A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula

Robin A. A. Ince[1*], Bruno L. Giordano[1], Christoph Kayser[1], Guillaume A. Rousselet[1], Joachim Gross[1] and Philippe G. Schyns[1]

[1] Institute of Neuroscience and Psychology, University of Glasgow, 58 Hillhead Street, Glasgow, G12 8QB, UK

[*] Corresponding author:
robin.ince@glasgow.ac.uk
+44 7939 203 596

Abstract

We begin by reviewing the statistical framework of information theory as applicable to neuroimaging data analysis. A major factor hindering wider adoption of this framework in neuroimaging is the difficulty of estimating information theoretic quantities in practice. We present a novel estimation technique that combines the statistical theory of copulas with the closed form solution for the entropy of Gaussian variables. This results in a general, computationally efficient, flexible, and robust multivariate statistical framework that provides effect sizes on a common meaningful scale, allows for unified treatment of discrete, continuous, uni- and multi-dimensional variables, and enables direct comparisons of representations from behavioral and brain responses across any recording modality. We validate the use of this estimate as a statistical test within a neuroimaging context, considering both discrete stimulus classes and continuous stimulus features. We also present examples of novel analyses facilitated by these developments, including application of multivariate analyses to MEG planar magnetic field gradients, and pairwise temporal interactions in evoked EEG responses. We show the benefit of considering the instantaneous temporal derivative together with the raw values of M/EEG signals as a multivariate response, how we can separately quantify modulations of amplitude and direction for vector quantities, and how we can measure the emergence of novel information over time in evoked responses. Open-source Matlab and Python code implementing the new methods accompanies this article.

Abstract 1
Keywords 3
Highlights 3

1. Introduction 4

2. Review of information theory for neuroimaging 5
2.1. Entropy 5
2.2. Mutual information 7
2.3. Existing methods for calculating mutual information 10
Binned methods 10
Continuous methods 12
2.4. Estimating the entropy and MI of Gaussian variables 12
2.5. Estimating MI within different types of experimental design 13
Event related design 14
Continuous design 15
Hybrid design 15
2.6. Higher order information theoretic quantities 16
Conditional Mutual Information 16
Directed Information (Transfer entropy) 18
Interaction information 18
Directed Feature Information 21
2.7. Relation between information theoretic quantities and other statistical approaches 21

3. A novel method for MI estimation using a Gaussian copula 23
3.1. Estimating MI between two continuous variables with a Gaussian copula 23
3.2. Estimating MI between a discrete and a continuous variable using a copula based approach 26
3.3. Estimating MI in spectral data: phase and power 28

4. Results 30
4.1. Discrete experimental condition with continuous EEG response: Face detection 31
4.2. Continuously varying experimental condition with continuous MEG response: listening to speech 34
4.3. Pairwise temporal interactions reveal modulation of gradient in EEG 37

5. Discussion 41
Application to multi-dimensional spaces 41
Second level analyses on MI values 42
Quantifying pairwise interactions 42
Broader applications 43
Conclusion 44

Acknowledgements 44

References 45

Keywords

Multivariate statistics, Mutual information, conditional mutual information, interaction information, redundancy, synergy

Highlights

- Novel estimator for mutual information and other information theoretic quantities
- Provides general, efficient, flexible and robust multivariate statistical framework
- Validated statistical performance on EEG and MEG data
- Applications to spectral power and phase, 2D magnetic field gradients, temporal derivatives
- Interaction information relates information content in different responses

1. Introduction

Mutual Information (MI) measures the statistical dependence between two random variables (Cover and Thomas, 1991; Shannon, 1948). It can be viewed as a statistical test against a null hypothesis that two variables are statistically independent, but in addition its effect size (measured in bits) has a number of useful properties and interpretations, that we review here.

There is a long history of applications of MI for the study of neural activity (Borst and Theunissen, 1999; Eckhorn and Pöpel, 1974; Fairhall et al., 2012; Nelken and Chechik, 2007; Rolls and Treves, 2011; Victor, 2006). Its consistent effect size allows comparison of different neural response codes (Ince et al., 2013; Kayser et al., 2009; Reich et al., 2001), characterization of different neurons (Sharpee, 2014) as well as quantification of the effect of correlations between neurons (Ince et al., 2010, 2009; Moreno-Bote et al., 2014) and of the importance of spike timing (Kayser et al., 2010; Nemenman et al., 2008; Panzeri et al., 2001). Recent studies have begun to explore its application to neuroimaging (Caballero-Gaudes et al., 2013; Gross et al., 2013; Guggenmos et al., 2015; Ostwald and Bagshaw, 2011; Panzeri et al., 2008; Salvador et al., 2007; Saproo and Serences, 2010; Schyns et al., 2011; Serences et al., 2009).

Despite its useful properties, a possible reason for why MI has not been more widely adopted, particularly within the neuroimaging community, is the difficulty of accurately estimating MI from limited quantities of experimental data (Steuer et al., 2002). The most common approach involves quantizing the data into a number of bins, and estimating MI over the resulting discrete spaces. However, this method is sensitive to the problem of limited sampling bias, which is particularly acute when considering multi-dimensional responses. Several continuous methods are available (outlined briefly in Materials and Methods), but while these measures are often less sensitive to sampling bias effects, they can be computationally expensive and often require the estimation (or ad-hoc setting) of additional parameters.

Here we present a novel approach to estimating MI with continuous variables (Ince et al., 2015; Kayser et al., 2015). Our method is rank-based, robust and makes no assumptions on the marginal distributions of each variable. It does make an assumption on the form of the relationship between the variables, which results in the estimate being a lower bound to the true MI. It is computationally efficient and statistically powerful, particularly when combined with permutation testing. We highlight the benefits resulting from the ability of the estimator to extend to multivariate response spaces, which are often intractable with other methods. This improved multivariate performance allows accurate estimation of quantities such as Conditional Mutual Information (CMI) (Ince et al., 2012), Directed Information (also called transfer entropy) (Ince et al., 2015; Massey, 1990; Schreiber, 2000) as well as measures quantifying pairwise interactions between variables (Chicharro, 2014; Panzeri et al., 2008). We believe these higher-order information-theoretic quantities have the potential to provide transformative new interpretations of neuroimaging data, by providing a unified framework for

analyses based on the information content of neural signals (Kriegeskorte and Bandettini, 2007; Naselaris et al., 2011; Schyns et al., 2009). The methods we present enable study of the representation, processing, and communication in the brain of multiple features of the external world (Ince et al., 2015). Furthermore, they also enable the study of the relationships between representation in different signals (Kriegeskorte et al., 2008). Overall, this new estimator provides the basis for a useful and flexible multivariate statistical framework for the analysis of neuroimaging data.

In this paper, we first introduce the concepts of entropy and MI within a neuroimaging context, and briefly review current MI estimation methods. We also describe higher-order information theoretic quantities and possible neuroimaging applications (Section 2). We then present our novel MI estimator (Section 3) and demonstrate its performance with example analyses and statistical simulations on several data sets (Section 4).

2. Review of information theory for neuroimaging

In this section we review information theoretic methods from a neuroimaging perspective. Readers familiar with information theory can skip this section and proceed to Section 3 where we present our novel MI estimator.

