Directed Spectral Measures Improve Latent Network Models Of Neural Populations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Systems neuroscience aims to understand how networks of neurons distributed throughout the brain mediate computational tasks. One popular approach to identify those networks is to first calculate measures of neural activity (e.g. power spectra) in multiple different brain regions, and then apply a linear latent factor model to those measures. Critically, despite the established role of directed communication between brain regions in neural computation, measures of directed communication have been rarely utilized in network estimation because they are incompatible with the implicit assumptions of the linear latent factor model approach. Here, we develop a novel spectral measure of directed communication called the Directed Spectrum (DS). We prove that it is compatible with the implicit assumptions of linear latent factor models, and we provide a method to estimate the DS. We demonstrate that latent linear factor models of DS measures better capture underlying brain networks in both simulated and real neural recording data compared to available alternatives. Thus, linear latent factor models of the DS offer neuroscientists a simple and effective way to explicitly model directed communication in networks of neural populations.

1 Introduction

A major goal in neuroscience is to characterize how populations of neurons work together to carry out computational tasks [1]. A well-known example is the biological neural network that identifies low-level visual features such as edges. It first sends signals from the retina to the lateral geniculate nucleus, then from there to the primary visual cortex, then finally to higher cortical visual processing regions [2, 3]. While some biological neural networks are well known, we expect that the vast majority remain undiscovered due to the enormous variety of tasks the brain performs. Many methods have been developed to help discover latent networks of neural populations (i.e., brain networks) [4–6]. A key aspect of such methods is that they should be interpretable, meaning that a neuroscientist must be able to use the model to draw conclusions about the brain [7]. Unfortunately, many types of common models in the machine learning literature do not offer this type of interpretability [8].

One class of models that is considered interpretable for brain networks and widely used is Linear Latent Factor Models (LLFMs) [9, 10]. This family includes popular methods like Principal Component Analysis (PCA) and Independent Component Analysis (ICA), and are used across almost all modalities in neuroscience to model brain networks. For example, ICA is regularly used in functional magnetic resonance imaging (fMRI) applications to identify latent brain networks such as the default mode network [11, 12]. A common approach in neuroscience is to apply an LLFM to some measure or extracted feature from the recorded signal, rather than to the recorded signals themselves, in order to get a model of brain networks that are characterized in the desired measure. For example, power spectra are often calculated from the raw electroencephalogram (EEG) recordings before an LLFM

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

is applied, yielding brain networks defined by power spectra [13]. In multi-site local field potential (LFP) data, LLFMs have also been applied to identify latent brain networks that are defined by a cross-spectral covariance matrix [14].

One type of measure that could be especially informative for defining latent brain networks is directed communication between brain regions. However, LLFMs that directly incorporate measures of directed communication are lacking. We believe that this is because existing measures of directed communication, such as Granger causality [15], are incompatible with the implicit assumptions of using LLFMs to identify brain networks, as we discuss in Sections 2 and 4. There are a few exceptions. For example, Gallagher et al. [14] attempted to capture this effect by modeling networks in terms of phase shifts in spectral content between neural populations, but it is unclear if such phase shifts are an appropriate proxy for directed transmission of signals, and those models are bottlenecked by significant computational time.

In response to this methodological gap, we introduce a novel measure of directed communication that we will refer to as the Directed Spectrum (DS). The DS estimates directed communication between time series in the frequency domain. We prove that these measures are a linear function of latent brain networks under reasonable assumptions, making them compatible with LLFMs in this application. We then demonstrate that an LLFM of DS measures recovers latent networks in a simulated dataset where the ground truth networks are known. We compare the performance to several competing directed communication measures, including Spectral Granger Causality [16], and show that using the DS results in significantly higher quality reconstruction of the true brain networks. Finally, we show that latent brain networks identified from real neural data via the DS can decode behaviorally relevant information with much higher fidelity.

2 Linear latent factor models (LLFMs) for identifying brain networks

The term *linear latent factor model* describes any model that seeks to approximate a data vector $\boldsymbol{x}_n$ as a linear sum of latent factor vectors plus an additional noise term $\boldsymbol{\epsilon}_n$,

$$\boldsymbol{x}_n = \sum_{j=1}^{J} Z_n^{(j)} \boldsymbol{x}^{(j)} + \boldsymbol{\epsilon}_n. \tag{1}$$

$\boldsymbol{x}_n$ represents a single sample from a larger dataset, $[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$. We assume that $\boldsymbol{x}_n$ contains measures calculated from multi-channel time-series data, $V_n = [\boldsymbol{v}_{1,n}, \ldots, \boldsymbol{v}_{K,n}]'$, where K is the number of channels and $\boldsymbol{v}_{c,n} \in \mathbb{R}^T$ is the row of V_n corresponding to a recording from a single channel. We are often interested in the frequency domain representations of these quantities, which will be marked by a tilde (e.g., $\tilde{v}_{c,n}(\omega)$ is the frequency domain representation of $\boldsymbol{v}_{c,n}$). The vectors $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(J)}$ represent the J different latent factor loadings, and $Z_n^{(j)}$ represents the activation level, or score, of the jth factor in the nth sample. Many well-known models, such as principal component analysis [17], independent component analysis [18], and nonnegative matrix factorizaton [19], are LLFMs.

LLFMs are frequently used for modeling and discovering brain networks [10]. In that context, it is straightforward to interpret the factor scores, $Z_n^{(j)}$, as the activation levels for a set of brain networks. Likewise, we can interpret the factor loading $\boldsymbol{x}^{(j)}$ as the observable signature of the jth brain network. LLFMs are also quite flexible. A wide variety of different assumptions can be made regarding the structure of the factors, the scores, and the noise distribution [6, 9, 20]. The fact that LLFMs are interpretable and adaptable to a wide variety of data types has made them a standard model class for discovering latent brain networks.

2.1 Compatibility between linear latent factor models and measures of neural activity

It is common to model measures of neural activity recordings rather than modeling the recording directly. For example, a typical choice is to convert the observed data into a set of power spectra (S_{cc}) for each channel [13, 21],

$$S_{cc}(\omega) \equiv E[|\tilde{v}_{c,n}(\omega)|^2], \tag{2}$$

where $\tilde{v}_{c,n}(\omega)$ is the element of the frequency domain representation for the cth channel at frequency ω in sample n. Modeling these measures in an LLFM leads to latent networks described in terms of power spectra associated with each channel. The chosen measures dictate the representation of the discovered networks, so we want to use relevant and easy-to-interpret measures.

An implicit assumption in this application of LLFMs is that the modeled measures are *linear* functions of the latent brain networks. This is seen in (1) where the observed measure is a linear function of the latent factor scores $Z_n^{(j)}$ within some noise tolerance. We let $\mathcal{X}(\cdot)$ represent a function that produces the set of measures from our observed neural data V_n.

