

Causal density and integrated information as measures of conscious level

BY ANIL K. SETH*, ADAM B. BARRETT AND LIONEL BARNETT

*Sackler Centre for Consciousness Science, and School of Informatics,
University of Sussex, Brighton BN1 9QJ, UK*

An outstanding challenge in neuroscience is to develop theoretically grounded and practically applicable quantitative measures that are sensitive to conscious level. Such measures should be high for vivid alert conscious wakefulness, and low for unconscious states such as dreamless sleep, coma and general anaesthesia. Here, we describe recent progress in the development of measures of dynamical complexity, in particular *causal density* and *integrated information*. These and similar measures capture in different ways the extent to which a system's dynamics are simultaneously differentiated and integrated. Because conscious scenes are distinguished by the same dynamical features, these measures are therefore good candidates for reflecting conscious level. After reviewing the theoretical background, we present new simulation results demonstrating similarities and differences between the measures, and we discuss remaining challenges in the practical application of the measures to empirically obtained data.

Keywords: consciousness; causal density; integrated information

1. Introduction

A key objective for consciousness science is to develop and test what can be called 'explanatory correlates': neural processes that not only *correlate with* but also *account for* fundamental properties of conscious experience [1]. One such property is that conscious scenes are simultaneously *integrated* (i.e. they are experienced 'all of a piece') and *differentiated* (i.e. they are composed of many different parts such that each conscious scene is one among a vast repertoire of possible scenes). Having a measure of conjoined integration and differentiation (more generally, *dynamical complexity*) in neural dynamics could account for this fundamental property of consciousness, in much the same way that measures of synchrony and coherence may account for and not merely correlate with the binding of different modalities in visual perception [2,3]. Such measures would most readily apply to conscious *level* (a position on a scale from total unconsciousness as in brain death or coma to full alert awake consciousness) rather than conscious *contents* (the components or qualia comprising a given conscious scene), though extension to the latter case represents an important objective [4].

*Author for correspondence (a.k.seth@sussex.ac.uk).

One contribution of 11 to a Theme Issue 'The complexity of sleep'.

Over recent years, several measures of dynamical complexity have been proposed and related to consciousness. These include *neural complexity*, *causal density* and *integrated information* (Φ) [5–7]. However, convincing application of these measures to experimental data remains challenging. Our aim in this paper is to review the theory underpinning causal density (CD) and integrated information, emphasizing recent work encouraging practical application (we discuss neural complexity only briefly). We describe simulations investigating the relations between the different measures, and we discuss some remaining obstacles to their efficient and interpretable use in common neuroimaging contexts such as magneto/electroencephalography (M/EEG) and functional magnetic resonance imaging (fMRI).

(a) A note on notation

We use a standard mathematical vector/matrix notation in which bold type generally denotes vector quantities and upper-case type denotes matrices or random variables, according to context (for random variables, lower case is used to indicate a particular realization). All vectors are considered to be *column* vectors. ‘ \oplus ’ denotes *vertical concatenation*, so that for $\mathbf{x} = (x_1, \dots, x_n)^T$ and $\mathbf{y} = (y_1, \dots, y_m)^T$, $\mathbf{x} \oplus \mathbf{y}$ is the vector $(x_1, \dots, x_n, y_1, \dots, y_m)^T$, where ‘ T ’ denotes the transpose operator. Given random vectors \mathbf{X} and \mathbf{Y} , we denote by $\Sigma(\mathbf{X})$ the $n \times n$ matrix of covariances $\text{cov}(X_i, X_j)$, and by $\Sigma(\mathbf{X}, \mathbf{Y})$ the $n \times m$ matrix of cross-covariances $\text{cov}(X_i, Y_\alpha)$. We shall make use of the quantity

$$\Sigma(\mathbf{X}|\mathbf{Y}) =: \Sigma(\mathbf{X}) - \Sigma(\mathbf{X}, \mathbf{Y})\Sigma(\mathbf{Y})^{-1}\Sigma(\mathbf{X}, \mathbf{Y})^T, \quad (1.1)$$

which we call the partial covariance of \mathbf{X} given \mathbf{Y} [8]. (If \mathbf{X} and \mathbf{Y} are both multi-variate Gaussian variables, then the partial covariance $\Sigma(\mathbf{X}|\mathbf{Y})$ is precisely the covariance matrix of the conditional variable $\mathbf{X}|\mathbf{Y} = \mathbf{y}$, for any \mathbf{y}). If \mathbf{X}_t is a random vector in discrete time, we use $\mathbf{X}_t^{(p)} \equiv \mathbf{X}_t \oplus \mathbf{X}_{t-1} \oplus \dots \oplus \mathbf{X}_{t-p+1}$ to denote \mathbf{X}_t itself, along with $p - 1$ lags. Given the lag p , we often use the shorthand $\mathbf{X}_t^- \equiv \mathbf{X}_{t-1}^{(p)}$ for the lagged variable, and drop the subscript ‘ t ’ if there is no confusion. Other notations will be introduced as they appear.

2. Causal density

(a) Granger causality and univariate causal density

Causal density is a measure of the overall causal interactivity sustained by a system. It leverages the econometric concept of Granger causality (G-causality), which is a measure of causal influence based on time-series inference. According to G-causality, given variables X and Y , Y G-causes X if, in an appropriate statistical sense, Y assists in predicting the future of X beyond the degree to which X already predicts its own future. In the more general, conditional case [9], Y is said to G-cause X , conditional on Z , if Y assists in predicting the future of X beyond the degree to which X and Z together already predict the future of X .

Given a set of time series, G-causality is typically implemented using the framework of linear autoregression. To measure the G-causality from Y (‘predictor’ variable) to X (‘predictee’ variable) given Z (conditional variable),

we compare the following multi-variate autoregressive (MVAR) models [10]:

$$\left. \begin{aligned} X_t &= A(X_{t-1}^{(p)} \oplus Z_{t-1}^{(r)}) + \varepsilon_t \\ \text{and} \quad X_t &= A'(X_{t-1}^{(p)} \oplus Y_{t-1}^{(q)} \oplus Z_{t-1}^{(r)}) + \varepsilon'_t. \end{aligned} \right\} \quad (2.1)$$

Thus, the ‘predictee’ variable X is regressed firstly on the previous p lags of itself plus r lags of the conditioning variable Z and secondly, in addition, on q lags of the predictor variable Y (p, q and r can be selected according to the Akaike or Bayesian information criterion [11]). The magnitude of the G-causality interaction is then given by the logarithm of the ratio of the residual variances,

$$\mathcal{F}_{Y \rightarrow X|Z} =: \ln \left(\frac{\text{var}(\varepsilon_t)}{\text{var}(\varepsilon'_t)} \right) = \ln \left(\frac{\Sigma(\varepsilon_t)}{\Sigma(\varepsilon'_t)} \right) = \ln \left(\frac{\Sigma(X|X^- \oplus Z^-)}{\Sigma(X|X^- \oplus Y^- \oplus Z^-)} \right), \quad (2.2)$$

where the final term expresses G-causality in terms of partial covariances. The statistical significance of a G-causality value can be assessed either by a χ^2 test on \mathcal{F} or by examining the F -statistic for the regressions (2.1), in either case using appropriate corrections for multiple comparisons (alternatively, permutation or bootstrap resampling can be used instead) [11].

Given a set of G-causality values among elements of a system \mathbf{X} , a simple version of causal density (CD) can be defined as the average of all pairwise G-causalities between elements (conditioning on all remaining elements),

$$\text{CD}(\mathbf{X}) =: \frac{1}{n(n-1)} \sum_{i \neq j} \mathcal{F}_{X_i \rightarrow X_j | \mathbf{X}_{[ij]}}, \quad (2.3)$$

where $\mathbf{X}_{[ij]}$ denotes the subsystem of \mathbf{X} with variables X_i and X_j omitted, and n is the total number of variables. Causal density provides a principled measure of dynamical complexity inasmuch as elements that are completely independent will score zero, as will elements that are completely integrated in their dynamics. High values will only be achieved when elements behave somewhat differently from each other, in order to contribute novel potential predictive information, and at the same time are globally integrated, so that the potential predictive information is in fact useful [7,12].