2.1. Entropy

Entropy is the foundational quantity of information theory, and is a measure of the uncertainty, or variability, of a random variable. For any particular value of a random variable, a low probability means that outcome is less likely to occur, and so an observer would be more surprised to see it. A high probability value would be less surprising. This notion can be formulated mathematically[1], and entropy is then defined as the expected (average) surprise over the distribution (Figure 1). If an observer draws samples from a distribution, a lower entropy distribution means the observer will be less surprised (or uncertain) about the outcome of any particular sample – i.e. they would be able to make a more accurate guess. Spread out distributions (with high variability) will have high entropy since all potential outcomes have similar probabilities and so the outcome of any particular draw is very uncertain. On the other hand, concentrated distributions (with low variability) will have lower entropy, since some outcomes will have high probability, allowing a reasonable guess to be made about the outcome of any particular draw. Figure 1 shows some examples of entropy of some continuous and discrete variables.

[1] The surprise for value x drawn from a distribution $P(x)$ is defined as $-\log_2 P(x)$

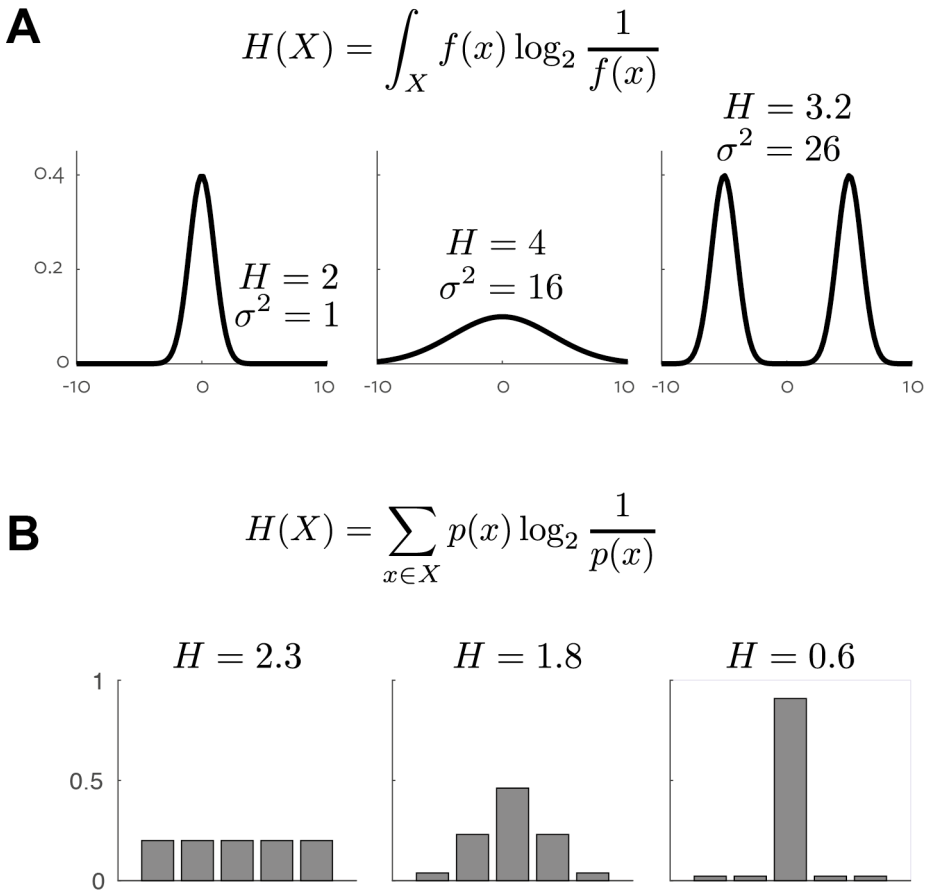

Figure 1 *Entropy values for some example distributions.* **A.** Variance (σ^2) and entropy (H) for three example continuous 1D distributions. **B.** Entropy (H) for three example discrete distributions.

When the logarithm used in the definition of surprise is base-2 (Figure 1), the resulting entropy value has units of *bits*. In this case the entropy value has a useful interpretation. If an ideal observer with knowledge of the true distribution has to guess the value of a particular sample by asking a series of yes/no questions, the entropy in bits gives the average number of questions required. Equivalently, a reduction of entropy by 1 bit corresponds to halving of the uncertainty.

If the random variable considered is discrete, the distribution with maximal possible entropy is the uniform distribution (Figure 1B). If the variable is continuous, entropy[2] can be thought of as a generalized form of variance although unlike variance, which implicitly assumes uni-modality, entropy can give meaningful values for any form of distribution. This analogy with variance can be useful to keep in mind when considering other information theoretic quantities (Garner and McGill, 1956). In the case of continuous variables, for a specified mean and variance the distribution with maximal entropy is a Gaussian. Further, for Gaussian distributions the entropy is proportional to the logarithm of the variance.

[2] For continuous variables, the term *differential entropy*, is often used – but here for conciseness we use *entropy* for both types of variable.

Entropy alone can form a useful measure of the complexity of a signal (Abásolo et al., 2006; Inouye et al., 1991; Overath et al., 2007). However here we are interested primarily its relation to mutual information, which quantifies the relationship between two variables (for example an external stimulus and a recorded signal) in terms of differences in entropies.

2.2. Mutual information

Mutual information (MI) is a measure of the statistical dependence between two random variables (Cover and Thomas, 1991; Latham and Roudi, 2009). It is the most general such measure because MI makes no assumptions on the distribution of the variables, or the nature of the relationship between them and is sensitive to non-linear and non-monotonic effects (Figure 2).

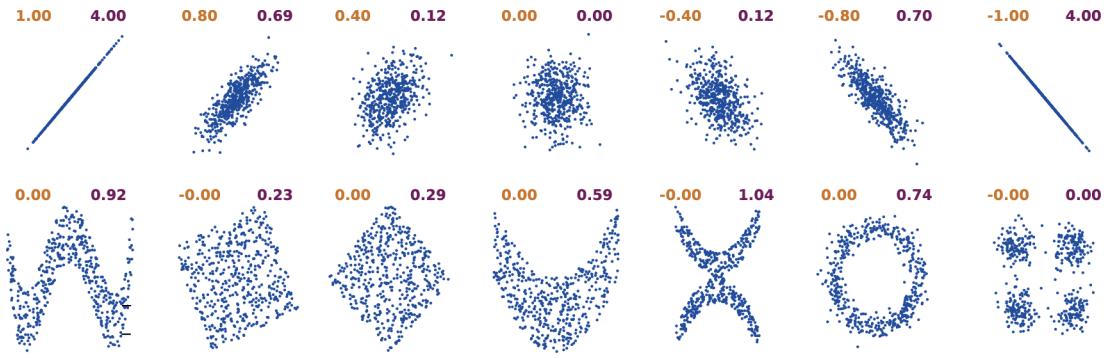

Figure 2 *Examples of correlation vs mutual information.* Each panel illustrates a scatter plot of samples drawn from a particular bi-variate distribution. For each example, the correlation between the two variables is shown in orange (left) and the MI is shown in purple (right; discrete method, 16 bins, 100,000 samples, no bias correction). The top row shows linear relationships, for which MI and correlation both detect a relationship (although on different scales, and note that as MI is always positive it does not reveal the direction of the relationship). The bottom row shows a series of distributions for which the correlation is zero.