Definition 1. *We call $\mathcal{X}(\cdot)$ a linear function of latent brain networks if*

$$\mathcal{X}(V_n) = \sum_{j=1}^{J} Z_n^{(j)} \mathcal{X}(V_n^{(j)}), \tag{3}$$

where $V_n^{(j)}$ represents the data that would be observed if only the jth latent network was present with an activation of 1.

It can be shown that power spectra are theoretically consistent with this requirement (see Supplemental Section E), but there are many measures that cannot obey this assumption. For example, the Pearson correlation coefficient and Granger causality are two measures that cannot obey this assumption (see Section 4). Including such measures will lead to suboptimal representations of latent brain networks.

3 Modeling directed communication in brain networks with LLFMs

Measures capturing directed communication between neural populations are rarely included in LLFMs despite their frequent use in neuroscience [22–24]. We believe this is because such measures, including Granger causality [15, 16], are not linear functions of brain networks and therefore are poorly modeled by LLFMs (see Section 4 and Supplemental Section C). Nonetheless, measures of directed communication such as Granger causality are vital for the study of brain networks [25, 26].

Below, we introduce a novel measure of directed communication within a network, which we refer to as the Directed Spectrum (DS). The DS is a spectral measure, meaning it describes the strength of communication associated with different frequencies. In order to derive the DS measures, we first build a general time-series model of latent brain networks in Section 3.1, and then define how latent networks combine to produce the observed signal in Section 3.2. The DS is formally defined in Section 3.3. These DS measures capture communication in a similar manner to Granger causality while also being compatible with a linear model of brain networks.

3.1 Modeling brain networks as independent VAR processes

We outline a model where the output for each latent network is defined by a Vector Autoregressive (VAR) process. For additional background on VAR models see Supplemental Section D. The model outlined below represents a general framework for understanding what the DS measures capture, even though inferring it directly would be quite challenging.

We begin by assuming that we have J latent brain networks that each generate vector timeseries outputs. Rather than slicing by channel as in Section 2, we define the output series as $V_n^{(j)} = [\boldsymbol{v}_{1,n}^{(j)}, \ldots, \boldsymbol{v}_{T,n}^{(j)}]$, where $\boldsymbol{v}_{t,n}^{(j)} \in \mathbb{R}^K$ is the column of V_n representing signal for all channels at time t within sample n. The output series is associated with a VAR process,

$$\boldsymbol{v}_{t,n}^{(j)} = \sum_{\tau=1}^{p_j} A_\tau^{(j)} \boldsymbol{v}_{(t-\tau),n}^{(j)} + \boldsymbol{\sigma}_{t,n}^{(j)}, \qquad \boldsymbol{\sigma}_{t,n}^{(j)} \sim \mathcal{N}(0, Z_n^{(j)} \Sigma^{(j)}). \tag{4}$$

We refer to $\boldsymbol{\sigma}_{t,n}^{(j)}$ as the *innovations* in the jth network at time t, and assume that they are drawn *iid* from a zero mean Gaussian distribution. Each network has a single covariance structure over all samples $\Sigma^{(j)}$ that is scaled by $Z_n^{(j)}$ (the activation level) to define the innovation distribution of the nth sample. The innovation terms represent new signal introduced into the network that cannot be explained by the past network outputs $\boldsymbol{v}_{(t-1),n}^{(j)}, \ldots, \boldsymbol{v}_{(t-p_j),n}^{(j)}$. We can represent this network in the frequency domain given standard assumptions regarding the stability of VAR models [27],

$$\tilde{\boldsymbol{v}}_n^{(j)}(\omega) = H^{(j)}(\omega) \tilde{\boldsymbol{\sigma}}_n^{(j)}(\omega), \qquad H^{(j)}(\omega) \equiv \left(I - \sum_{\tau=1}^{p_j} A_\tau^{(j)} e^{-i\tau\omega} \right)^{-1}, \qquad 0 \leq \omega \leq 2\pi, \tag{5}$$

where $\tilde{\boldsymbol{v}}_n^{(j)}(\omega) \in \mathbb{R}^K$ and $\tilde{\boldsymbol{\sigma}}_n^{(j)}(\omega) \in \mathbb{R}^K$ are the frequency domain representations of the series $\boldsymbol{v}_{1,n}^{(j)}, \ldots, \boldsymbol{v}_{T,n}^{(j)}$ and $\boldsymbol{\sigma}_{1,n}^{(j)}, \ldots, \boldsymbol{\sigma}_{T,n}^{(j)}$, respectively.

3.2 Modeling observed signal as a superposition of latent transmitted signals

In order to complete our model of latent brain networks, we must relate it to the observed signal. We model each observed sample by a VAR in the frequency domain, analogous to the way we modeled latent network outputs,

$$\tilde{\boldsymbol{v}}_n(\omega) = H_n(\omega)\tilde{\boldsymbol{\sigma}}_n(\omega). \tag{6}$$

We assume that the K channels in our data are partitioned into a set of non-overlapping groups, $\mathcal{G} = \{\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}, \dots\}$. For example, each group can be a single channel to model all inter-channel relationships, or could be all channels within a given brain region. We partition the observed data and the transfer matrix $H_n(\omega)$ and innovations $\tilde{\boldsymbol{\sigma}}_n(\omega)$ into these groups,

$$\begin{aligned}
\tilde{\boldsymbol{v}}_n(\omega) &= \begin{bmatrix} \tilde{\boldsymbol{v}}_{\boldsymbol{b},n}(\omega) \\ \tilde{\boldsymbol{v}}_{\boldsymbol{c},n}(\omega) \\ \tilde{\boldsymbol{v}}_{\boldsymbol{d},n}(\omega) \\ \vdots \end{bmatrix}, \quad H_n(\omega) = \begin{bmatrix} H_{\boldsymbol{bb},n}(\omega) & H_{\boldsymbol{bc},n}(\omega) & H_{\boldsymbol{bd},n}(\omega) & \\ H_{\boldsymbol{cb},n}(\omega) & H_{\boldsymbol{cc},n}(\omega) & H_{\boldsymbol{cd},n}(\omega) & \cdots \\ H_{\boldsymbol{db},n}(\omega) & H_{\boldsymbol{dc},n}(\omega) & H_{\boldsymbol{dd},n}(\omega) & \\ & \vdots & & \end{bmatrix}, \\
\tilde{\boldsymbol{\sigma}}_n(\omega) &= \begin{bmatrix} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{b},n}(\omega) \\ \tilde{\boldsymbol{\sigma}}_{\boldsymbol{c},n}(\omega) \\ \tilde{\boldsymbol{\sigma}}_{\boldsymbol{d},n}(\omega) \\ \vdots \end{bmatrix}, \qquad \Sigma_n = \begin{bmatrix} \Sigma_{\boldsymbol{bb},n} & \Sigma_{\boldsymbol{bc},n} & \Sigma_{\boldsymbol{bd},n} & \\ \Sigma_{\boldsymbol{cb},n} & \Sigma_{\boldsymbol{cc},n} & \Sigma_{\boldsymbol{cd},n} & \cdots \\ \Sigma_{\boldsymbol{db},n} & \Sigma_{\boldsymbol{dc},n} & \Sigma_{\boldsymbol{dd},n} & \\ & \vdots & & \end{bmatrix}.
\end{aligned} \tag{7}$$

We assume that the networks outputs and corresponding VAR parameters are partitioned in the same way. The ordering of these groups is arbitrary. Without loss of generality, we will focus on modeling communication between groups $\boldsymbol{b}$ and $\boldsymbol{c}$.