(b) Multi-variate G-causality and extended causal density

As with most time-series measures, G-causality is standardly assessed between single (univariate) variables, perhaps conditioned on a set of other variables. However, relevant causal interactions within a system may take place between *groups* of variables. For example, in neural systems, one may wish to examine causal interactions among ‘Hebbian’ ensembles of neurons [13] or, at a macroscopic level, among networks of regions-of-interest (ROIs) distributed throughout the brain. More generally, measured variables (observables) are constrained by methods of data acquisition and need not map cleanly onto explanatorily relevant decompositions of the studied system.

Fortunately, it is straightforward to extend G-causality to the multi-variate case in which G-causality interactions are assessed among sets of variables ($\mathbf{X}, \mathbf{Y}, \mathbf{Z}$) rather than only among univariate variables (X, Y, Z). Following

Geweke [9], we can define *multi-variate G-causality* (MVGC) as

$$\mathcal{F}_{Y \rightarrow X|Z} =: \ln \left(\frac{|\Sigma(\boldsymbol{\varepsilon}_t)|}{|\Sigma(\boldsymbol{\varepsilon}'_t)|} \right) = \ln \left(\frac{|\Sigma(\mathbf{X}|\mathbf{X}^- \oplus \mathbf{Z}^-)|}{|\Sigma(\mathbf{X}|\mathbf{X}^- \oplus \mathbf{Y}^- \oplus \mathbf{Z}^-)|} \right), \quad (2.4)$$

where $|\cdot|$ represents the matrix determinant and $|\Sigma(\boldsymbol{\varepsilon})|$ is the *generalized variance* of the residual covariance matrix $\Sigma(\boldsymbol{\varepsilon})$, which quantifies the volume in which the residuals lie. As we have discussed in detail previously [14], the determinant (generalized variance) formulation of MVGC has important advantages over an alternative formulation [15] based on the *trace* (total variance) of the residual covariance matrix. In brief, the determinant formulation is fully equivalent to transfer entropy (see §2*c*) under Gaussian assumptions, is invariant under a wider range of variable transformations, is expandable as a sum of standard univariate G-causalities, and admits a satisfactory spectral decomposition. Numerically, evidence indicates that it is just as stable as the trace formulation [14].

MVGC suggests an extension to CD in which G-causality interactions are assessed across bipartitions of a system. For a system \mathbf{X} , we define $\text{CD}_{k \rightarrow r}(\mathbf{X})$, as the average MVGC from a subset of size k to a subset of size r , conditioned on the rest of the system,

$$\text{CD}_{k \rightarrow r}(\mathbf{X}) =: \frac{1}{n_{k,r}} \sum_{i=1}^{n_{k,r}} \mathcal{F}_{\mathbf{V}_i^k \rightarrow \mathbf{U}_i^r | \mathbf{W}_i^{n-k-r}}, \quad (2.5)$$

where $\mathbf{X} = \mathbf{V}_i^k \cup \mathbf{U}_i^r \cup \mathbf{W}_i^{n-k-r}$ denotes the i th of the $n_{k,r} \equiv \binom{n}{k} \binom{n-k}{r}$ distinct tripartitions of \mathbf{X} into disjoint subsystems of respective sizes k , r and $(n - k - r)$. The BCD is then the average of $\text{CD}_{k \rightarrow (n-k)}(\mathbf{X})$ over predictor size k ,

$$\text{BCD}(\mathbf{X}) =: \frac{1}{n-1} \sum_{k=1}^{n-1} \text{CD}_{k \rightarrow (n-k)}(\mathbf{X}). \quad (2.6)$$

This quantity may provide a more principled measure of dynamical complexity than CD in virtue of analysing a target system at multiple scales.¹ As we explain below (see §2*d*), it is closely related to the well-known ‘neural complexity’ measure [5], which averages *mutual information* across bipartitions.²

We offer two final remarks about causal density. First, because MVGC has a spectral decomposition, both CD and BCD can be evaluated within specific frequency bands, which could be useful in cases where such bands have distinct neurophysiological interpretations. Second, in any complex system, it is usually possible to sample only a subset of relevant variables, which can lead to spurious causal inferences arising from hidden common causes. One approach to this problem is to ‘partial out’ hidden influences (by analogy with partial correlation) by introducing an additional term into the Granger equations that is sensitive to correlations among residuals [14,16].

¹A further extension to a full ‘tripartition’ causal density is described in [14].

²This measure has also been referred to as Tononi–Sporns–Edelman complexity.

(c) *Transfer entropy*

A common criticism of G-causality is that its standard implementation in terms of linear MVAR models apparently excludes sensitivity to nonlinear interactions. Nonlinear extensions of G-causality do exist (e.g. [17]), however they are often complex and unwieldy to apply in practice. An alternative framework is provided by *transfer entropy*, which is a measure of directed information transfer based on conditional mutual information [18]. Transfer entropy \mathcal{T} is defined by the difference in entropies

$$\mathcal{T}_{Y \rightarrow X|Z} =: H(\mathbf{X}|\mathbf{X}^- \oplus \mathbf{Z}^-) - H(\mathbf{X}|\mathbf{X}^- \oplus \mathbf{Y}^- \oplus \mathbf{Z}^-), \quad (2.7)$$

and quantifies, in a naturally nonlinear way, the degree to which knowledge of the past of \mathbf{Y} reduces uncertainty in the future of \mathbf{X} , conditional on \mathbf{Z} .

Although it has long been recognized that G-causality and transfer entropy must be related, only recently has an equivalence been formally established [19]. For Gaussian variables, it turns out that a simple factor of 2 relates the two quantities,

$$\mathcal{F}_{Y \rightarrow X|Z} = 2\mathcal{T}_{Y \rightarrow X|Z}. \quad (2.8)$$

The equivalence (2.8) rests on relations between conditional entropy, partial covariance and linear regression prediction error. The essential relations are as follows. Firstly, for Gaussian variables, conditional entropy is a function of the determinant of the corresponding partial covariance matrix [19],

$$H(\mathbf{X}|\mathbf{Y}) = \frac{1}{2} \ln(|\Sigma(\mathbf{X}|\mathbf{Y})|) + \frac{1}{2}n \ln(2\pi e), \quad (2.9)$$

where n is the dimension of \mathbf{X} . This follows from the conditional distribution of \mathbf{X} given any outcome for \mathbf{Y} being Gaussian with covariance matrix $\Sigma(\mathbf{X}|\mathbf{Y})$, and the Gaussian entropy formula

$$H(\mathbf{X}) = \frac{1}{2} \ln(|\Sigma(\mathbf{X})|) + \frac{1}{2}n \ln(2\pi e). \quad (2.10)$$

Second, the partial covariance of \mathbf{X} given \mathbf{Y} is precisely the covariance matrix of the residuals of a linear regression of \mathbf{X} on \mathbf{Y} ,

$$\Sigma(\boldsymbol{\varepsilon}) \equiv \Sigma(\mathbf{X}|\mathbf{Y}). \quad (2.11)$$

Note that the equivalence (2.11) holds for *any* (stationary) \mathbf{X} and \mathbf{Y} , Gaussian or otherwise. Together, these expressions allow \mathcal{T} to be written in terms of linear regression residuals, and therefore to be related directly to \mathcal{F} .

The equivalence between \mathcal{F} and \mathcal{T} is important because it implies that, for Gaussian variables, linear regression accounts for *all* the dependence among variables, further justifying CD as a measure of dynamical complexity (see [20] for a comprehensive review of nonlinear causality measures). Importantly, in the multi-variate case, the equivalence (2.8) holds for the preferred determinant version (MVG) but not for the alternative trace version [14].

(d) Neural complexity and its relation to causal density

The neural complexity \mathcal{C} of a system \mathbf{X} with n elements is given by

$$\mathcal{C}(\mathbf{X}) =: \sum_{k=1}^{n-1} \left(\frac{1}{n_k} \sum_{j=1}^{n_k} H(\mathbf{U}_k^j) - \frac{k}{n} H(\mathbf{X}) \right), \quad (2.12)$$

where \mathbf{U}_k^j is the state of the j th sub-system with k elements and $n_k = \binom{n}{k}$. This measure quantifies the extent to which the entropy of sub-systems is greater than the normalized entropy of the whole, where normalization is by the ratio of the size of the sub-system to the size of the whole. The expected differences for each sub-system size are summed.