MI is defined in terms of entropy differences. As a motivating example, consider a roll of a fair 6-sided die. The outcome of any particular roll follows a uniform distribution over the 6 possible values, with entropy $\log_2 6$. If an observer is told that the result of a particular roll is an even number, there are now 3 possible values, all equally likely, and the entropy of this distribution is $\log_2 3$. The difference between these entropies[3] is 1 bit; this quantifies the amount of information conveyed by the knowledge "this roll is even", and corresponds to a halving of the uncertainty about the outcome (from 6 possibilities to 3).

There are three mathematically equivalent ways to define MI based on entropy differences, each providing a different perspective on the interpretation of the resulting measure. For illustration within a neuroimaging context, S denotes a random variable representing some stimulus feature that is varied

[3] $\log_2 6 - \log_2 3 = \log_2 \frac{6}{3} = \log_2 2 = 1$

across multiple presentations (e.g. in the visual domain, edge contrast, or orientation, or opacity; in the auditory domain loudness, pitch and so forth) and *R* some neural response (e.g. EEG voltage, MEG source amplitude or fMRI bold voxel response measured at a specific site at a specific post-stimulus latency).

$$\begin{aligned} I(R;S) &= H(S) - H(S \mid R) \\ &= H(R) - H(R \mid S) \\ &= H(R) + H(S) - H(R,S) \end{aligned} \tag{1}$$

Here, $H(S \mid R)$ is the conditional entropy: the entropy of *S* when the value of *R* is known[4]. $H(R,S)$ represents the entropy of the joint distribution of *R* and *S* (i.e. the two dimensional variable obtained by combining R and S).

The first expression in Eq. (1) shows that MI quantifies the average reduction in uncertainty about which stimulus *S* (e.g. edge contrast, or auditory pitch) was presented after observation of a response *R* (e.g. MEG source amplitude or fMRI bold response). The second expression demonstrates the symmetry of MI, and shows that it equally quantifies the average reduction in uncertainty about the neural response when the stimulus is known. Here it is useful to revisit the analogy with variance – MI measures the *entropy explained* by knowledge of the second variable, which is conceptually similar to the notion of *variance explained* from regression models. However, unlike a regression, MI makes no assumption on the form of the relationship and is symmetric in the two variables considered.

The third expression in Eq. (1) shows that MI quantifies the difference in entropy between a model in which the two variables are statistically independent[5] and the true joint distribution ($H(R,S)$). In fact, in the discrete case, MI is proportional (with a scale factor depending on the number of data points) to the effect size for the log-likelihood ratio test of independence, often called the G-test (Sokal and Rohlf, 1981). The Neyman-Pearson lemma (Neyman and Pearson, 1933) states that for a given significance level, the likelihood ratio test is the most powerful statistical test for comparing two nested models (used here to test for independence). This motivates perhaps the most useful interpretation of MI from a neuroimaging perspective: a statistical test for independence. It is worth repeating that all three expressions above are mathematically equivalent, but considering them separately explains the different interpretations that can be applied to MI.

---

[4] This is defined as the expectation (over values *r* of *R*) of the entropy of the distribution of *S* conditional on *r*.

[5] The statistically independent model is given by the product of the marginal distributions of the two variables, with entropy $H(R)+H(S)$. In fact MI can be written in another form – as the Kullback-Leibler (KL) divergence (a measure of distance between probability distributions) between the statistically independent model and the true joint distribution (Akaike, 1992).

MI has several useful properties that are worth highlighting. As discussed above, it is symmetric in the variables considered. It is also additive for independent variables. Additivity derives directly from the mathematical properties of the logarithm: if two variables are statistically independent their joint probability is, by definition, the product of their individual probabilities and therefore the log joint probability is the sum of the individual log probabilities. This means the joint entropy of two independent variables is the sum of the individual entropies. Similarly, the MI between independent pairs of variables is added when they are considered jointly[6]. This is a crucial property that is not shared by the effect sizes of other statistical tests and which enables direct quantification of pairwise interaction effects (see Section 2.6). MI measures dependence on a common scale (bits), which provides a meaningful effect size (Friston, 2012) and allows direct comparisons across different responses, experimental modalities or with behavior.

Within the field of information theory there are many mathematical results revealing further properties of the MI measure. A theorem called the Channel Coding Theorem (Cover and Thomas, 1991) relates MI to the transmission capacity of noisy communication channels. A noisy channel can be represented by a conditional probability distribution quantifying the relationship between the output symbols *y* and the input signals *x*: $P_{Y|X}(y \mid x)$. This is a fixed property of the channel, but the MI between *x* and *y* depends also on the distribution of the inputs $P_X(x)$. The maximum rate at which information can be transmitted over the noisy channel without errors (the channel capacity) is given by the maximal value of MI over all possible input distributions. MI was originally developed within this coding framework, which represents communication as the transmission of a set of discrete symbols over a noisy channel. MI provides theoretical limits on communication efficiency and helped to formulate coding principles that are now pervasive in modern communication systems. This interpretation also motivated much of the early application of MI within neuroscience, viewing the neural pathway from the stimulus receptors to the recorded brain activity as a noisy communication channel, and using MI to quantify properties of this putative communication channel as well as investigating encoding and decoding schemes at different neural levels. Another theorem called the Data Processing Inequality (Cover and Thomas, 1991) states that post-processing cannot increase information[7]. This is a desirable property for a neuroimaging statistic as it ensures that any signal processing or feature extraction applied to the recorded responses (for example spectral analysis) cannot artificially inflate the measured effect size, provided it is applied across the whole data set without incorporating knowledge of the stimulus.

---

[6] Formally, if $P_{ABCD}(a,b,c,d) = P_{AB}(a,b)P_{CD}(c,d)$ then $I(A,C;B,D) = I(A;B) + I(C,D)$

[7] Formally, if the response *R* is transformed to a new representation *P*, where *P* is a probabilistic function of *R* only (and does not depend on *S*), then $I(R;S) \geq I(P;S)$.

Given the above, we suggest there are two views that can be adopted when applying MI in practice (Nelken and Chechik, 2007). The first relies on the coding interpretations of MI, and therefore requires accurate and bias-free estimates; this view has driven most neuroscience applications of MI to date. The second view considers MI more like a conventional statistical hypothesis test of independence, comparable to a t-test or correlation. With this view, accurate bias-free values are less important, but determining statistical significance, including accounting for the problem of multiple comparisons, is crucial. We suggest that this second view is more useful for neuroimaging. In comparison to other statistical tests, MI brings the advantages of sensitivity and robustness (demonstrated in Results), as well as additivity, which allows direct comparisons of the MI from different neural responses. With the novel estimator presented here, MI allows treatment of discrete, continuous and possibly multidimensional variables within a common framework with directly comparable effect sizes on a common and meaningful scale.

2.3. Existing methods for calculating mutual information

There are several different practical approaches for calculating MI from experimental observations, which we briefly review here from a neuroimaging perspective.