We examine how our VAR models represent communication between groups by noting that

$$\tilde{\boldsymbol{v}}_{\boldsymbol{c},n}(\omega) = \textstyle\sum_{\boldsymbol{g}\in\mathcal{G}} H_{\boldsymbol{cg},n}(\omega)\tilde{\boldsymbol{\sigma}}_{\boldsymbol{g},n}(\omega). \tag{8}$$

The observed signal in $\boldsymbol{c}$ is a sum of contributions from each group $\boldsymbol{g}$, including the self-contribution from $\boldsymbol{c}$. If the innovations in $\boldsymbol{b}$ and $\boldsymbol{c}$ are uncorrelated, $H_{\boldsymbol{cb},n}(\omega)\tilde{\boldsymbol{\sigma}}_{\boldsymbol{b},n}(\omega)$ would unambiguously represent the signal in $\boldsymbol{c}$ that can be attributed to $\boldsymbol{b}$. If the innovations in $\boldsymbol{b}$ and $\boldsymbol{c}$ are correlated, then there is ambiguity regarding which innovations are responsible for which components of the observed signal. For example, the data $\tilde{\boldsymbol{v}}_{\boldsymbol{c},n}(\omega)$ could be still be explained by $\boldsymbol{b}$ contributing $H_{\boldsymbol{cb},n}(\omega)\tilde{\boldsymbol{\sigma}}_{\boldsymbol{b},n}(\omega)$, or alternatively by $\boldsymbol{b}$ only contributing to $\boldsymbol{c}$ via innovations that are uncorrelated with innovations in $\boldsymbol{c}$. We prefer this second explanation and define the contribution of $\boldsymbol{b}$ to the observed signal in $\boldsymbol{c}$ as the *transmitted signal*,

$$\mathcal{TS}_{\boldsymbol{b}\to\boldsymbol{c},n}(\omega) = \begin{cases} H_{\boldsymbol{cb},n}(\omega)\left(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{b},n}(\omega) - \Sigma_{\boldsymbol{bc},n}\Sigma_{\boldsymbol{cc},n}^{-1}\tilde{\boldsymbol{\sigma}}_{\boldsymbol{c},n}(\omega)\right), & \boldsymbol{b} \neq \boldsymbol{c} \\ \left(H_{\boldsymbol{cc},n}(\omega) + \sum_{k=1}^{K} H_{\boldsymbol{c}k,n}(\omega)\Sigma_{k\boldsymbol{c},n}\Sigma_{\boldsymbol{cc},n}^{-1}\right)\tilde{\boldsymbol{\sigma}}_{\boldsymbol{c},n}(\omega), & \boldsymbol{b} = \boldsymbol{c}. \end{cases} \tag{9}$$

The choice to restrict $\mathcal{TS}_{\boldsymbol{b}\to\boldsymbol{c},n}(\omega)$ to only convey uncorrelated innovations is important because it results in a more conservative estimate of the effect of $\boldsymbol{b}$ on $\boldsymbol{c}$ and agrees with the principle of efficient communication that likely governs brain networks [28]. The innovations $\tilde{\boldsymbol{\sigma}}_{\boldsymbol{b},n}(\omega)$ must account for any signals that are induced in $\boldsymbol{b}$ due to the influence of some other unobserved neural population (the so-called "common input" problem [29]). By using $\mathcal{TS}_{\boldsymbol{b}\to\boldsymbol{c},n}(\omega)$ to model that communication, we are saying that the brain would not waste energy to transmit signal that is already common to both $\boldsymbol{b}$ and $\boldsymbol{c}$.

We complete our model of the observed signal $\tilde{\boldsymbol{v}}_n(\omega)$ with the following two assumptions.

Assumption 1. *The innovations associated with any network are independent of the innovations in all other networks.*

Assumption 2. *The transmitted signals of the observed data are the sum of the corresponding transmitted signals for the latent networks,*

$$\mathcal{TS}_{k\to\boldsymbol{c},n}(\omega) = \textstyle\sum_{j=1}^{J} \mathcal{TS}_{k\to\boldsymbol{c},n}^{(j)}(\omega), \tag{10}$$

where the transmitted signal is defined for the latent networks as

$$\mathcal{TS}_{\boldsymbol{b}\to\boldsymbol{c},n}^{(j)}(\omega) = \begin{cases} H_{\boldsymbol{cb}}^{(j)}(\omega)\left(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{b},n}^{(j)}(\omega) - \Sigma_{\boldsymbol{bc}}^{(j)}{\Sigma_{\boldsymbol{cc}}^{(j)}}^{-1}\tilde{\boldsymbol{\sigma}}_{\boldsymbol{c},n}^{(j)}(\omega)\right), & \boldsymbol{b} \neq \boldsymbol{c} \\ \left(H_{\boldsymbol{cc}}^{(j)}(\omega) + \sum_{k=1}^{K} H_{\boldsymbol{c}k}^{(j)}(\omega)\Sigma_{k\boldsymbol{c}}^{(j)}{\Sigma_{\boldsymbol{cc}}^{(j)}}^{-1}\right)\tilde{\boldsymbol{\sigma}}_{\boldsymbol{c},n}^{(j)}(\omega), & \boldsymbol{b} = \boldsymbol{c} \end{cases}. \tag{11}$$

Assumption 1 can be thought of as enforcing the independence of inputs to each network. Assumption 2 is equivalent the following two statements: 1) the propagation of signal within each network is independent of propagation in each other network, which we believe is reasonable approximation of the biological reality, and 2) the observed outputs occur in some medium where simultaneously occurring phenomena obey the laws of superposition, which is a property of many neural activity measurement modalities (e.g. electrical potentials, fluorescence). By combining these two assumptions with (8), we get the result that the observed data are a superposition of the network outputs,

$$\tilde{\boldsymbol{v}}_{\boldsymbol{c},n}(\omega) = \sum_{\boldsymbol{g}\in\mathcal{G}} H_{\boldsymbol{cg},n}(\omega)\tilde{\boldsymbol{\sigma}}_{\boldsymbol{g},n}(\omega) \qquad = \sum_{\boldsymbol{g}\in\mathcal{G}} \mathcal{TS}_{\boldsymbol{g}\to\boldsymbol{c},n}(\omega) \tag{12}$$

$$= \sum_{\boldsymbol{g}\in\mathcal{G}} \sum_{j=1}^{J} \mathcal{TS}^{(j)}_{\boldsymbol{g}\to\boldsymbol{c},n}(\omega) \quad = \sum_{\boldsymbol{g}\in\mathcal{G}} \sum_{j=1}^{J} H^{(j)}_{\boldsymbol{cg}}(\omega)\tilde{\boldsymbol{\sigma}}^{(j)}_{\boldsymbol{g},n}(\omega) = \sum_{j=1}^{J} \tilde{\boldsymbol{v}}^{(j)}_{\boldsymbol{c},n}(\omega). \tag{13}$$

We note that because the relationship above does not contain a noise term, non-physiological sources of signal would simply be represented as additional networks in the model to adapt to additive noise. This completes our model of the relationship between latent brain networks and observed signals.