A key difference between \mathcal{C} and causal density is that \mathcal{C} is concerned only with the stationary distributions of the states of the system and its parts, whereas causal density is concerned with predicting the present of a system based on its past. However, for Gaussian variables, it can be shown that BCD is equivalent to a modified version of \mathcal{C} and \mathcal{C}' , in which entropies are replaced by the conditional entropies of present states given past states, i.e.

$$\mathcal{C}'(\mathbf{X}) =: \sum_{k=1}^{n-1} \left(\frac{1}{n_k} \sum_{j=1}^{n_k} H[\mathbf{U}_k^j | (\mathbf{U}_k^j)^-] - \frac{k}{n} H[\mathbf{X} | \mathbf{X}^-] \right). \quad (2.13)$$

To see the equivalence, we rearrange \mathcal{C}' in terms of bipartitions of \mathbf{X} . We label the bipartitions of \mathbf{X} such that $\{\mathbf{U}_k^j, \mathbf{V}_k^j\}$ is the j th bipartition with the smaller component \mathbf{U}_k^j consisting of k elements. Then, we have

$$\mathcal{C}'(\mathbf{X}) \equiv \sum_{k=1}^{\lfloor n/2 \rfloor} \frac{1}{n_k} \sum_{j=1}^{n_k} (H[\mathbf{U}_k^j | (\mathbf{U}_k^j)^-] + H[\mathbf{V}_k^j | (\mathbf{V}_k^j)^-] - H[\mathbf{X} | \mathbf{X}^-]). \quad (2.14)$$

We assume that there are no hidden or exogenous elements affecting \mathbf{X} , so that, given the past \mathbf{X}^- , the present state of elements of X are independent, i.e.

$$H(\mathbf{X} | \mathbf{X}^-) = H(\mathbf{U}_k^j | \mathbf{X}^-) + H(\mathbf{V}_k^j | \mathbf{X}^-). \quad (2.15)$$

We can now express \mathcal{C}' in terms of \mathcal{T} , and hence \mathcal{F} ,

$$\mathcal{C}'(\mathbf{X}) = \sum_{k=1}^{n-1} \frac{1}{n_k} \sum_{j=1}^{n_k} \mathcal{T}_{\mathbf{V}_k^j \rightarrow \mathbf{U}_k^j} = \sum_{k=1}^{n-1} \frac{1}{2n_k} \sum_{j=1}^{n_k} \mathcal{F}_{\mathbf{V}_k^j \rightarrow \mathbf{U}_k^j}, \quad (2.16)$$

where now $\{\mathbf{U}_k^j, \mathbf{V}_k^j\}$ is the j th bipartition with \mathbf{U}_k^j consisting of k elements. The direct equivalence between causal density and neural complexity is then given by

$$\text{BCD}(\mathbf{X}) = \frac{1}{2(n-1)} \mathcal{C}'(\mathbf{X}). \quad (2.17)$$

3. Integrated information

An alternative information-theoretic approach to measuring dynamical complexity has been developed by Giulio Tononi, under the banner of the ‘information integration theory of consciousness’ (IITC) [6]. The IITC identifies conscious level with the quantity of *integrated information* (Φ) generated by a system. Several versions of Φ have now been described. The first was conceived as a measure of the *capacity* of a system to integrate information [21], however, it did not take into account time or changing dynamics. A more recent version, Φ_{DM} , was designed to measure the information generated when a system transitions to one particular state out of a repertoire of possible states, to the extent that this information is generated by the whole system, over and above that generated independently by the parts [22]. However, Φ_{DM} is defined only for idealized discrete, Markovian (memoryless) systems (hence the subscript ‘DM’), an in-principle restriction that severely limits in-practice applicability because neural systems are often measured using continuous variables, and also have memory (i.e. dynamics that depend on more than just the previous state). Here, we describe two recent alternative measures, Φ_{E} (‘empirical Φ ’) and Φ_{AR} (‘autoregressive (AR) Φ ’), which overcome the limitations of Φ_{DM} , and which are generally applicable to time-series data [8].

(a) *Integrated information for stationary, continuous systems*

The primary difference between Φ_{DM} and Φ_{E} (and Φ_{AR}) has to do with the probability distributions assumed to characterize the target system. Φ_{DM} measures information with respect to a hypothetical *maximum entropy* distribution, corresponding to the *potential* behaviour of a system (i.e. its capacity). By contrast, Φ_{E} measures information with respect to the *stationary distribution* describing the system’s dynamics, reflecting the *actual* behaviour of a system. Explicitly, Φ_{E} is concerned with the (average) information generated by the current state \mathbf{X}_t of the system about some past state $\mathbf{X}_{t-\tau}$,³

$$I(\mathbf{X}_{t-\tau}; \mathbf{X}_t) \equiv H(\mathbf{X}_{t-\tau}) - H(\mathbf{X}_{t-\tau} | \mathbf{X}_t). \quad (3.1)$$

To measure the extent to which this information is integrated, we use the concept of *effective information* (φ), which refers to the information generated by the whole system, minus the information generated independently by the parts (sub-systems) [21]. Considering only bipartitions $\mathcal{B} = \{M^1, M^2\}$,⁴ the effective information at a time scale τ is given by

$$\varphi[\mathbf{X}; \tau, \mathcal{B}] =: I(\mathbf{X}_{t-\tau}; \mathbf{X}_t) - \sum_{k=1}^2 I(M_{t-\tau}^k; M_t^k). \quad (3.2)$$

Φ_{E} is then defined as the effective information with respect to the *minimum information bipartition* (MIB). The MIB, \mathcal{B}^{MIB} , is the bipartition that minimizes φ after normalization to penalize asymmetric bipartitions.⁵ Intuitively, the MIB

³Since mutual information is symmetric in its two arguments, this can equally well be read as the information generated by the past state about the current state.

⁴The formalism applies straightforwardly to general partitions, see [8].

⁵The normalization factor is determined by the smallest stationary entropy of a subsystem [8].

can be thought of as the ‘informational weakest link’. Thus,

$$\Phi_E[\mathbf{X}; \tau] =: \varphi[\mathbf{X}; \tau, \mathcal{B}^{\text{MIB}}(\tau)]. \quad (3.3)$$

Note that the value of $\Phi_E[\mathbf{X}; \tau]$ is given by the *non-normalized* effective information.

Because we assume stationary statistics, φ and therefore Φ_E can be measured empirically from time-series data without needing a generative model. However, accurately estimating entropies directly from time series can be challenging. Fortunately, for Gaussian systems, Φ_E can be calculated straightforwardly from empirical covariance matrices. Equations (2.9) and (2.10) allow effective information to be written as

$$\varphi[\mathbf{X}; \tau, \mathcal{B}] = \frac{1}{2} \ln \left(\frac{|\Sigma(\mathbf{X})|}{|\Sigma(\mathbf{X}_{t-\tau} | \mathbf{X}_t)|} \right) - \sum_{k=1}^2 \frac{1}{2} \ln \left(\frac{|\Sigma(\mathbf{M}^k)|}{|\Sigma(\mathbf{M}_{t-\tau}^k | \mathbf{M}_t^k)|} \right). \quad (3.4)$$

The partial covariances in equation (3.4) can be obtained by using equation (1.1).

Together, the above formulae permit the straightforward computation of integrated information from time-series data, an important step not possible for previous measures [6,21]. We emphasize that the construction of Φ_E is very similar to that of Φ_{DM} . The important differences are that (i) Φ_E uses the stationary, rather than maximum entropy distribution, (ii) Φ_E uses the average information generated, and so is state-independent, and (iii) Φ_E enables a choice of time scale (τ) over which integrated information is measured. These differences carry substantial implications beyond practical applicability. Most notably, Φ_E is a measure of *process*, whereas Φ_{DM} remains in part a measure of *capacity* or *potential*, in virtue of the maximum entropy distribution. As we discuss later (see §5) Φ_E (and also Φ_{AR} , discussed in §3b) corresponds to a Jamesian view of consciousness-as-process, and thus entails a departure from the IITC that interprets conscious level in terms of capacity.