Binned methods

Although neuroimaging signals typically take continuous values (voltage in EEG, magnetic field strength in MEG, fMRI BOLD signal amplitude) one commonly used approach for such signals consists of quantizing the continuous valued observations into a set of discrete categories or bins. We therefore briefly review methods for estimating information values on discrete spaces. There are two main strategies for the quantization step: either bins are set with equal spacing, or bins are sized to have approximately equal occupancy (i.e. if using four bins, each data sample is labeled according to the quartile of the empirical distribution in which it lies). For data that are approximately normally distributed, the second method is preferable, as fixed width bins result in the extreme bins having few samples, which exacerbates the limited sampling problem. This is the simplest approach; much work has considered different methods to optimize this quantization step (Darbellay and Vajda, 1999; Endres and Foldiak, 2005; Fraser and Swinney, 1986; Reshef et al., 2011).

Once the continuous signal is quantized, a common approach is to apply multinomial maximum likelihood estimation (the histogram estimate) of the underlying distributions and calculate entropy and MI from their definitions using these distributions (e.g. combining the entropy definition of Figure 1A with Eq (1)). This approach is variously called the naïve, direct or plug-in estimate (Cover and Thomas, 1991). The resulting information estimate is linearly related to the G-test,

a log-likelihood ratio test of independence (Sokal and Rohlf, 1981)[8]. In fact, it is suggested that the chi-square test was originally developed as an approximation to the G-test to avoid the requirement for the laborious calculation of logarithms before the advent of modern computers (Sokal and Rohlf, 1981). The Neymann-Pearson lemma proves that the likelihood-ratio test (and hence MI) is the most powerful hypothesis test for a given significance level (Neyman and Pearson, 1933), a fact which, coupled with the computational efficiency of calculating binned MI, demonstrates the potential usefulness of this quantity as an exploratory statistical test for use in neuroimaging. Other properties of the MI effect size, such as additivity, provide additional advantages (Section 2.6).

However, when trying to obtain accurate MI estimates for interpretation from a coding channel perspective, there is a problem that the plug-in estimate described above is biased upwards; estimates calculated from finite numbers of samples will be higher than the true value even when averaged over many sample sets. A wide variety of approaches have been proposed to address this problem (Paninski, 2003; Panzeri et al., 2007). The simplest approach is to subtract the mean of the distribution expected under the null hypothesis that there is no relationship between the two variables[9]. This is called the Miller-Madow correction (Miller, 1955). Various extensions have been proposed to deal with this in different situations such as Bayesian approaches (Nemenman et al., 2004; Panzeri and Treves, 1996), specific methods to estimate entropy and information rates in ongoing processes (Kennel et al., 2005; Shlens et al., 2007) and to deal specifically with the sparse binary probability spaces that result from measuring single neuron spiking activity (Archer et al., 2013; Montemurro et al., 2007). However, we stress again that if the goal of the analysis is classical statistical inference, bias correction is not necessary, and can actually reduce statistical power due to the increased variance of bias corrected estimators (Ince et al., 2012).

A further consideration with binned methods is that they suffer from the curse of dimensionality (Geman et al., 1992). The number of parameters that must be estimated for the multinomial distributions grows exponentially with the number of variables considered. This renders it practically impossible to apply this approach to calculate MI and higher order information-theoretic quantities (Sections 2.8, 2.9) from multivariate spaces with the amounts of data that can be realistically collected from neuroimaging experiments. It may be possible to exploit techniques such as dimensionality reduction or clustering approaches, often referred to as vector quantization in this context (Wilcox and Niles, 1995), to directly quantize multivariate spaces into a small number of representative

[8]The G-test statistic is equal to MI multiplied by a factor $2N\log(2)$, and is chi-square distributed with the same degrees of freedom as the corresponding chi-square test: $2N\log(2)I \sim \chi^2(df)$, $df = (|R|-1)(|S|-1)$

[9] As can be seen from the relationship with the chi-square distribution described in the previous paragraph this is given by $(|R|-1)(|S|-1)/2N\log(2)$

symbols. However, such an approach also removes the possibility of investigating the effects of the different variables in the multivariate space, for example by considering the MI of each variable individually, or investigating the effect of correlations between them (Chicharro, 2014; Magri et al., 2009; Panzeri and Treves, 1996).

Continuous methods

Several methods exist for estimating MI between continuous variables without the quantization step. The most direct way is to first estimate the continuous probability distributions with a Kernel Density Estimation (KDE) technique, and then numerically integrate those estimates to obtain MI (Moon et al., 1995). An alternative approach, which bypasses explicit estimation of the joint distributions, exploits the relationship between probability density and local neighborhood structure. These methods estimate entropy and MI using k-nearest neighbor structure (Faivishevsky and Goldberger, 2009; Kraskov et al., 2004; Victor, 2002). While these methods are accurate and relatively unbiased, they are computationally intensive. This is particularly problematic when combined with permutation testing in a mass-univariate neuroimaging context (Groppe et al., 2011).

An alternative approach to dealing with continuous data is to assume a parametric form for the distribution. For example, Local Field Potentials (and similarly M/EEG data) are often approximately Gaussian (Magri et al., 2009). The parameters of this assumed distribution can be estimated from the data and the entropy and MI values estimated directly from the parametric model, for which there are often closed form expressions solving the integral definition of these quantities. While different parametric models can be used (for example, a t-distribution might be more appropriate for M/EEG data), and this approach is computationally efficient, it is not clear what effect a violation of the parametric assumptions would have on the estimate. Further, there are many variables of interest, for example stimulus features obtained from dynamic naturalistic stimuli, which do not have an obviously appropriate parametric form.

2.4. Estimating the entropy and MI of Gaussian variables

For Gaussian variables, the integral definition (Figure 1A) can be solved analytically resulting in a closed form expression for the entropy (in bits) as a function of the determinant of the covariance matrix Σ:

$$H(X) = \frac{1}{2\ln 2}\ln\left[(2\pi e)^k |\Sigma|\right] \tag{2}$$

This measure still exhibits some bias due to the estimation of the covariance matrix from limited data, but there exists an analytic correction to remove much

of this effect[10] (Goodman, 1963; Magri et al., 2009; Misra et al., 2005). From equation (1), the MI between two Gaussian variables is therefore given by:

$$I(X;Y) = \frac{1}{2\ln 2}\ln\left[\frac{|\Sigma_X||\Sigma_Y|}{|\Sigma_{XY}|}\right]$$

where Σ_X, Σ_Y are the covariance matrices of X and Y respectively and Σ_{XY} is the covariance matrix for the joint variable *(X,Y)*.

Figure 3 shows the MI between two 1-D Gaussian variables as a function of their correlation. This reveals two key properties. First, the symmetric shape of the graph demonstrates how, since MI is an unsigned quantity, it can reveal the strength but not the direction of a relationship; this is an important aspect to keep in mind. Second, the relationship is clearly non-linear. We suggest that this non-linearity is an advantage, especially for neuroimaging studies, since it results in an enhanced contrast of strong effects with respect to background values in mass-univariate analyses.

While the parametric Gaussian approach is data robust due to the relatively low number of parameters that need to be estimated, it is not clear how the estimator might perform if the Gaussian distribution assumption was violated, and it cannot be employed in many cases where the distribution of stimulus values is highly non-Gaussian.