3.3 Defining the Directed Spectrum

The model defined above provides a framework for determining whether a measure of neural recordings can be considered a linear function of latent brain networks. We define our measure, the Directed Spectrum (DS), as the second moment of the *transmitted signal*,

$$\mathcal{DS}_{\boldsymbol{b}\to\boldsymbol{c}}(\omega) \equiv \mathbb{E}\left[\widetilde{\mathcal{TS}}_{\boldsymbol{b}\to\boldsymbol{c},n}(\omega)\widetilde{\mathcal{TS}}^{*}_{\boldsymbol{b}\to\boldsymbol{c},n}(\omega)\right] = H_{\boldsymbol{cb}}(\omega)\Sigma_{\boldsymbol{b}|\boldsymbol{c}}H^{*}_{\boldsymbol{cb}}(\omega), \tag{14}$$

$$\Sigma_{\boldsymbol{b}|\boldsymbol{c}} = \Sigma_{\boldsymbol{bb}} - \Sigma_{\boldsymbol{bc}}\Sigma^{-1}_{\boldsymbol{cc}}\Sigma^{*}_{\boldsymbol{bc}},$$

where $\Sigma_{\boldsymbol{b}|\boldsymbol{c}}$ represents the innovation variance for $\boldsymbol{b}$ conditioned on the innovations in $\boldsymbol{c}$. The DS represents the portion of the power spectrum of $\boldsymbol{c}$ that is explained by signal transmitted from $\boldsymbol{b}$ (see Supplemental Section E for more details). It is a linear function of the latent networks when data is appropriately modeled as described above (for proof, see Supplemental Section A).

The DS can be estimated efficiently via factorization of the cross-spectral matrix associated with the observed data [30, 31]. Additional algorithmic details can be found in Supplemental Section B.1.

These DS measures can be calculated according to (14) for data partitioned into any number of groups greater than one. In some situations it is desirable to calculate the DS values by factorizing a reduced cross-spectral matrix for each pair of groups, with all other groups removed. We refer to this method as the *Pairwise Directed Spectrum* (PDS). One benefit of the PDS is that it simplifies the handling of missing channels from the data, since the PDS for the non-missing groups of channels can still be calculated without any adjustments.

4 Related measures of directed communication

The Directed Spectrum is a necessary development because related measures typically used in neuroscience are not linear functions of latent brain networks. We first examine Granger causality, which measures the degree to which one signal or group of signals helps predict future values of another signal or group [15]. Spectral Granger causality is a modification that separates Granger causality into components associated with individual frequencies [16]. It is a leading spectral measure of directed communication in neuroscience [22, 23], and we view it as conceptually the closest measure to the DS. The spectral Granger causality from b to c can be defined in terms of the DS,

$$\boldsymbol{f}_{\boldsymbol{b}\to\boldsymbol{c}}(\omega) = \ln\left(\frac{|S_{\boldsymbol{cc}}(\omega)|}{|S_{\boldsymbol{cc}}(\omega) - \mathcal{DS}_{b\to c}(\omega)|}\right). \tag{15}$$

Spectral Granger causality is not a linear function of the network models described in Section 3. To see this, consider the case where only the jth network is present without any other signals being added to it. Note that changing $Z_n^{(j)}$ would then have no impact on $\boldsymbol{f}_{\boldsymbol{b}\to\boldsymbol{c}}(\omega)$ because both the numerator and denominator in (15) would scale equally with $Z_n^{(j)}$. Standard (non-spectral) Granger causality is equivalent to a scaled integral over all frequencies of the spectral Granger causality [16], meaning that it also would remain constant as $Z_n^{(j)}$ changes.

There are several other spectral measures of directed communication used in neuroscience, including phase slope index [32], partial directed coherence [33], and the directed transfer function [34]. Each

of these can not be a linear function of latent brain networks under reasonable assumptions. We offer the following brief explanations here, with more details in Supplemental Section C. The phase slope index cannot capture bidirectional transmission of signal (all signal is considered unidirectional) [35]. Partial directed coherence and the directed transfer function are both based in VAR models like the DS, but they ignore the innovation term $\tilde{\sigma}_n(\omega)$, which accounts for unobserved components of the networks and for network activation level in our model.

There are a number of other non-spectral measures of directed communication used in neuroscience [23, 24]. The cross-correlation function between activity in different neural populations has been used to identify directed communication between neuronal populations [36]. In these applications it has been used to identify the lag associated with unidirectional transmission between populations, and is considered unreliable in situations involving bidirectional communication [23]. Transfer entropy provides an information theoretic measure of directed communication. In our application transfer entropy suffers from the same problem as Granger causality, where scaling the activation of a single network results in the same value [37, 38]. We discuss additional methods for measuring directed communication between neural populations in Supplemental Section C.

One final method worth mentioning here is dynamic causal modeling (DCM) [39]. DCM produces a generative model to explain observed neural activity based on latent connectivity. Directed connectivity is represented by parameters of the model, which contrasts with our application where measures of directed connectivity are calculated first before being used as features in an LLFM. DCM has even been extended in order to represent spectral content in neural signals [40]. One substantial difference between DCM and the approach we suggest here is that DCM is not designed to segregate the directed connectivity out into subnetworks in an unsupervised manner. Instead, DCM is typically used to identify differences in connectivity between two known conditions [41].

5 Directed Spectrum improves identification of latent networks

To test the effectiveness of the DS and LLFMs, we generated a simulated dataset composed of known latent networks and corresponding activation levels. It contains 10,000 independent recordings, each containing 5 channels sampled for five seconds at 500 Hz. The recorded signals were generated as a sum of contributions from three independent networks. A vector autoregressive (VAR) model for each network generated those network contributions as a different random draw for each recording with parameters that were fixed over all recordings. Each VAR model was designed to generate a particular pattern of oscillations and delayed directional transmission of signals. See Figure 1 for the network properties and Figure 2 for a visualization of the superposition. See Supplemental Section B.2 for additional details on the design of those VAR models.