(b) Autoregressive Φ

As noted, for Gaussian variables, Φ_E can be computed efficiently from empirical covariance matrices. Explicit calculation of Φ_E for (stationary) non-Gaussian variables requires estimation of entropies directly from time series, which is computationally expensive. However, in such cases, we can still use the same formulae to calculate a quantity that remains readily interpretable in terms of integrated information. We have called this quantity Φ_{AR} (for AR Φ). To see why, recall the equivalence between linear regression prediction error and partial covariance (2.11), which allows us to rewrite equation (3.4) as

$$\varphi[\mathbf{X}; \tau, \mathcal{B}] = \frac{1}{2} \ln \left(\frac{|\Sigma(\mathbf{X})|}{|\Sigma(\boldsymbol{\varepsilon}^{\mathbf{X}})|} \right) - \sum_{k=1}^2 \frac{1}{2} \ln \left(\frac{|\Sigma(\mathbf{M}^k)|}{|\Sigma(\boldsymbol{\varepsilon}^{\mathbf{M}^k})|} \right), \quad (3.5)$$

where $\boldsymbol{\varepsilon}^{\mathbf{M}^k}$, $k = 1, 2$, and $\boldsymbol{\varepsilon}^{\mathbf{X}}$ are the residuals in the following regressions:

$$\mathbf{M}_{t-\tau}^k = A^{\mathbf{M}^k} \cdot \mathbf{M}_t^k + \boldsymbol{\varepsilon}_t^{\mathbf{M}^k} \quad (3.6)$$

and

$$\mathbf{X}_{t-\tau} = A^{\mathbf{X}} \cdot \mathbf{X}_t + \boldsymbol{\varepsilon}_t^{\mathbf{X}}. \quad (3.7)$$

Note that, in the above regressions, the *past* of a variable (τ time steps ago) is regressed on its *present* value (contrast with equation (2.1)).⁶ Although the relation between covariance and entropy (2.10) holds only in the Gaussian case, the relation between linear regression prediction error and partial covariance (2.11) holds for any stationary distribution. Thus, we can take equation (3.5) to define a new version of effective information, φ_{AR} , applicable to both Gaussian and non-Gaussian systems. Φ_{AR} is then the non-normalized φ_{AR} across the bipartition that minimizes (normalized) φ_{AR} . For Gaussian systems, Φ_{AR} and Φ_{E} are equivalent, but for non-Gaussian systems, they may differ.

Note that Φ_{AR} can be computed directly from empirical covariance matrices (owing to equation (2.11)); explicit calculation of the regressions is not needed. However, it is the formulation in terms of linear regression that gives Φ_{AR} its meaning in terms of integrated information. Specifically, Φ_{AR} can be understood as a measure of how well the *present* of a system predicts its *past*, to the extent that these predictions improve over predictions based on the parts acting independently. When Gaussian conditions are satisfied, this interpretation becomes exactly equivalent to the interpretation of Φ_{E} in terms of information theory.

(c) *The variants $\tilde{\Phi}_{\text{E}}$ and $\tilde{\Phi}_{\text{AR}}$*

Further measures of integrated information, $\tilde{\Phi}_{\text{E}}$ and $\tilde{\Phi}_{\text{AR}}$, can be defined, respectively, using the following alternative definitions for effective information [8]:

$$\varphi[\mathbf{X}; \tau, \mathcal{B}] =: \sum_{k=1}^2 H(\mathbf{M}_{t-\tau}^k | \mathbf{M}_t^k) - H(\mathbf{X}_{t-\tau} | \mathbf{X}_t) \quad (3.8)$$

and

$$\varphi[\mathbf{X}; \tau, \mathcal{B}] =: \sum_{k=1}^2 \frac{1}{2} \ln \left(|\Sigma(\boldsymbol{\varepsilon}^{\mathbf{M}^k})| \right) - \frac{1}{2} \ln \left(|\Sigma(\boldsymbol{\varepsilon}^{\mathbf{X}})| \right), \quad (3.9)$$

where $\boldsymbol{\varepsilon}^{\mathbf{M}^k}$, $k = 1, 2$, and $\boldsymbol{\varepsilon}^{\mathbf{X}}$ are as in equations (3.6) and (3.7). Rather than being formulated in terms of mutual information, the effective information for $\tilde{\Phi}_{\text{E}}$ (3.8) is the average Kullback–Leibler (KL) divergence between the total distribution for the past state and the product of the distributions of the past states of the parts, all conditioned on respective present states. Note that, for a maximum entropy stationary distribution, $\tilde{\Phi}_{\text{E}}$ and Φ_{E} are equivalent; hence they are equally relatable to Φ_{DM} . $\tilde{\Phi}_{\text{AR}}$ is analogous to Φ_{AR} , and is equivalent to $\tilde{\Phi}_{\text{E}}$ for Gaussian systems. In the examples presented below, all versions of Φ behave similarly.

4. Simulations

We present results from computing CD, BCD, Φ_{E} and $\tilde{\Phi}_{\text{E}}$ (time scale $\tau = 1$), for some example Markovian Gaussian systems. The example systems are

⁶Note that, by symmetry of mutual information, φ could be equivalently expressed in terms of regressions of the present state on the past state.

characterized by the MVAR(1) dynamics

$$\mathbf{X}_t = A \cdot \mathbf{X}_{t-1} + \mathbf{E}_t, \quad (4.1)$$

where \mathbf{X}_t contains n variables, A is the connectivity matrix, and each component of \mathbf{E}_t is an independent Gaussian random variable of mean 0 and variance 1. We considered seven systems with $n = 8$ and connectivity as shown in figure 1*a–g*; we refer to these systems as ‘1(*a*)’, ‘1(*b*)’ and so on. The corresponding values of CD, BCD, Φ_E and $\tilde{\Phi}_E$ are given in figure 1*h–k*. These values were computed analytically, using the generative model (4.1) to obtain the necessary cross- and auto-covariance matrices.⁷

The values of all measures mostly correspond with expectations for these examples. A ring of reciprocal connections (figure 1*c*) generates higher values than a ring of unidirectional connections (figure 1*b*), which itself generates higher values than an open chain of unidirectional connections (figure 1*a*). Also as expected, the homogeneous system figure 1*d* has a low value for all measures (Φ_E and $\tilde{\Phi}_E$ are low compared to other reciprocally connected networks). Perhaps in contrast to expectations, adding sparse long-range ‘short-cut’ connections to a reciprocal ring (figure 1*e–g*), in the style of a so-called ‘small-world’ network [12,23,24], does not increase the value of any of the measures (compare with network figure 1*c*). (We also tested networks with $n = 16$ nodes, finding essentially the same results. Small-world properties are of course more prominent in larger networks; given adequate computational resources, analysis of still larger networks may reveal differences.)

To examine general trends in behaviour of the measures as a function of connection density, we performed the following procedure 100 times (trials). Starting with a fully disconnected network of $n = 8$ elements, in which each element has only a self-connection (strength 0.5), we filled in the (directed) connectivity matrix in a random order. Each time a connection was added, connection strengths were normalized such that (i) there was a constant total afferent (including self-connection) of 0.5 to each element and (ii) all connections to a given element were equal. We then computed CD, BCD, Φ_E and $\tilde{\Phi}_E$, as well as the average stationary instantaneous correlation between elements (the ‘correlation index’), assuming MVAR(1) dynamics (as above). Results can be analysed in multiple ways. First, for each given number of (non-self-) connections k , we plot the inter-trial mean of each measure (solid line, figure 2*a–e*). These plots show that both causal density measures exhibit peaks at intermediate values of k , whereas Φ_E and $\tilde{\Phi}_E$ and the correlation index do not. Second, for each value of k , we plot the inter-trial maximum of each measure (dashed line, figure 2*a–e*). This leads to all measures exhibiting a peak at an intermediate value of k . (Note correlation index shows a peak because, for intermediate k , there is greater inter-trial variety in network architecture.) Finally, we noted, for each trial, the number of connections present when each measure peaked, and compared this with the point at which the network became connected (i.e. the point at which every element can be reached, on some path, from every other element). Figure 2*f* shows the mean of this for each measure, showing that, on a given trial, the expectation is that CD and BCD will peak after the network becomes connected

⁷For the restricted regressions for computing CD and BCD, we used five lags, which was sufficient to approximate the infinite lag limit.