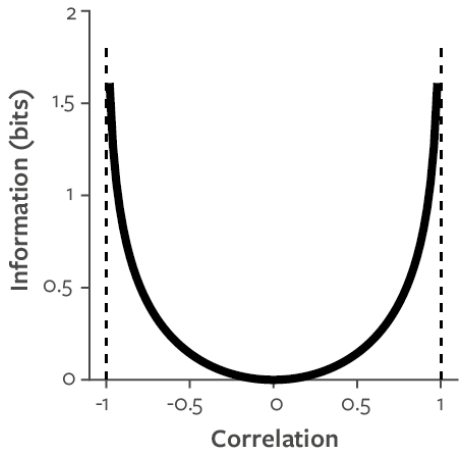

Figure 3 *Relationship between correlation and information for two 1-D Gaussian variables.*

2.5. Estimating MI within different types of experimental design

MI is, like Pearson correlation, a function of two variables that can be applied in practice to many different sorts of data. In order to correctly interpret a particular information theoretic analysis it is important to understand how the

[10] The bias-corrected entropy estimate is given by

$$H(X) = \frac{1}{2\ln 2}\left(\ln\left[(2\pi e)^k|\Sigma|\right] - p\ln\frac{2}{N-1} - \sum_{i=1}^{k}\Psi\left(\frac{N-i}{2}\right)\right)$$

samples used to estimate MI were obtained. Given the diversity of experimental designs employed in neuroimaging, the different approaches to obtaining samples are often a point of confusion. In this section we describe some common experimental designs and detail how MI is estimated in each case.

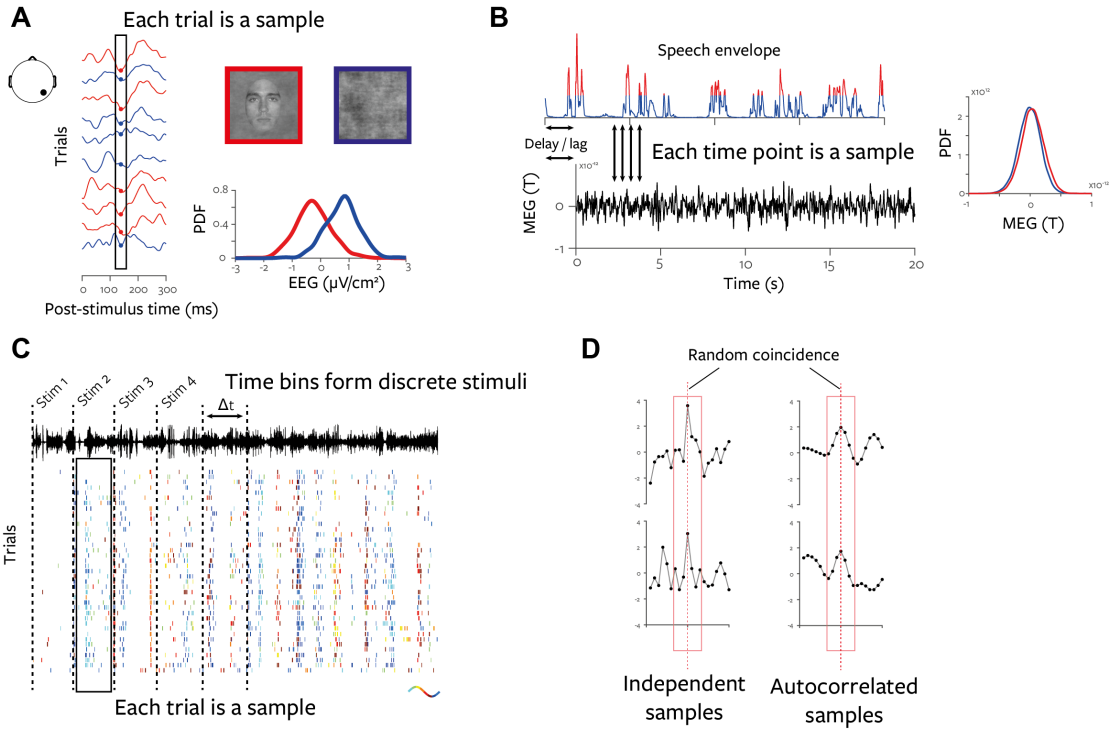

Figure 4 *MI calculation for different experimental designs.* Schematic illustrations show how samples used to estimate MI are obtained from different neuroimaging experimental paradigms. **A.** Event related design. Example data from a single sensor are recorded to repeated presentations (trials) of two classes of stimuli, faces (red) or noise images (blue) (Rousselet et al., 2014a). Values are extracted across presentations for a specific post-stimulus time at a specific sensor; these form the samples used for MI calculation. Kernel smoothed PDF estimates are shown for the example time point. **B.** Continuous design. Here the amplitude envelope of a speech stimulus is effectively cross-correlated with an MEG sensor signal (Gross et al., 2013). **C.** Hybrid design. A short section of a dynamic stimulus is presented many times. The stimulus is divided into bins, each of which is treated as a separate discrete stimulus. The responses over the repeated presentations (trials) of the stimulus are used as samples (Kayser et al., 2012). **D.** Two pairs of random signals with no autocorrelation (left) and with autocorrelation induced by low-pass filtering (right) are shown. The dashed red line indicates a random coincidence of high values, the red box highlights the additional relationship between the neighbouring points induced by the autocorrelation.

Event related design

An event-related design consists of serial presentation of stimuli, possibly from different classes or with parametrically varying features, and possibly requiring a behavioral response. These presentations are separated in time and the analysis begins by extracting sections of neuroimaging recordings following each presentation. Event-related experiments are typically analyzed by averaging response epochs to different stimulus classes (resulting in an Event-Related Potential or ERP) and performing group statistics. However, using MI we can also quantify modulation of the M/EEG response by a continuous stimulus feature (e.g. a variation of stimulus orientation) that varies across trials. To apply MI in this

paradigm, each post-stimulus time point and sensor/source is treated independently within a mass-univariate framework (Groppe et al., 2011). The MI calculation is repeated for each time point and sensor/source, using the repeated presentations of the stimulus as samples (Figure 4A). Multiple comparison correction is required over time points and sensors/sources: this can be achieved using permutation testing (repeating the calculation with shuffled stimulus values) combined with the method of maximum statistics (Holmes et al., 1996; Nichols and Holmes, 2002), or cluster sum statistics (Maris and Oostenveld, 2007) possibly with threshold-free cluster enhancement (Pernet et al., 2015; Smith and Nichols, 2009). An advantage of this design is that, because each time point is analyzed separately, there is no assumption that the signal is stationary.

Continuous design

In a continuous design an ongoing, usually naturalistic, dynamic stimulus is presented, for example a visual movie or auditory speech. The goal of the analysis is to determine a relationship between time varying stimulus features and the M/EEG signal. The analysis is performed separately for each sensor/source and is similar to a cross-correlation. A particular delay or lag is chosen and the MI is calculated using the values of the lagged signals over time as samples (Figure 4B). Any inference requires multiple comparison correction over features, delays and sensors/sources. Since the signals usually exhibit strong autocorrelation, the permutation strategy needs to take this into account. Autocorrelation can strongly alter the distribution of the MI under the permutation null hypothesis, because if a pair of peaks in two signals happened to coincide by chance, many neighbouring points would also coincide (Figure 4D). This structure is lost if the time-domain samples are permuted without preserving the autocorrelation. We therefore suggest using a blockwise permutation approach when calculating MI in this sort of experimental design (Adolf et al., 2011). The continuous design also imposes an implicit assumption that the neural processes under consideration are stationary for the duration of the presented stimulus.