We generated a model of the latent networks in the simulated dataset by applying an LLFM to DS measures. We calculated the DS for frequencies between 1 Hz and 50 Hz at 1 Hz intervals, between all directed pairs of channels. A non-negative matrix factorization model (i.e. LLFM) was trained using the Itakura-Saito (IS) divergence loss, with an L1 penalty on the factor loadings and scores, and multiplicative update steps for optimization [20]. The IS divergence loss was chosen because it is related to the gamma distribution negative log-likelihood and is more appropriate than mean square error loss for power. After training, the NMF factors were re-ordered to correspond to the matching network in the simulation, based on maximizing the average Spearman's correlation between the estimated and true network activation levels. For comparison, we repeated this process with the Pairwise Directed Spectrum (PDS), spectral Granger causality (GC) [16], phase-slope index (PSI) [32], directed transfer function (DTF) [34], and partial directed coherence (PDC) [33] features substituted for the DS features. Because PSI represents the direction of shared information with sign (i.e. positive/negative), we set negative PSI values to zero and leave the opposite direction as positive; this allowed us to continue to use non-negative matrix factorization to identify latent networks.

We evaluated how well each model recovered the underlying network activation levels using Spearman's correlation. The model of DS features performed significantly better in this regard than the 'non-linear' features (see Table 1). The PDS model also performed significantly better than the 'non-linear' models, and gave comparable results to the DS model. We have visualized the estimated scores along with the true scores for each window for the DS features and Granger causality features in Figure 3. We see that the score estimates of the DS model are much more tightly spread around the

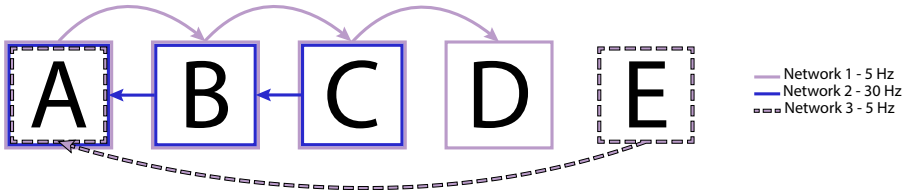

Figure 1: Graphical representation of the simulated latent networks. The five channels are represented by the letters A, B, C, D, and E. A box around a channel indicates the corresponding network induces oscillations in that channel. Arrows indicate propagation of signal at some delay. Networks 1 and 3 have a predominant frequency of 5 Hz (both in purple), and network 2 has a predominant frequency of 30 Hz (in blue).

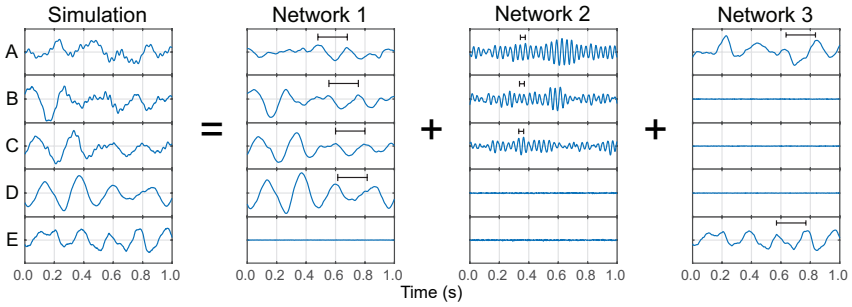

Figure 2: Simulated recordings are a sum of latent network contributions. The simulated recording (left) is a sum of signals generated by each of the three networks. Each panel contains 5 signals associated with the 5 regions over the same one second period. Black scale bars indicate the period for oscillations associated with each network, and are placed to show that the distance between peaks in the signal is approximately one period.

Table 1: Spearman's correlation between latent network activation estimates and true activation levels. GC: Granger causality; PSI: phase slope index; DTF: directed transfer function; PDC: partial directed coherence; DS: directed spectrum; PDS: pairwise directed spectrum. Values in *[brackets]* represent the 95% confidence interval [42].

Measure	Network 1	Network 2	Network 3
GC	0.485 *[0.468, 0.502]*	0.442 *[0.424, 0.459]*	0.281 *[0.261, 0.300]*
PSI	0.387 *[0.368, 0.405]*	0.135 *[0.115, 0.155]*	0.248 *[0.229, 0.268]*
DTF	0.426 *[0.408, 0.443]*	0.131 *[0.111, 0.151]*	0.542 *[0.526, 0.557]*
PDC	0.560 *[0.545, 0.575]*	0.154 *[0.134, 0.174]*	0.445 *[0.427, 0.462]*
DS	0.920 *[0.916, 0.923]*	0.905 *[0.901, 0.909]*	0.927 *[0.924, 0.930]*
PDS	0.908 *[0.904, 0.912]*	0.917 *[0.913, 0.920]*	0.916 *[0.913, 0.920]*

true score values, with the relationship between true scores and estimated scores being very weak in the model of GC features.

Finally, we tested whether the models produced accurate and interpretable representations of the true latent networks from the dataset. The model of DS features successfully recovers all directional influences of each network in the corresponding factors (see Figures 4, S1). Spurious detection of network components was limited to connections between true nodes in the network at the frequency associated with the network, and these spurious influences were small relative to the true detected influences. While the model of GC features did also recover all directional influences of the networks underlying the simulated data, it incorrectly spread the directional influences associated with each network out over multiple estimated networks. This resulted in the GC model identifying more spurious influences with greater relative amplitudes.

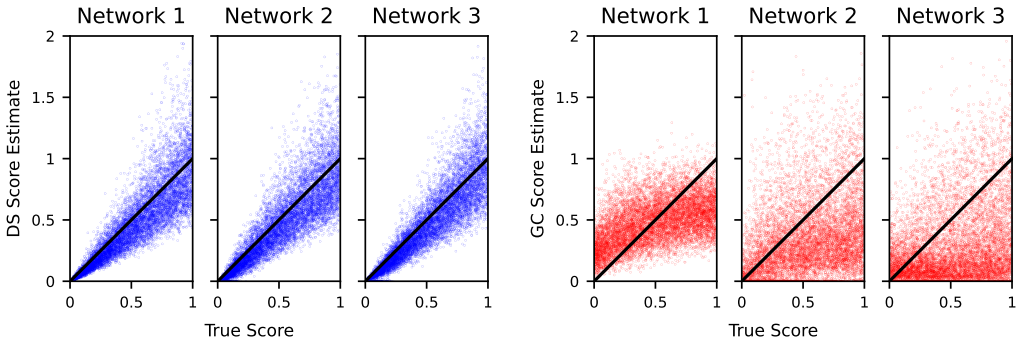

Figure 3: True vs. estimated network scores. The estimated scores for the model trained using DS features are plotted in blue, the estimated scores for a model trained on Granger causality features are plotted in red. The line in black demonstrates the expected trend for a model that perfectly recovers the latent network activations.