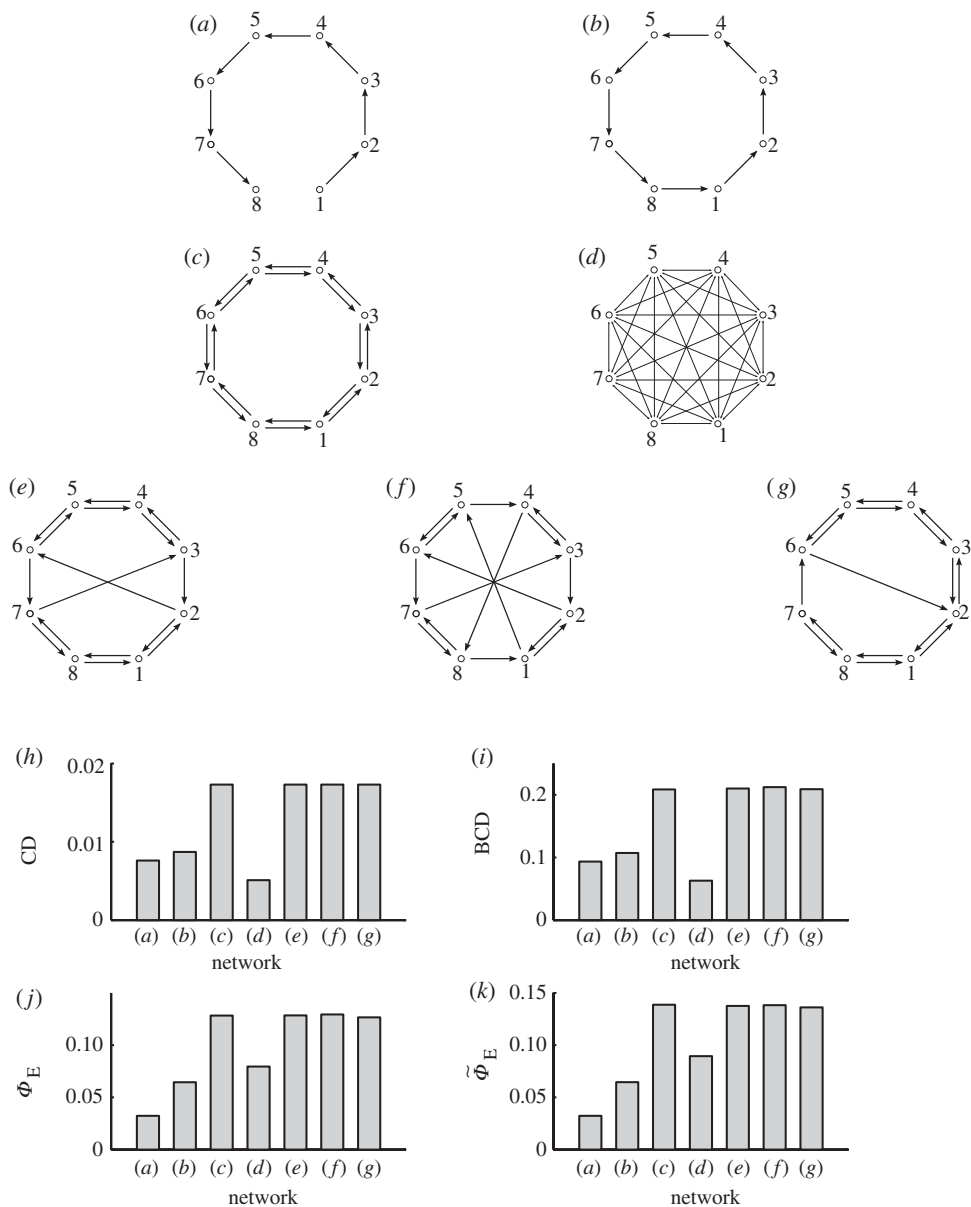


Figure 1. Measures CD, BCD, Φ_E and $\tilde{\Phi}_E$ for Markovian Gaussian systems. (a–g) Connectivity diagrams for seven systems as specified by the corresponding connectivity matrices A . Arrow widths reflect connection strengths: for (a–c) and (e–g), all connection strengths are 0.25; for system (d), each connection strength is $1/14$, thus the total afferent connection to each element is 0.5. (h) CD, (i) BCD, (j) Φ_E ($\tau = 1$), (k) $\tilde{\Phi}_E$ ($\tau = 1$), for each of the systems (a–g), computed analytically.

but before correlation peaks; by contrast, Φ_E and $\tilde{\Phi}_E$ will on average peak at about the same k as the correlation index. It is interesting that, for each measure, there is a difference between the mean peak (figure 2*f*) and the peak of the mean (figure 2*a–e*).

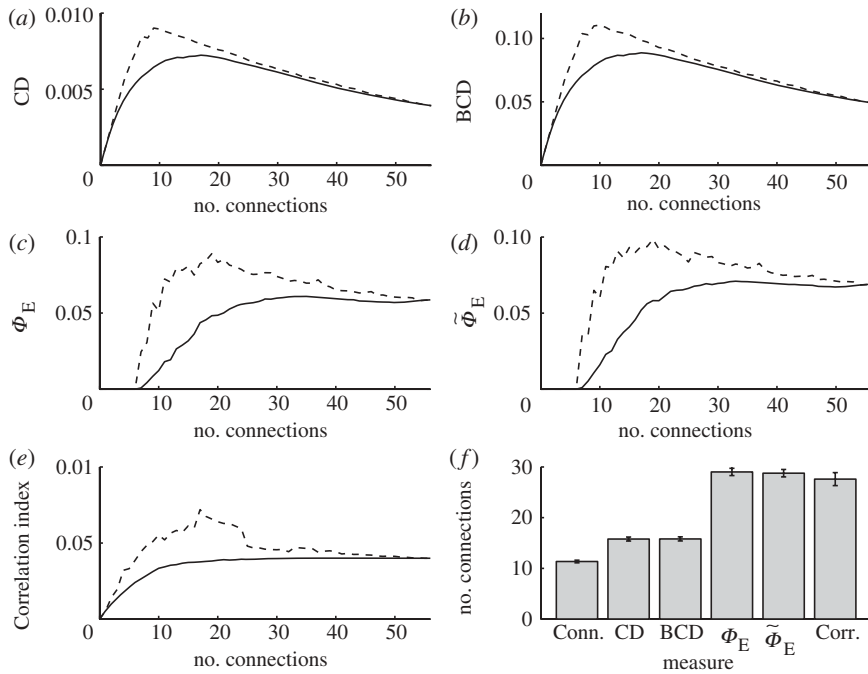


Figure 2. Dynamical complexity measures as a function of connection density, in a network of eight elements. On each trial, the connectivity matrix was filled in a random order, normalizing after each new connection. (a) CD, (b) BCD, (c) Φ_E , (d) $\tilde{\Phi}_E$, (e) correlation index. Solid line shows mean value of each measure over 100 trials; dashed line shows maximum. (f) Mean number of (non-self-) connections at which each measure peaks, taken across trials. ‘Conn.’ shows the mean number of (non-self-) connections present when the network becomes fully connected. ‘Corr.’ is the correlation index. Error bars show standard error.

To assess the generality of these results, we performed an additional set of simulations on networks with $n = 12$ and k randomly chosen connections (no self-connections). Connection strengths were first drawn from independent Gaussians with mean and standard deviation both equal to 0.2, then set as positive (i.e. excitatory) with 80 per cent probability (else negative, i.e. inhibitory), and finally normalized to yield a connectivity matrix with spectral radius $\rho = 0.8$.⁸ We note that for small k (roughly $k < n/2$), the obtained connection strengths frequently yielded zero spectral radius; such networks could not be normalized and were rejected. We calculated CD, Φ_E and $\tilde{\Phi}_E$ using three lags. The number of connected components of the underlying graph and the correlation index for the MVAR(1) were also calculated for each network. For CD, 20 000 trial networks were generated for each k . For Φ_E and $\tilde{\Phi}_E$, there were 1000 trials for each k . Results are shown in figure 3. Figure 3a shows mean CD and correlation index plotted against connection density; figure 3b shows mean Φ_E and $\tilde{\Phi}_E$. The mean number of connected components is plotted on the right-hand axes (graphs

⁸The spectral radius of a (square) matrix is the maximum of the absolute values of its eigenvalues [25]. Intuitively, it reflects the overall level of feedback in the corresponding dynamical process.