Hybrid design

A third approach is a hybrid design that combined elements of both the event-related and continuous designs described above (Figure 4c). Here, a short segment of a dynamic naturalistic stimulus is presented many times. The time-course of the dynamic stimulus is split into a number of fixed-width time windows, each of which is treated as a discrete categorical stimulus. The responses obtained during that time window across the repeated presentation are used as samples for the MI calculation. This approach has frequently been applied with electrophysiology data (Strong et al., 1998) because it results in an efficient use of experimental time, and requires no prior assumptions on the specific stimulus features driving the neural response. MI calculated in this design quantifies the overall reliability of the modulation of the neural response by the stimulus without considering specific stimulus features.

As a measure of dependence MI is a function of two paired sets of samples. However, in practice sets of samples can be obtained in different ways, depending

on the experimental designs just reviewed: across experimental trials, time points, or through some combination of the two. To enable meaningful interpretation of any estimated MI quantity, it is critical to properly understand how the samples were obtained via the experimental design. So, we recommend clear reporting of these design details whenever MI quantities are used.

2.6. Higher order information theoretic quantities

We here review some higher order information theoretic quantities and their application to brain imaging that we believe provide particularly useful and novel insights in the analyses of neuroimaging data. We describe Conditional Mutual Information, which can isolate the specific effect of a stimulus feature on neural response, while controlling the potential contribution of other correlated stimulus features; Directed Information which quantifies the time-lagged causal transfer of information between two neural responses; Interaction Information which quantifies the similarity (or synergy) of representation of the same stimulus feature between two neural responses; and Directed Feature Information which measures the time-lagged causal communication of a specific stimulus feature between two neural responses.

Conditional Mutual Information

Conditional Mutual Information (CMI) (Cover and Thomas, 1991) quantifies the relationship between two variables while removing any effect of a third variable[11]. It is the information-theoretic analogue of partial correlation. However, while partial correlation removes only the linear effects of the third variable, CMI controls for effects of all orders and so allows for stronger conclusions to be drawn. With many types of naturalistic stimuli, extracted stimulus features are highly correlated (for example, luminance of neighboring pixels of a natural image or the acoustic features of speech). Given an analysis of each feature alone, it is difficult to determine whether a specific feature is genuinely encoded in a neural response, or whether the response is actually modulated by a different correlated stimulus feature. CMI provides a rigorous way to address this issue (Ince et al., 2012), allowing strong conclusions to be drawn about the relationship between neural responses and multiple correlated stimulus features. Figure 5 demonstrates with a simulation the use of CMI to dissociate two possible situations where a response is modulated by two correlated stimulus features.

[11] CMI between *X* and *Y*, conditioning out *Z* is usually denoted $I(X;Y \mid Z)$

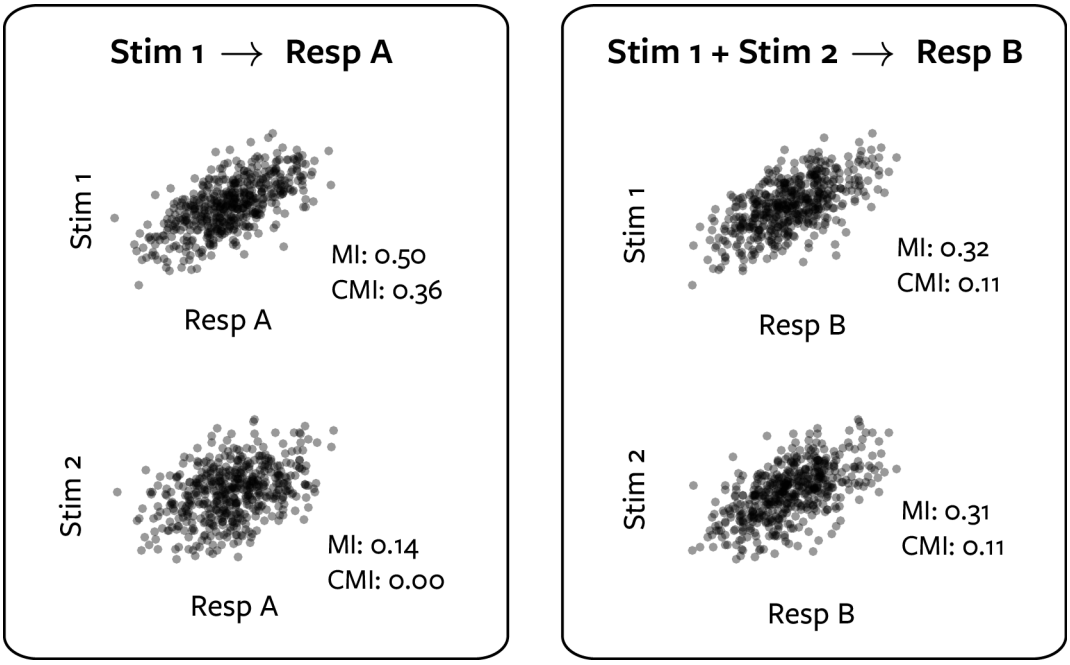

Figure 5 *CMI reveals genuine encoding of correlated stimulus features.* In this simulation we generated two correlated Gaussian stimulus features (covariance = 0.6), Stim 1 and Stim 2. We generated responses from two different models. Response A (left) was obtained from stimulus 1 plus Gaussian noise (top plot). In the bottom plot, MI reveals a relationship with stimulus 2 (MI = 0.14), but CMI reveals this is due only to the relationship between the features (CMI = 0). Response B (right) was obtained on each trial from the sum of stimulus 1 and stimulus 2 plus Gaussian noise. MI again reveals response dependence with both Stim 1 and Stim 2, but CMI (= 0.11) shows that now each stimulus is genuinely represented in the signal.

The ability to combine continuous and discrete variables allows for CMI to provide a novel approach to group statistics. We can construct a discrete variable representing participant identity (*P*) and use this as the conditioning variable in the definition of CMI. We can calculate the MI from the data pooled over participants (with normalization done on a per-participant basis to account for signal differences), and also the CMI conditioned on participant identity. This CMI is the average MI effect size within each participant. These quantities are then closely related to the quantities used in the replicated G-test for independence (Sokal and Rohlf, 1981), which provides three useful inferences for group studies: First, is there a significant effect over the group (*total-G*, equivalent to CMI $I(S;R|P)$)? This is closest to classical group inference and tells that overall the members of the group deviate from the null hypothesis (but they may not all do so and they may not deviate in the same way). Second, is there a significant difference in the effect between participants (*heterogeneity-G*, equivalent to CMI – MI, or alternatively $I(S,R;P)$)? If so, this indicates the effect is not consistent between participants and so the data should not be pooled. This is particularly useful given that MI is an unsigned quantity – all participants could have similar levels of MI but with opposite signs of effect. It can also help to identify situations where group significance is driven by a strong effect in one or a few participants, rather than a consistent effect across the group. This is something that is not considered with most existing group statistics. Third, is the effect significant if the data is pooled (*pooled-G*, equivalent to MI)? This means overall, the data recorded across all participants deviates from the null hypothesis. This allows the

identification of cases where there is a weak but consistent effect across participants, which might not suffice to produce a significant CMI value. While this approach requires further development and testing we mention it here to motivate some of the wide range of potential applications for CMI within neuroimaging.