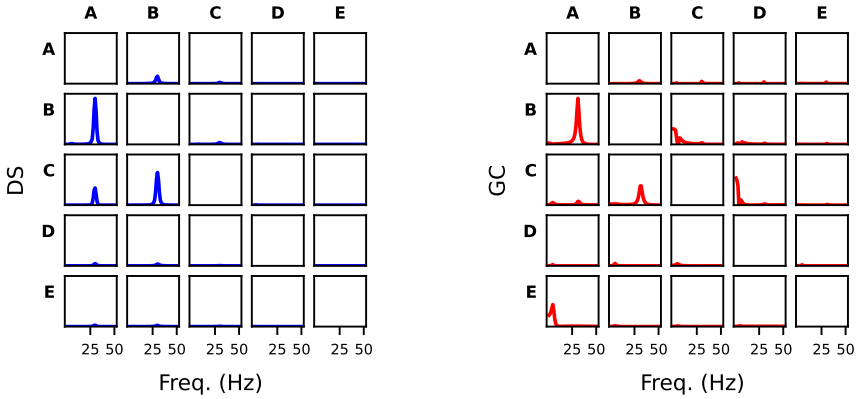

Figure 4: Estimated directional influences in Network 2 for a model of DS features (blue) and Granger causality features (red). Within each grid, a plot corresponds to signal that is being transmitted from the channel listed on the left to the channel listed above.

6 Directed Spectrum improves brain state decoding on neural recordings

We next tested how effectively networks identified by the Directed Spectrum decode behaviorally relevant variables from neural recordings, as a proxy for determining whether they reflect 'real' networks. The dataset consists of local field potentials simultaneously recorded from 11 brain regions in 26 mice (originally published in [25]). In each recording session local field potentials were recorded from a mouse while it was exposed to three different behavioral contexts of successively increasing stress levels: resting in the home cage, exploring an open field, and a tail suspension test. The tail suspension test is commonly used to investigate learned helplessness [43].

In order to train networks that could be used to decode behavioral context, we first divided the dataset into time windows with a duration of 1 second. For each time window, we calculated the DS for all directed pairs of brain regions, for frequencies from 1 to 56 Hz at 1 Hz intervals. A nonnegative matrix factorization model with L1 factorization was trained using the Itakura-Saito divergence objective [44] in order to identify latent brain networks from the DS measures. A multinomial logistic regression classifier with L1 regularization was then applied to the latent factor scores for decoding the behavioral context. In order to obtain estimates of the spread of the decoding performance and choose optimal hyperparameter values, a 5-fold nested cross-validation procedure was used, where each mouse was associated with only one split. The average one-vs-all area under the reciever operating characteristic curve (AUC) across all three behavioral contexts was used as the evaluation metric. The hyperparameters tuned during cross-validation were the number of NMF factors (20, 40, 80), NMF regularization strength (1000, 100, 10, 0), and logistic regression regularization strength

(10, 1). Similar to the experiment described in Section 5, the procedure was repeated using the Pairwise Directed Spectrum (PDS) and the comparison measures listed in that section.

Our training procedure resulted in a one-vs-all AUC for each of the behavioral contexts, for each of the 5 splits, for each of the measure types (see Table 2). We performed a 2-factor repeated measures ANOVA and determined that there was a significant difference associated with the measure types ($F = 92.9$, $p = 4 \times 10^{-13}$). A Tukey's HSD post-hoc test revealed that there was not a significant difference between models of the pairwise and non-pairwise versions of DS ($p = 0.90$), but that both performed significantly better than models of the other measures ($p \leq 0.001$ for all).

Table 2: Behavioral Context Decoding Performance. The columns 'HC AUC', 'OF AUC', and 'TS AUC' report the mean and standard error of the one-vs-all AUC across 5 splits for the homecage, open field, and tail suspension behavioral contexts, respectively. The 'Mean AUC' column reports the average across the mean AUCs reported for each behavioral context.

Measure	*Mean AUC*	HC AUC	OF AUC	TS AUC
GC	0.830	0.828 ± 0.018	0.825 ± 0.021	0.837 ± 0.009
PSI	0.674	0.676 ± 0.014	0.684 ± 0.015	0.661 ± 0.016
DTF	0.758	0.772 ± 0.017	0.781 ± 0.014	0.720 ± 0.022
PDC	0.721	0.730 ± 0.021	0.782 ± 0.016	0.652 ± 0.012
DS	$\mathbf{0.908}$	0.894 ± 0.016	0.916 ± 0.012	0.915 ± 0.007
PDS	$\mathbf{0.919}$	0.909 ± 0.014	0.915 ± 0.014	0.932 ± 0.005

7 Discussion

We have shown that our novel measure of directed communication, the Directed Spectrum (DS), is a linear function of latent brain networks under reasonable assumptions, and so is compatible with LLFMs for characterizing latent brain networks. We saw drastically improved recovery of networks in simulated data with a known ground truth. In real neural recordings, the DS improved decoding of neurally relevant environmental variables. These results demonstrate that the DS allows for more accurate and interpretable models of latent networks defined by directed communication.

A limitation that could be explored in future work is that the DS assumes that the network states are stationary over a single time window (see Supplemental Section D). This limitation is common to almost all methods that assess spectral content of neural activity. Because of this, the DS is only theoretically grounded when applied to time windows that are shorter than the time expected for substantial change to be occur in the latent brain state. In practice, time window lengths of one to five seconds have been considered a relatively stable period in studies of emotional processing [45, 46].

We have only explored LLFMs here, but a variety of nonlinear latent factor models exist for modeling latent brain networks [5, 47]. Nonlinear models are especially useful in applications where predictive performance is the only priority, such as brain-computer interfaces. When the primary goal is to drive scientific understanding of the brain, nonlinear models are less desirable because it is challenging to relate the parameters of such models to relevant conclusions about brain function [8]. We believe that LLFMs and the DS are an appropriate, efficient, and reproducible baseline approach in both scientific and prediction-based applications, while noting that other more expressive models may lead to even better predictive performance at the cost of interpretability.

The development of the DS provides a straightforward way to generate linear models of latent brain networks defined by directed communication. We view this as an important advancement because directed communication between neural populations is a critical component of the way neuroscientists understand brain networks. Furthermore, the DS is likely useful for studying latent networks in many other fields as well (e.g. latent networks of directed internet traffic). We view the major strength of the Directed Spectrum is that it enhances the expressiveness of LLFMs by allowing them to accurately model directed communication within networks. Thus, we have expanded the capabilities of LLFMs while retaining the level of interpretability that make them attractive models in practice.

References

[1] Bratislav Mišić and Olaf Sporns. From regions to connections and networks: new bridges between brain and behavior. *Current Opinion in Neurobiology*, 40:1–7, October 2016. ISSN 0959-4388. doi: 10.1016/j.conb.2016.05.003. URL https://www.sciencedirect.com/science/article/pii/S095943881630054X.

[2] Javier Cudeiro and Adam M. Sillito. Looking back: corticothalamic feedback and early visual processing. *Trends in Neurosciences*, 29(6):298–306, June 2006. ISSN 0166-2236. doi: 10.1016/j.tins.2006.05.002. URL https://www.sciencedirect.com/science/article/pii/S0166223606000907.

[3] Dale Purves, George J Augustine, David Fitzpatrick, William C Hall, Anthony-Samuel LaMantia, James O McNamara, and Leonard E White. Neuroscience. 4th. *Sunderland, Mass.: Sinauer. xvii*, 857:944, 2008.