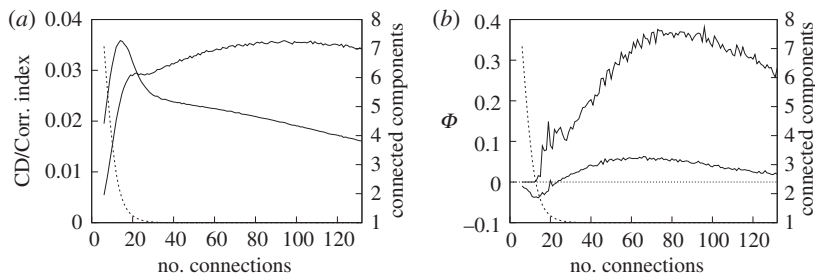


Figure 3. Dynamical complexity measures as a function of connection density in networks of 12 elements, with connection strength normalized by spectral radius (see text for details): (a) mean CD (thick line) and mean correlation index (thin line), (b) mean Φ_E (thick line) and mean $\tilde{\Phi}_E$ (thin line). The mean number of connected components (dotted line) is plotted against the right-hand axes.

become connected as the number of connected components approaches 1). We see that CD peaks (sharply) just prior to emergence of connectivity, whereas Φ_E and $\tilde{\Phi}_E$ peak at higher connectivity levels. We note that, in general, for networks with disconnected components, Φ_E may be negative, while $\tilde{\Phi}_E$ remains at zero.

Taken together, these simulations indicate that both CD and integrated information track non-trivial properties of network topology and dynamics, as suggested by their theoretical formulation. Consistent with intuition, CD is sensitive to a dynamical regime in-between independence of elements and strong correlation among elements (see also [12]). Integrated information, as measured by Φ_E and $\tilde{\Phi}_E$, also shows a dissociation from overall correlation, at least when networks are normalized by spectral density. The two measures also dissociate from each other, with CD (but not integrated information) showing clear peaks in both simulations, and with these peaks appearing at lower connection densities, corresponding to lower levels of overall correlation.

5. Discussion

We have described measures of dynamical complexity based on causal density and integrated information, which are easily applicable to time-series data and which are theoretically grounded in reflecting conjoined functional differentiation and functional integration. This grounding encourages their use as potential measures of conscious level, given neural data contrasting normal waking consciousness with unconscious (or lowered/abnormal) conscious conditions including dreamless sleep, general anaesthesia, coma, the vegetative state, absence epilepsy, somnambulism and perhaps even certain psychiatric disorders.

Measures of causal density are based on predictive ability among time series, operationalized using G-causality, and they attempt to capture the overall amount of causal interactivity exhibited by a system. We have described causal density measures, CD and BCD, which characterize, respectively, (i) overall causal interactivity between pairs of elements, conditioned on the rest of the system and (ii) overall causal interactivity between bipartitions of a system. Measures of integrated information reflect the extent to which the information generated by

the whole system exceeds that generated by its parts. Our measures Φ_E and Φ_{AR} operationalize this concept for Gaussian and non-Gaussian systems, respectively, and unlike the previous measure Φ_{DM} , they are easy to apply to empirical time-series data. Our key innovations over Φ_{DM} are (i) to measure information with respect to the stationary (as opposed to maximum entropy) distribution (Φ_E) and (ii) to interpret integrated information in terms of the predictive ability of the present with respect to the past (Φ_{AR}). The variants $\tilde{\Phi}_E$ and $\tilde{\Phi}_{AR}$ are equally valid as alternatives to Φ_{DM} ; they differ from Φ_E and Φ_{AR} only by characterizing effective information in terms of average KL divergence rather than by mutual information. We emphasize that the simulation results we have presented pertain to Φ_E and $\tilde{\Phi}_E$, not to Φ_{DM} , though we have previously shown that the measures do behave qualitatively identically in simple examples where calculation of Φ_{DM} is tractable [8].

(a) *Comparison among measures*

Causal density (CD, BCD) integrated information (Φ_{DM} , Φ_E , Φ_{AR}) and neural complexity \mathcal{C} embody theoretical differences that have empirical consequences as revealed in our simulations. At the theoretical level, measures of integrated information differ from measures of causal density most prominently in that the former are based on a value (of effective information, φ) across a single partition (the MIB), whereas the latter are based on values (of G-causality, \mathcal{F}) averaged across the whole system. The former approach guarantees that a disconnected system will not score positive, but raises problems of stability in systems that have several partitions with similar normalized φ but different non-normalized φ . In previous work, we have discovered examples of such networks for which Φ_E changes abruptly as a single connection strength is varied continuously [8]. On the other hand, a partition-based measure ensures that the measured value does not reflect only generic properties of a system, such as overall connection strength. Indeed, we have shown that simple eight-element networks optimized for Φ_E exhibit highly heterogeneous topologies [8]. This difference aside, the measures have several theoretical connections. Leveraging an equivalence between G-causality and transfer entropy, we have shown here that BCD is equivalent (for Gaussian systems) to a version of \mathcal{C} based on (time-directed) conditional entropies. Hence, causal density can be thought of as a ‘transfer entropy density’. Φ_{AR} bears further relation to causal density since both measures are based on AR models, the former employing a model order of 1, with varying time lag, and the latter employing, in addition, a variable model order.

Our simulation results support the intuitions underlying causal density and integrated information, while emphasizing that the measures can behave differently. We examined two sets of simulations, both of which involved ensembles of networks parameterized by overall connection density, with connections arranged at random. The simulations differed primarily by using different normalization factors; the first set maintained a constant afference to each element, while the second set normalized by spectral radius (roughly, the overall level of feedback in the corresponding AR process). The first set found that, as connection density increased, CD and BCD peaked after the network became connected, but before the overall correlation among network elements reached its maximum. By contrast, Φ_E and $\tilde{\Phi}_E$ either peaked later, more-or-less with

the onset of maximum correlation (when measured by the maximum across each ensemble) or did not peak (when measured by the mean). Generalizing these observations, the second set showed that causal density again peaked prior to maximum correlation, though in this case, more sharply and also prior to the emergence of a single connected component. Information integration, by contrast, peaked at higher connection densities and exhibited a noisier profile, likely due to the instabilities incurred by a partition-based method. As well as exemplifying the intuitive behaviour of our measures, these simulations highlight the importance of selecting an appropriate normalization method to ensure that measures do not reflect only trivial network properties. Since there is no *a priori* ‘best’ normalization for a network considered an abstraction of a neural system (e.g. [26]), we decided to compare two plausible alternatives. Further work is needed to establish the general relations among normalization, network structure and dynamical measures.

Other groups have described related measures without explicitly relating them to consciousness. For example, Ay and colleagues compare the whole system to the sum of individual elements [27] in terms of information geometry and conditional temporal information, furnishing a link to \mathcal{C} . They show an interesting link between these ideas and timing dependent plasticity, raising the interesting possibility that neural plasticity may be an essential prerequisite for consciousness in virtue of shaping dynamical complexity properties [28].

(b) *Practical application*

Despite the theoretical advances described in this paper, application of causal density and integrated information to neural data remains challenging for several reasons. For numerical ease (i.e. for calculation directly from empirical covariance matrices), the data must be covariance (wide-sense) stationary; i.e. means and variances must remain constant over time. However, neural data are often noisy and contaminated by artefacts and drift, introducing non-stationarities. Non-stationary data can sometimes be made stationary by *differencing* (i.e. $X'_t = X_t - X_{t-1}$), repeatedly if necessary (differencing can however change the interpretation of subsequent calculations). Alternatively, one can analyse shorter time windows, each of which may be locally stationary [29]. For some types of data (notably M/EEG), bandpass and/or notch filtering are commonly used in order to remove artefacts and drift. Although AR modelling is in principle invariant to filtering [30], the temporal correlations induced by convolution with a filter introduce challenges for accurate and stable estimation of AR models [8].

