017. ISSN 10974199. doi: 10.1016/j.neuron.2017.09.033.

15. G Bard Ermentrout, Roberto F Galán, and Nathaniel N Urban. Reliability, synchrony and noise. *Trends in neurosciences*, 31(8):428–434, 2008.

16. Stuart Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, pages 252–261, 1980.

17. Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

18. Xue Han, Brian Y. Chow, Huihui Zhou, Nathan C. Klapoetke, Amy Chuong, Reza Rajimehr, Aimei Yang, Michael V. Baratta, Jonathan Winkle, Robert Desimone, and Edward S. Boyden. A high-light sensitivity optical neural silencer: Development and application to optogenetic control of non-human primate cortex. *Frontiers in Systems Neuroscience*, 5, 2011. doi: 10.3389/fnsys.2011.00018. URL `https://doi.org/10.3389/fnsys.2011.00018`.

19. Miguel A Hernan and James M Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

20. Aaron Ho Pui Ho, Donghyun Kim, and Michael G Somekh. *Handbook of photonics for biomedical engineering*. Springer Netherlands, 2017.

21. CJ Honey, O Sporns, Leila Cammoun, Xavier Gigandet, Jean-Philippe Thiran, Reto Meuli, and Patric Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040, 2009.

22. Guido W Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635, 2008.

23. Eric Jonas and Konrad Paul Kording. Could a Neuroscientist Understand a Microprocessor? *PLoS Comput. Biol.*, 13(1):1–24, 2017. ISSN 15537358. doi: 10.1371/journal.pcbi.1005268.

24. Gary King and Richard Nielsen. Why propensity scores should not be used for matching. *Copy at http://j.mp/1sexgVw Download Citation BibTex Tagged XML Download Paper*, 378, 2016.

25. Christof Koch. *Biophysics of computation: information processing in single neurons*. Oxford university press, 2004.

26. Benjamin Lansdell and Konrad Kording. Spiking allows neurons to estimate their causal effect. *bioRxiv*, page 253351, 2018.

27. Gilad M Lerman, Jonathan V Gill, Dmitry Rinberg, and Shy Shoham. Two photon holographic stimulation system for cellular-resolution interrogation of olfactory coding. In *Optics and the Brain*, pages BrM3B–5. Optical Society of America, 2017.

28. David Marc Anton Mehler and Konrad Paul Kording. The lure of causal statements: Rampant mis-inference of causality in estimated connectivity. *arXiv preprint arXiv:1812.03363*, 2018.

29. Volodymyr Nikolenko, Kira E Poskanzer, and Rafael Yuste. Two-photon photostimulation and imaging of neural circuits. *Nature methods*, 4(11):943, 2007.

30. Liam Paninski, Jonathan W. Pillow, and Eero P. Simoncelli. Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model. *Neural Computation*, 16(12):2533–2561, December 2004. doi: 10.1162/0899766042321797. URL `https://doi.org/10.1162/0899766042321797`.

31. Denis Paré, Eric Shink, Hélène Gaudreau, Alain Destexhe, and Eric J Lang. Impact of spontaneous synaptic activity on the resting properties of cat neocortical pyramidal neurons in vivo. *Journal of neurophysiology*, 79(3):1450–1460, 1998.

32. Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

33. Judea Pearl. *Causality*. Cambridge university press, 2009.

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

35. David Pfau, Eftychios A Pnevmatikakis, and Liam Paninski. Robust learning of low-dimensional dynamics from large neural ensembles. In *Advances in neural information processing systems*, pages 2391–2399, 2013.

36. Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995, 2008.

37. Didier Pinault. A novel single-cell staining procedure performed in vivo under electrophysiological control: morpho-functional features of juxtacellularly labeled thalamic cells and other central neurons with biocytin or neurobiotin. *Journal of neuroscience methods*, 65(2):113–136, 1996.

38. Yasser Roudi, Joanna Tyrcha, and John Hertz. Ising model for neural data: model quality and approximate methods for extracting functional connectivity. *Physical Review E*, 79(5):051915, 2009.

39. Michelle Rudolph and Alain Destexhe. On the use of analytical expressions for the voltage distribution to analyze intracellular recordings. *Neural computation*, 18(12):2917–2922, 2006.

40. Elad Schneidman, Michael J. Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, April 2006. doi: 10.1038/nature04701. URL `https://doi.org/10.1038/nature04701`.

41. Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

42. Eran Stark and Moshe Abeles. Unbiased estimation of precise temporal correlations between spike trains. *J. Neurosci. Methods*, 179:90–100, 2009. doi: 10.1016/j.jneumeth.2008.12.029.

43. Ian H Stevenson, James M Rebesco, Lee E Miller, and Konrad P Körding. Inferring functional connections between neurons. *Current opinion in neurobiology*, 18(6):582–588, 2008.

44. Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

45. Martin Thunemann, Torbjørn V Ness, Kivilcim Kilic, Christopher G Ferri, Sava Sakadzic, Anders M Dale, Yeshaiahu Fainman, David A Boas, Gaute T Einevoll, and Anna Devor. Does light propagate better along pyramidal apical dendrites in cerebral cortex? In *Optics and the Brain*, pages JW3A–56. Optical Society of America, 2018.

46. H Wang, J Peca, M Matsuzaki, K Matsuzaki, J Noguchi, L Qiu, D Wang, F Zhang, E Boyden, K Deisseroth, et al. High-speed mapping of synaptic connectivity using photostimulation in channelrhodopsin-2 transgenic mice. *Proceedings of the National Academy of Sciences*, 104(19):8143–8148, 2007.

47. Charles Wilson. Up and down states. *Scholarpedia*, 3(6):1410, 2008. doi: 10.4249/scholarpedia.1410. URL `https://doi.org/10.4249/scholarpedia.1410`.

48. Philip G Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

49. Boris V Zemelman, Georgia A Lee, Minna Ng, and Gero Miesenböck. Selective photostimulation of genetically charged neurons. *Neuron*, 33(1):15–22, 2002.