Directed Information (Transfer entropy)

CMI also forms the basis for an information theoretic approach to the analysis of causal relationships between neural responses. Calculating CMI between the values of a signal *Y*, and the values of a signal *X* earlier in time, conditioning on the earlier values of *Y* itself produces a measure originally termed Directed Information (DI) (Massey, 1990) but frequently referred to as Transfer Entropy (TE) (Schreiber, 2000). DI measures the time-lagged dependence between two signals, over and above the dependence with the past of the signal itself (its self-predictability; Figure 6). This is the information theoretic analogue of Granger causality (Barnett et al., 2009; Granger, 1969) and following the arguments developed by Wiener and Granger (Granger, 1969; Wiener, 1956) can be used to infer causal relationships between brain signals, with some caveats (Bressler and Seth, 2011; Chicharro and Ledberg, 2012; Chicharro and Panzeri, 2014; Quinn et al., 2011; Wibral et al., 2014). The Gaussian copula estimate we present below (Section 3) provides a robust and computationally efficient method to estimate DI (Ince et al., 2015).

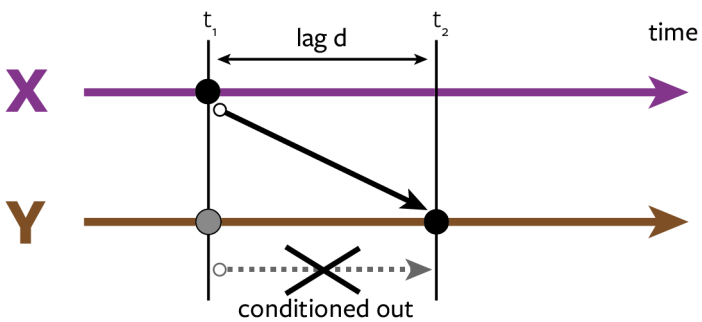

Figure 6 *Schematic of Directed Information (DI) calculation.* DI from *X* to *Y* is calculated as the CMI between the activity of *X* at time t_1, and the activity of *Y* at a later time t_2, conditioned on the activity of *Y* at t_1: $I\left(X_{t_1};Y_{t_2} \mid Y_{t_1}\right)$.

Interaction information

It is now widely accepted that rather than operating as a number of separate functional units, the brain functions as a highly interactive distributed network. For neuroimaging studies to fully embrace this perspective we require tools to relate experimental modulations (e.g. effect of different experimental conditions) or stimulus modulations (effect of changing stimulus features) across multiple responses. For example, univariate MI analyses might reveal stimulus modulation in distinct spatial, temporal or spectral regions (Figure 7A). A natural question that arises is: Are the modulations we observe in both regions similar, or different? In other words, are the responses in both regions representing the stimuli in a similar, or a different manner?

Focusing on pairs of responses, we can schematize the stimulus MI in each response in a Venn diagram (Figure 7B). One quantity of interest is the overlap, termed the *redundancy* (Panzeri et al., 2008; Schneidman et al., 2003; Timme et al., 2013), that quantifies the MI which is shared between the two responses. Because of the additivity of MI we can obtain this quantity directly: summing the MI available in each response separately counts the overlapping region twice. We then subtract the MI available when considering both responses together, which counts the overlapping region once. The resulting value quantifies the redundancy and is equal to the negative Interaction Information (II) (McGill, 1954). Negative II corresponds to redundancy as described above, but II can also be positive, indicating *synergy* between the variables. In this case, the MI in the pair of responses when considered jointly is greater than the MI when they are considered separately. This implies that the relationship between the responses on individual trials is itself modulated by the stimulus feature considered. Redundancy is bounded above by three quantities, the MI between the stimulus and each response and the MI between the responses themselves, so it can be normalized by the minimum of these three values.

Within a neuroimaging context, high redundancy would suggest the two responses reflect the same aspects of the stimulus, and therefore likely reflect the same processing pathway or mechanisms. Alternatively, independence (zero interaction information) would suggest different processing pathways produce the observed responses. This approach can also be applied to compare different response representations or to compare responses from different experimental modalities (e.g. simultaneously recorded fMRI + EEG, Figure 7A). Similarly, interaction information can be applied in the opposite direction, considering two stimulus features (possibly from different modalities) and a single neural response, and quantifying whether they modulate the neural response in a synergistic or redundant fashion. The multivariate performance of the Gaussian copula estimate we present below (Section 3) is crucial to allow accurate estimation of the joint MI in pairs of neural responses (which themselves can be multivariate) required for computing interaction information.

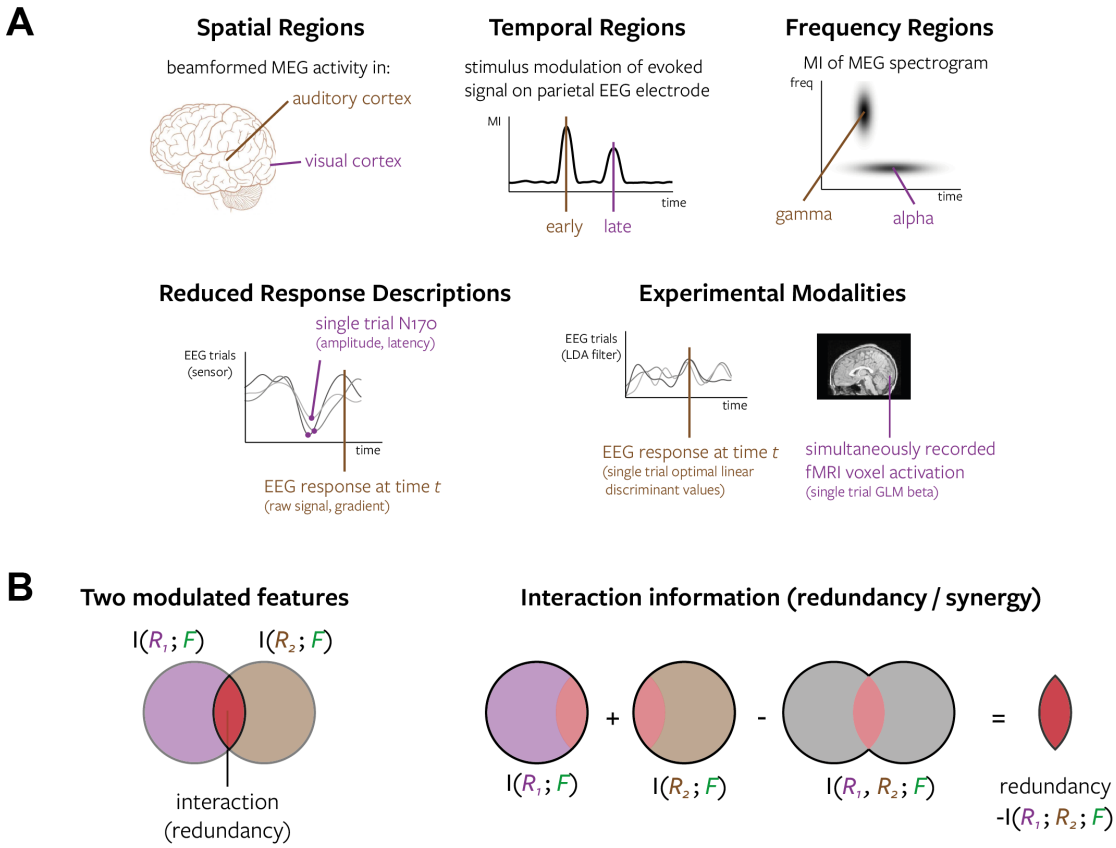

Figure 7 *Interaction Information: Redundancy and Synergy between neuroimaging responses*. **A**. Example situations where two different neuroimaging responses are modulated by a stimulus or task condition, and it would be of interest to relate the modulation, or information content, of the two signals. **B**. The additivity of MI allows us to quantify the redundancy (overlap) directly (see main text).