Neural signals are frequently nonlinear as well as non-stationary. While information-theoretic measures are naturally nonlinear, AR models are typically estimated by linear models. However, as we have shown (§2c; see [14] for an extended discussion), a Gaussian equilibrium state distribution implies a linear AR model; moreover, information-theoretic quantities typically depend on Gaussian distributions in order to be calculable in practice. Together, these points underline the importance of a Gaussian assumption. Fortunately, Gaussian approaches appear to be very powerful in neuroscience. For example, Palùs and colleagues have shown (in theory and when applied to preictal EEG) that a Gaussian approach can distinguish different states of nonlinear chaotic attractors in ways similar to Lyapunov exponents or nonlinear entropy rates

[31,32]; they also show that Gaussian descriptions are sufficient to describe resting state activity as measured by fMRI [33]. These results build on earlier work indicating the utility of linear approximations, especially for modelling large-scale interaction in neuroscience [34].

Neural data relevant to conscious level are typically obtained from neuroimaging methods such as M/EEG and fMRI. Each method poses its own distinctive challenges. M/EEG signals offer high temporal resolution suitable for time-series analysis, however, they require non-unique inverse modelling to move from sensor space (at the scalp) to an underlying source space, conferring ambiguity on subsequent interpretations. The spatial resolution of EEG is also low as compared to fMRI, and additional challenges related to filtering have already been mentioned. fMRI contrasts with M/EEG by offering high spatial resolution at the expense of temporal resolution. The blood oxygen level dependent (BOLD) signal measured by fMRI reflects slow metabolic processes related to neural activity (these relations remain incompletely understood [35]), and is typically sampled of the order of 0.3–1 Hz. The correspondingly sparse time series pose challenges for accurate estimation of AR models and even of covariance matrices. Moreover, variability in haemodynamic responses in different brain regions and different subjects [36] may confound causal inference [37], undermining subsequently derived complexity measures. However, causal inference based on fMRI time series is an area of extremely active research [38], and promising new approaches are emerging, for example, embedding AR models into state–space models that include haemodynamic parameters [39] and integrating perturbational approaches using transcranial magnetic stimulation [40].

Our simulations have focused on small networks ($n = 8, 12$). One reason for this is computational feasibility. For Φ , there is an exponential growth in the number of partitions to consider as n increases (causal density is less demanding, scaling roughly as n^2). However, raw computer power is continuing to increase, raising the possibility of calculation of these measures for much larger systems. A second reason is statistical. When computing causal density or Φ from data, the number of parameters to estimate scales roughly as n^2 . This is a problem, not for computational reasons, but because estimating a large number of parameters requires a correspondingly large amount of data (i.e. longer time series) in order to avoid overfitting, and to obtain reasonable constraints on the estimated parameters. However, a system will only remain statistically stationary for a limited time period, restricting the number of nodes that can be considered in practice. One approach to this problem is to impose priors on the estimated parameters, for example, via regularization or sparsification [41].

Even for small n , causal density and Φ may be informative about consciousness. Abundant evidence indicates that coarse-grained low-dimensional information can differentiate states and contents of consciousness. For example, when analysing EEG signals from subjects experiencing binocular rivalry, synchrony among a few well-separated electrodes indicates perceptual dominance [2,42]. Sleep, anaesthesia and other states involving global loss of consciousness also have neural signatures detectable at coarse-grained levels (see [43] for a review). We emphasize that the measures described in this paper are motivated because they propose to *explain* rather than merely *correlate* with consciousness [1,44], a relationship that may hold at multiple levels of coarse graining. Although (to our

knowledge) neither causal density nor Φ have yet been convincingly applied to experimental data, there have been several attempts to apply neural complexity \mathcal{C} to EEG data, with mixed results [45–47]. These studies have also used relatively small n (e.g. $n = 18$ in [45]); their divergence underlines the need for careful attention to details of practical application and may reflect such differences and/or a limitation of \mathcal{C} in depending exclusively on zero-lag covariance.

(c) *Consciousness and complexity*

William James famously described consciousness as a ‘stream’ or ‘process’ [48]. The measures of dynamical complexity described in this paper—causal density, Φ_E and Φ_{AR} —are consistent with this view. Because these measures depend on the stationary statistics of the underlying neural dynamics, when considered as potential measures of conscious level they imply (i) the conscious level is constant during each stationary epoch in brain activity and (ii) the conscious level changes when functional connectivity changes, modifying the stationary statistics. Despite sharing a common ground in emphasizing dynamical complexity, this perspective is different from the one advocated by the IITC. According to the IITC, consciousness is best characterized as a ‘potential’ or ‘capacity’, as reflected by Φ_{DM} in virtue of its reliance on the maximum entropy distribution [6]. A challenging implication of this property is that conscious experience is modulated by states that the brain never in fact encounters, if these states are part of the maximum entropy but not the stationary distribution. Another feature of the IITC is that integrated information is *identified* with consciousness, implying a relation of sufficiency. In our view, dynamical complexity (information integration) may be necessary, but is unlikely to be sufficient for generating consciousness. A challenge to the IITC in this context is that all measures of integrated information so far described exhibit instabilities due to normalization (see above, and [8]), undermining the ascription of physical meaning to the quantity. Furthermore, Φ and causal density are not invariant under changes of coordinates or state–space representations. On the one hand, this further emphasizes that the measures should not be identified with consciousness; on the other, this raises the interesting possibility that there may be a particular state–space description that maximizes causal density or Φ [8], and which may have neurobiological relevance in doing so.

The differences among the various measures described in this paper underscore the importance of operationalizing intuitive properties of a target phenomenon (in this case, consciousness). As we have seen, a similar intuition (conjoined functional integration and functional differentiation) can be operationalized in significantly different ways, carrying major implications not just for practical applicability but also for underlying theoretical principles. Future challenges lie in continuing this programme of operationalization for other fundamental properties of consciousness, such as the existence of a first-person perspective, the shaping of conscious contents by mood, volition and agency, and the sense of subjective reality (presence) [44]. At the same time, additional theory and modelling is needed to fully disclose the relations among integrated information, causal density and neural complexity, and to understand how these quantities are affected by embedding a neural system within a sensorimotor loop involved in generating behaviour. But the most pressing concern is to close the gap between theory

and practice, to examine whether theoretically grounded measures, such as those described here, offer additional empirical purchase in virtue of, and not in spite of, their theoretical properties.

A.K.S. and A.B.B. are supported by EPSRC Leadership Fellowship EP/G007543/1 to A.K.S. A.K.S. and L.B. are supported by the Dr Mortimer and Theresa Sackler Foundation. We are grateful to four anonymous reviewers for their useful comments, which helped improve the manuscript.

References

- 1 Seth, A. K. 2010 The grand challenge of consciousness. *Front. Psychol.* **1**, 1–2. (doi:10.3389/fpsyg.2010.00005)
- 2 Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W. & Rodriguez, E. 2007 Synchronization of neural activity across cortical areas correlates with conscious perception. *J. Neurosci.* **27**, 2858–2865. (doi:10.1523/JNEUROSCI.4623-06.2007)
- 3 Seth, A. K., McKeinsty, J. L., Edelman, G. M. & Krichmar, J. L. 2004 Visual binding through reentrant connectivity and dynamic synchronization in a brain-based device. *Cereb. Cortex* **14**, 1185–1199. (doi:10.1093/cercor/bhh079)
- 4 Balduzzi, D. & Tononi, G. 2009 Qualia: the geometry of integrated information. *PLoS Comput. Biol.* **5**, e1000462. (doi:10.1371/journal.pcbi.1000462)
- 5 Tononi, G., Sporns, O. & Edelman, G. M. 1994 A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl Acad. Sci. USA* **91**, 5033–5037. (doi:10.1073/pnas.91.11.5033)
- 6 Tononi, G. 2008 Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* **215**, 216–242. (doi:10.2307/25470707)
- 7 Seth, A. K., Izhikevich, E., Reeke, G. N. & Edelman, G. M. 2006 Theories and measures of consciousness: an extended framework. *Proc. Natl Acad. Sci. USA* **103**, 10 799–10 804. (doi:10.1073/pnas.0604347103)
- 8 Barrett, A. B. & Seth, A. K. 2011 Practical measures of integrated information for time series data. *PLoS Comput. Biol.*, **7**, e1001052. (doi:10.1371/journal.Pcbi.1001052)
- 9 Geweke, J. 1982 Measurement of linear dependence and feedback between multiple time series. *J. Am. Statist. Assoc.* **77**, 304–313. (doi:10.2307/2287238)
- 10 Granger, C. W. J. 1969 Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438. (doi:10.2307/1912791)
- 11 Seth, A. K. 2010 A MATLAB toolbox for Granger causal connectivity analysis. *J. Neurosci. Meth.* **186**, 262–273. (doi:10.1016/j.jneumeth.2009.11.020)
- 12 Shanahan, M. 2008 Dynamical complexity in small-world networks of spiking neurons. *Phys. Rev. E* **78**, 041924. (doi:10.1103/PhysRevE.78.041924)
- 13 Harris, K. D. 2005 Neural signatures of cell assembly organization. *Nat. Rev. Neurosci.* **6**, 399–407. (doi:10.1038/nrn1669)
- 14 Barrett, A. B., Barnett, L. & Seth, A. K. 2010 Multivariate granger causality and generalized variance. *Phys. Rev. E* **81**, 041907. (doi:10.1103/PhysRevE.81.041907)
- 15 Ladroue, C., Guo, S., Kendrick, K. & Feng, J. 2009 Beyond element-wise interactions: identifying complex interactions in biological processes. *PLoS ONE* **4**, e6899. (doi:10.1371/journal.pone.0006899)
- 16 Guo, S., Seth, A. K., Kendrick, K. M., Zhou, C. & Feng, J. 2008 Partial Granger causality—eliminating exogenous inputs and latent variables. *J. Neurosci. Meth.* **172**, 79–93. (doi:10.1016/j.jneumeth.2008.04.011)
- 17 Marinazzo, D., Liao, W., Chen, H. & Stramaglia, S. In press, Nonlinear connectivity by Granger causality. *Neuroimage*. (doi:10.1016/j.neuroimage.2010.01.099)
- 18 Schreiber, T. 2000 Measuring information transfer. *Phys. Rev. Lett.* **85**, 461–464. (doi:10.1103/PhysRevLett.85.461)