[4] Heejong Bong, Zongge Liu, Zhao Ren, Matthew Smith, Valerie Ventura, and Kass E. Robert. Latent Dynamic Factor Analysis of High-Dimensional Neural Recordings. *Advances in Neural Information Processing Systems*, 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/beb04c41b45927cf7e9f8fd4bb519e86-Abstract.html.

[5] Shu Zhang, Qinglin Dong, Wei Zhang, Heng Huang, Dajiang Zhu, and Tianming Liu. Discovering hierarchical common brain networks via multimodal deep belief network. *Medical Image Analysis*, 54:238–252, May 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.03.011. URL https://www.sciencedirect.com/science/article/pii/S1361841518308119.

[6] Alex H. Williams, Tony Hyun Kim, Forea Wang, Saurabh Vyas, Stephen I. Ryu, Krishna V. Shenoy, Mark Schnitzer, Tamara G. Kolda, and Surya Ganguli. Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron*, 98(6):1099–1115.e8, June 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.05.015. URL https://www.sciencedirect.com/science/article/pii/S0896627318303878.

[7] Mai-Anh T. Vu, Tülay Adalı, Demba Ba, György Buzsáki, David Carlson, Katherine Heller, Conor Liston, Cynthia Rudin, Vikaas S. Sohal, Alik S. Widge, Helen S. Mayberg, Guillermo Sapiro, and Kafui Dzirasa. A Shared Vision for Machine Learning in Neuroscience. *Journal of Neuroscience*, 38(7):1601–1607, February 2018. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0508-17.2018. URL https://www.jneurosci.org/content/38/7/1601. Publisher: Society for Neuroscience Section: TechSights.

[8] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://www.nature.com/articles/s42256-019-0048-x. Number: 5 Publisher: Nature Publishing Group.

[9] Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized Low Rank Models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, June 2016. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000055. URL https://www.nowpublishers.com/article/Details/MAL-055. Publisher: Now Publishers, Inc.

[10] John P. Cunningham and Byron M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, November 2014. ISSN 1097-6256. doi: 10.1038/nn.3776. URL http://www.nature.com/neuro/journal/v17/n11/full/nn.3776.html.

[11] Vince D. Calhoun, Jingyu Liu, and Tülay Adalı. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage*, 45(1, Supplement 1):S163–S172, March 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.10.057. URL https://www.sciencedirect.com/science/article/pii/S1053811908012032.

[12] Vesa Kiviniemi, Tapani Vire, Jukka Remes, Ahmed Abou Elseoud, Tuomo Starck, Osmo Tervonen, and Juha Nikkinen. A Sliding Time-Window ICA Reveals Spatial Variability of the Default Mode Network in Time. *Brain Connectivity*, 1(4):339–347, September 2011. ISSN 2158-0014. doi: 10.1089/brain.2011.0036. URL https://www.liebertpub.com/doi/abs/10.1089/brain.2011.0036. Publisher: Mary Ann Liebert, Inc., publishers.

[13] Fumikazu Miwakeichi, Eduardo Martinez-Montes, Pedro A. Valdés-Sosa, Nobuaki Nishiyama, Hiroaki Mizuhara, and Yoko Yamaguchi. Decomposing EEG data into space–time–frequency components using Parallel Factor Analysis. *NeuroImage*, 22(3):1035–1045, July 2004. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2004.03.039. URL https://www.sciencedirect.com/science/article/pii/S1053811904001958.

[14] Neil Gallagher, Kyle R Ulrich, Austin Talbot, Kafui Dzirasa, Lawrence Carin, and David E Carlson. Cross-Spectral Factor Analysis. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6845–6855, 2017. URL http://papers.nips.cc/paper/7260-cross-spectral-factor-analysis.pdf.

[15] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, 1969. ISSN 0012-9682. doi: 10.2307/1912791. URL https://www.jstor.org/stable/1912791. Publisher: [Wiley, Econometric Society].

[16] John Geweke. Measurement of Linear Dependence and Feedback between Multiple Time Series. *Journal of the American Statistical Association*, 77(378):304–313, June 1982. ISSN 0162-1459. doi: 10.1080/01621459.1982.10477803. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477803.

[17] Hervé Abdi and Lynne J. Williams. Principal component analysis. *WIREs Computational Statistics*, 2(4):433–459, 2010. ISSN 1939-0068. doi: https://doi.org/10.1002/wics.101. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.101.

[18] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, June 2000. ISSN 0893-6080. doi: 10.1016/S0893-6080(00)00026-5. URL https://www.sciencedirect.com/science/article/pii/S0893608000000265.

[19] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, June 2013. ISSN 1558-2191. doi: 10.1109/TKDE.2012.51. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[20] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

[21] Mitul Kumar Ahirwal and Narendra Londhe. Power spectrum analysis of eeg signals for estimating visual attention. *International Journal of computer applications*, 42(15):22–25, 2012.

[22] Lionel Barnett and Anil K. Seth. The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. *Journal of Neuroscience Methods*, 223:50–68, February 2014. ISSN 1872-678X. doi: 10.1016/j.jneumeth.2013.10.018.

[23] André M. Bastos and Jan-Mathijs Schoffelen. A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls. *Frontiers in Systems Neuroscience*, 9, January 2016. ISSN 1662-5137. doi: 10.3389/fnsys.2015.00175. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4705224/.

[24] Karl J. Friston. Functional and Effective Connectivity: A Review. *Brain Connectivity*, 1(1):13–36, January 2011. ISSN 2158-0014. doi: 10.1089/brain.2011.0008. URL https://www.liebertpub.com/doi/abs/10.1089/brain.2011.0008. Publisher: Mary Ann Liebert, Inc., publishers.

[25] David Carlson, Lisa K. David, Neil M. Gallagher, Mai-Anh T. Vu, Matthew Shirley, Rainbo Hultman, Joyce Wang, Caley Burrus, Colleen A. McClung, Sunil Kumar, Lawrence Carin, Stephen D. Mague, and Kafui Dzirasa. Dynamically Timed Stimulation of Corticolimbic Circuitry Activates a Stress-Compensatory Pathway. *Biological Psychiatry*, June 2017. ISSN 0006-3223. doi: 10.1016/j.biopsych.2017.06.008. URL http://www.sciencedirect.com/science/article/pii/S0006322317316682.

[26] Alexander Z. Harris and Joshua A. Gordon. Long-Range Neural Synchrony in Behavior. *Annual Review of Neuroscience*, 38(1):171–194, 2015. doi: 10.1146/annurev-neuro-071714-034111. URL http://dx.doi.org/10.1146/annurev-neuro-071714-034111.

[27] Raquel Prado and Mike West. *Time Series: Modeling, Computation, and Inference*. CRC Press, May 2010. ISBN 978-1-4200-9336-0. Google-Books-ID: kUhKLLdGGZ4C.

[28] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, March 2009. ISSN 1471-0048. doi: 10.1038/nrn2575. URL https://www.nature.com/articles/nrn2575. Number: 3 Publisher: Nature Publishing Group.

[29] Jayant E Kulkarni and Liam Paninski. Common-input models for multiple neural spike-train data. *Network: Computation in Neural Systems*, 18(4):375–407, 2007.