- 19 Barnett, L., Barrett, A. B. & Seth, A. K. 2009 Granger causality and transfer entropy are equivalent for gaussian variables. *Phys. Rev. Lett.* **103**, 238701. (doi:10.1103/PhysRevLett.103.238701)
- 20 Hvlaváčková-Schindler, K., Palūs, M., Vejmelka, M. & Bhattacharya, J. 2007 Causality detection based on information-theoretic approaches in time series analysis. *Phys. Rep.* **441**, 1–46. (doi:10.1016/j.physrep.2006.12.004)
- 21 Tononi, G. & Sporns, O. 2003 Measuring information integration. *BMC Neurosci.* **4**, 31.
- 22 Balduzzi, D. & Tononi, G. 2008 Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput. Biol.* **4**, e1000091. (doi:10.1371/journal.pcbi.1000091)
- 23 Watts, D. J. & Strogatz, S. H. 1998 Collective dynamics of ‘small world’ networks. *Nature* **393**, 440–442. (doi:10.1038/30918)
- 24 Watts, D. J. 1999 *Small worlds*. Princeton, NJ: Princeton University Press.
- 25 Horn, R. A. & Johnson, C. R. 1985 *Matrix analysis*. New York, NY: Cambridge University Press.
- 26 Barnett, L., Buckley, C. L. & Bullock, S. 2009 Neural complexity and structural connectivity. *Phys. Rev. E* **79**, 051914. (doi:10.1103/PhysRevE.79.051914)
- 27 Wennekers, T. & Ay, N. 2001 Dynamical properties of strongly interacting markov chains. *Neural Netw.* **16**, 1483–1497.
- 28 Wennekers, T. & Ay, N. 2005 Finite state automata resulting from temporal information maximization and a temporal learning rule. *Neural Comput.* **17**, 2258–2290. (doi:10.1162/0899766054615671)
- 29 Ding, M., Bressler, S., Yang, W. & Liang, H. 2000 Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment. *Biol. Cybern.* **83**, 35–45. (doi:10.1007/s004229900137)
- 30 Geweke, J. 1984 Measures of conditional linear dependence and feedback between time series. *J. Am. Statist. Assoc.* **79**, 907–915. (doi:10.2307/2288723)
- 31 Palūs, M. 1997 On entropy rates of dynamical systems and Gaussian processes. *Physica A* **227**, 301–308. (doi:10.1016/S0375-9601(97)00079-0)
- 32 Palūs, M., Komarek, V., Hrnčir, Z. & Prochazka, T. 1999 Is nonlinearity relevant for detecting changes in EEG? *Theory Biosci.* **118**, 179–188.
- 33 Hlinka, J., Palūs, M., Vejmelka, M., Martini, D. & Corbetta, M. 2011 Functional connectivity in resting-state fMRI: is linear correlation sufficient? *Neuroimage* **54**, 2218–2225. (doi:10.1016/j.neuroimage.2010.08.042)
- 34 McIntosh, A. R. & Gonzalez-Lima, F. 1994 Structural equation modeling and its application to network analysis in functional brain imaging. *Hum. Brain Map.* **2**, 2–22. (doi:10.1002/hbm.460020104)
- 35 Ekstrom, A. 2010 How and when the fMRI BOLD signal relates to underlying neural activity: the danger in dissociation. *Brain Res. Rev.* **62**, 233–244. (doi:10.1016/j.brainresrev.2009.12.004)
- 36 Aguirre, G. K., Zarahn, E. & D’Esposito, M. 1998 The variability of human, BOLD hemodynamic responses. *Neuroimage* **8**, 360–369. (doi:10.1006/nimg.1998.0369)
- 37 David, O., Guillemain, I., Saittel, S., Reyt, S., Deransart, C., Segebarth, C. & Depaulis, A. 2008 Identifying neural drivers with functional MRI: an electrophysiological validation. *PLoS Biol.* **6**, 2683–2697. (doi:10.1371/journal.pbio.0060315)
- 38 Roebroek, A., Formisano, E. & Goebel, R. In press. The identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *Neuroimage*. (doi:10.1016/j.neuroimage.2009.09.036)
- 39 Ryalí, S., Supekar, K., Chen, T. & Menon, V. 2011 Multivariate dynamical systems models for estimating causal interactions in fMRI. *Neuroimage* **54**, 807–823. (doi:10.1016/j.neuroimage.2010.09.052)
- 40 Massimini, M., Boly, M., Casali, A., Rosanova, M. & Tononi, G. 2009 A perturbational approach for evaluating the brain’s capacity for consciousness. *Prog. Brain Res.* **177**, 201–214. (doi:10.1016/S0079-6123(09)17714-2)

- 41 Martínez-Montes, E., Vega-Hernández, M., Sánchez-Bornot, J. M. & Valdés-Sosa, P. A. 2008 Identifying complex brain networks using penalized regression methods. *J. Biol. Phys.* **34**, 315–323. (doi:10.1007/s10867-008-9077-0)
- 42 Srinivasan, R., Russell, D. P., Edelman, G. M. & Tononi, G. 1999 Increased synchronization of magnetic responses during conscious perception. *J. Neurosci.* **19**, 5435–5448.
- 43 Tononi, G. & Koch, C. 2008 The neural correlates of consciousness: an update. *Ann. NY Acad. Sci.* **1124**, 239–261. (doi:10.1196/annals.1440.004)
- 44 Seth, A. K. 2009 Explanatory correlates of consciousness: theoretical and computational challenges. *Cogn. Comput.* **1**, 50–63. (doi:10.1007/s12559-009-9007-x)
- 45 Burgess, A. P., Rehman, J. & Williams, J. D. 2003 Changes in neural complexity during the perception of 3D images using random dot stereograms. *Int. J. Psychophysiol.* **48**, 35–42. (doi:10.1016/S0167-8760(03)00002-3)
- 46 Branston, N. M., El-Dereby, W. & McGlone, F. P. 2005 Changes in neural complexity of the EEG during a visual oddball task. *Clin. Neurophysiol.* **116**, 151–159. (doi:10.1016/j.clinph.2004.07.015)
- 47 van Putten, M. J. A. M. & Stam, C. J. 2001 Application of a neural complexity measure to multichannel eeg. *Phys. Lett. A* **281**, 131–141. (doi:10.1016/S0375-9601(01)00121-9)
- 48 James, W. 1904 Does consciousness exist? *J. Phil. Psychol. Sci. Meth.* **1**, 477–491. (doi:10.2307/2011942)