[30] G. Tunnicliffe Wilson. A convergence theorem for spectral factorization. *Journal of Multivariate Analysis*, 8(2):222–232, June 1978. ISSN 0047-259X. doi: 10.1016/0047-259X(78)90073-8. URL https://www.sciencedirect.com/science/article/pii/0047259X78900738.

[31] Mukeshwar Dhamala, Govindan Rangarajan, and Mingzhou Ding. Estimating Granger Causality from Fourier and Wavelet Transforms of Time Series Data. *Physical Review Letters*, 100(1):018701, January 2008. doi: 10.1103/PhysRevLett.100.018701. URL https://link.aps.org/doi/10.1103/PhysRevLett.100.018701. Publisher: American Physical Society.

[32] Guido Nolte, Andreas Ziehe, Vadim V. Nikulin, Alois Schlögl, Nicole Krämer, Tom Brismar, and Klaus-Robert Müller. Robustly Estimating the Flow Direction of Information in Complex Physical Systems. *Physical Review Letters*, 100(23):234101, June 2008. doi: 10.1103/PhysRevLett.100.234101. URL https://link.aps.org/doi/10.1103/PhysRevLett.100.234101. Publisher: American Physical Society.

[33] Luiz A. Baccalá and Koichi Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological Cybernetics*, 84(6):463–474, May 2001. ISSN 1432-0770. doi: 10.1007/PL00007990. URL https://doi.org/10.1007/PL00007990.

[34] M. J. Kaminski and K. J. Blinowska. A new method of the description of the information flow in the brain structures. *Biological Cybernetics*, 65(3):203–210, July 1991. ISSN 1432-0770. doi: 10.1007/BF00198091. URL https://doi.org/10.1007/BF00198091.

[35] Martin Vinck, Lisanne Huurdeman, Conrado A. Bosman, Pascal Fries, Francesco P. Battaglia, Cyriel M. A. Pennartz, and Paul H. Tiesinga. How to detect the Granger-causal flow direction in the presence of additive noise? *NeuroImage*, 108:301–318, March 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2014.12.017. URL https://www.sciencedirect.com/science/article/pii/S105381191401009X.

[36] Jose-Manuel Alonso, W. Martin Usrey, and R. Clay Reid. Precisely correlated firing in cells of the lateral geniculate nucleus. *Nature*, 383(6603):815–819, October 1996. ISSN 1476-4687. doi: 10.1038/383815a0. URL https://www.nature.com/articles/383815a0. Number: 6603 Publisher: Nature Publishing Group.

[37] Terry Bossomaier, Lionel Barnett, Michael Harré, and Joseph T. Lizier. *An introduction to transfer entropy: information flow in complex systems*. Springer International Publishing, January 2016. ISBN 978-3-319-43221-2 978-3-319-43222-9. doi: 10.1007/978-3-319-43222-9. URL https://researchoutput.csu.edu.au/en/publications/an-introduction-to-transfer-entropy-information-flow-in-complex-s.

[38] Lionel Barnett, Adam B. Barrett, and Anil K. Seth. Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Physical Review Letters*, 103(23):238701, December 2009. doi: 10.1103/PhysRevLett.103.238701. URL https://link.aps.org/doi/10.1103/PhysRevLett.103.238701. Publisher: American Physical Society.

[39] K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302, August 2003. ISSN 1053-8119. doi: 10.1016/S1053-8119(03)00202-7. URL https://www.sciencedirect.com/science/article/pii/S1053811903002027.

[40] K.J. Friston, A. Bastos, V. Litvak, K.E. Stephan, P. Fries, and R.J. Moran. DCM for complex-valued data: Cross-spectra, coherence and phase-delays. *Neuroimage*, 59(1):439–455, January 2012. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2011.07.048. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3200431/.

[41] K. E. Stephan, W. D. Penny, R. J. Moran, H. E. M. den Ouden, J. Daunizeau, and K. J. Friston. Ten simple rules for dynamic causal modeling. *NeuroImage*, 49(4):3099–3109, February 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2009.11.015.

[42] Douglas G. Bonett and Thomas A. Wright. Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65(1):23–28, March 2000. ISSN 1860-0980. doi: 10.1007/BF02294183. URL https://doi.org/10.1007/BF02294183.

[43] H. M. Abelaira, G. Z. Reus, and J. Quevedo. Animal models as tools to study the pathophysiology of depression. *Revista Brasileira de Psiquiatria*, 35:S112–S120, 2013.

[44] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis. *Neural Computation*, 21(3):793–830, March 2009. ISSN 0899-7667. doi: 10.1162/neco.2008.04-08-771. URL https://doi.org/10.1162/neco.2008.04-08-771.

[45] Rainbo Hultman, Kyle Ulrich, Benjamin D. Sachs, Cameron Blount, David E. Carlson, Nkemdilim Ndubuizu, Rosemary C. Bagot, Eric M. Parise, Mai-Anh T. Vu, Neil M. Gallagher, Joyce Wang, Alcino J. Silva, Karl Deisseroth, Stephen D. Mague, Marc G. Caron, Eric J. Nestler, Lawrence Carin, and Kafui Dzirasa. Brain-wide Electrical Spatiotemporal Dynamics Encode Depression Vulnerability. *Cell*, 173 (1):166–180.e14, March 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.02.012. URL `http://www.sciencedirect.com/science/article/pii/S0092867418301569`.

[46] W. Zheng and B. Lu. Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development*, 7(3): 162–175, September 2015. ISSN 1943-0604. doi: 10.1109/TAMD.2015.2431497.

[47] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, December 2000. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.290.5500.2319. URL `https://science.sciencemag.org/content/290/5500/2319`. Publisher: American Association for the Advancement of Science Section: Report.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The claims made in the abstract and introduction reflect the evidence provided in Sections 3, 5 and 6, and Supplemental Sections A and B.1.
   (b) Did you describe the limitations of your work? [Yes] We provide a discussion of limitations in the Discussion (Section 7).
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We include a brief discussion of negative societal impacts in Supplemental Section F. The neuroscience applications in this work offer very limited potential for negative societal impacts, unless we resort to substantial speculation about interaction with future technology. This makes the discussion somewhat trivial and we don't think it merits inclusion in the main text.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The necessary assumptions are stated in Section 3
   (b) Did you include complete proofs of all theoretical results? [Yes] The theoretical result that DS is linear with respect to our model of latent brain networks, given the model in Section 3, is proved in the Supplemental Section A
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code to generate the simulated data and to reproduce the results described in Section 5 will be provided as supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For the simulated data experiment we provide 95% confidence intervals of the Spearman correlation metric in Table 1. For the local field potential dataset we provide the standard error of the mean AUC across 5 cross-validation splits in Table 2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This is discussed in Supplemental Section B.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite the creators of the dataset used in Section 6. We also cite previous work that we incorporated in designing a method to estimate the Directed Spectrum.

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include code for calculating the Directed Spectrum from vector timeseries data that has been formated into time windows.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]