It should be noted that one issue with interaction information is that synergistic and redundant effects can cancel in the final average. While it is not clear to what degree such cancellation might occur in neuroimaging recordings this is a question that has recently received much interest (Bertschinger et al., 2014; Griffith and Koch, 2012; Harder et al., 2012; Olbrich et al., 2015; Timme et al., 2013; Williams and Beer, 2010). Further development and application of techniques to address the interplay between synergy and redundancy within a neuroimaging context is an important area for future work.

Interaction information can be applied to relate the information content of different neuroimaging responses, revealing redundant or synergistic representations. Alternative techniques to address these questions include classification images, representational similarity analysis and the temporal generalization method. Classification images (Murray, 2011) obtained from reverse correlation of different neural responses can be directly compared to quantify similarity in stimulus representation (redundancy) between areas (Smith et al., 2004). A similar approach is employed in Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008; Kriegeskorte and Kievit, 2013) which compares representational geometries between different neural responses (not the information content in the neural responses, as with classification images) by correlating dissimilarity matrices obtained from discrete category exemplar

stimuli. The temporal generalization method (King and Dehaene, 2014), in which a classification algorithm is trained with neural responses from one time point and then tested on another time point, can reveal similar representations between the two time points. All of these methods are conceptually similar to redundancy, but the information theoretic approach can also reveal synergistic effects, can combine discrete and continuous stimuli or responses, can be calculated while conditioning out correlated stimulus features and can be applied to univariate responses with the full spatial or temporal precision of the considered recording modality. It also provides results with a meaningful common effect size (bits), or alternatively redundancy can be normalized to a percentage that provides an intuitive measure of the degree of overlap.

Directed Feature Information

We recently developed a new measure of functional connectivity called Directed Feature Information (DFI) (Ince et al., 2015). It conceptually extends DI, a measure of the amount of causal communication between two regions, to quantify the communication that is *about* a specific stimulus feature. That is, as DI measures the time-lagged relationship between the responses of two regions, DFI quantifies the amount of DI that can be attributed to variations of a given stimulus feature (e.g. the graded presence of a face in the stimulus (Ince et al., 2015)). An alternative interpretation using redundancy is that DFI quantifies the amount of redundant MI about the stimulus that is shared between Y and the past of X, over and above that which is already present in the past of Y. Therefore, following the Wiener-Granger principle, DFI can be used to infer the communication of the specific information about the feature considered from X to Y. DFI enables the construction of networks based on the communication of specific, task-related stimulus features rather than the networks typically constructed from the overall dependence between activity in different areas that may or may not be directly task or stimulus related. Task effects on functional connectivity are increasingly addressed with tools such as psychophysiological interactions (PPI) analysis (Friston et al., 1997; O'Reilly et al., 2012) which reveals task induced changes in connectivity within the framework of Dynamic Causal Modelling (Friston et al., 2003). However, to our knowledge there is no other measure of functional connectivity that directly quantifies the specific content of communication as DFI does, and so it represents a transformative perspective for network-based analysis of neuroimaging data. The Gaussian copula method presented below (Section 3) is crucial to allow accurate estimation of DFI, since it requires an additional conditioning step, to calculate DI conditioned on the stimulus. The generality of the Gaussian copula estimate means this quantity can be applied to a range of situations, with discrete or continuous stimuli and potentially considering multivariate dynamic responses.

2.7. Relation between information theoretic quantities and other statistical approaches

Table 1 shows equivalent statistical approaches to address the same questions as the information theoretic quantities reviewed above. This illustrates how the

information theoretic framework unifies a wide variety of statistical approaches with effect sizes on a common scale across many different situations.

Information theoretic quantity	Other statistical approaches
Mutual Information (discrete; discrete)	Chi-square test of independence Fishers exact test
MI (univariate continuous; discrete)	2 classes: T-test, KS-test, Mann-Whitney U test ANOVA
MI (multivariate continuous; discrete)	2 classes: Hotelling T^2-test Predictive decoding approaches (machine learning classification with cross validation)
MI (univariate continuous; univariate continuous)	Pearson correlation Spearman rank correlation Kendall rank correlation
MI (multivariate continuous; univariate continuous)	Generalized Linear Model framework Predictive decoding approaches (machine learning regression with cross validation)
MI (multivariate continuous; multivariate continuous)	Canonical correlation analysis Distance correlation; Brownian covariance
Conditional Mutual Information	Partial correlation (continuous variables and linear effects only)
Directed Information	Granger causality (equivalent for Gaussian autoregressive processes)
Directed Feature Information	Dynamic Causal Modeling Psychophysiological Interactions
Interaction Information	Representational Similarity Analysis (redundancy only) Cross-temporal decoding (redundancy only, temporal interactions only) Mediation analysis

Table 1 *Relation between information theoretic quantities and other statistical approaches.*

3. A novel method for MI estimation using a Gaussian copula

3.1. Estimating MI between two continuous variables with a Gaussian copula

We present here a new estimator of MI that uses the concept of a statistical copula to provide the advantages of Gaussian parametric estimation (Section 2.4) for variables with any marginal distributions. A copula (Nelsen, 2007) is a statistical structure that expresses the relationship between two random

variables, independently of their marginal distributions. Sklar's theorem (Sklar, 1959) states that any multivariate distribution can be expressed as the combination of univariate marginal distributions and an appropriate copula – the copula links the individual variables and represents the statistical relationships between them. Copula based analyses have been widely applied in quantitative finance (Genest et al., 2009) and have recently been applied to estimate Granger causality in neuroimaging data (Hu and Liang, 2014).

Figures 8A and 8B illustrate the copula concept with simulated Gaussian data for uncorrelated and correlated variables respectively. Left hand scatter plots show 1000 simulated data points. Right hand scatter plots show the empirical copula of this simulated data: the empirical cumulative distribution function (CDF) value of each variable evaluated at each data point. The empirical CDF is calculated by ranking each sample (separately for each variable) and then rescaling the integer rank values to the range (0,1). Thus, the resulting bi-variate distribution (copula) is a probability density over the unit square. For independent variables (Fig. 8A) the copula is uniform, while for correlated variables (Fig. 8B) the copula has non-uniform density.

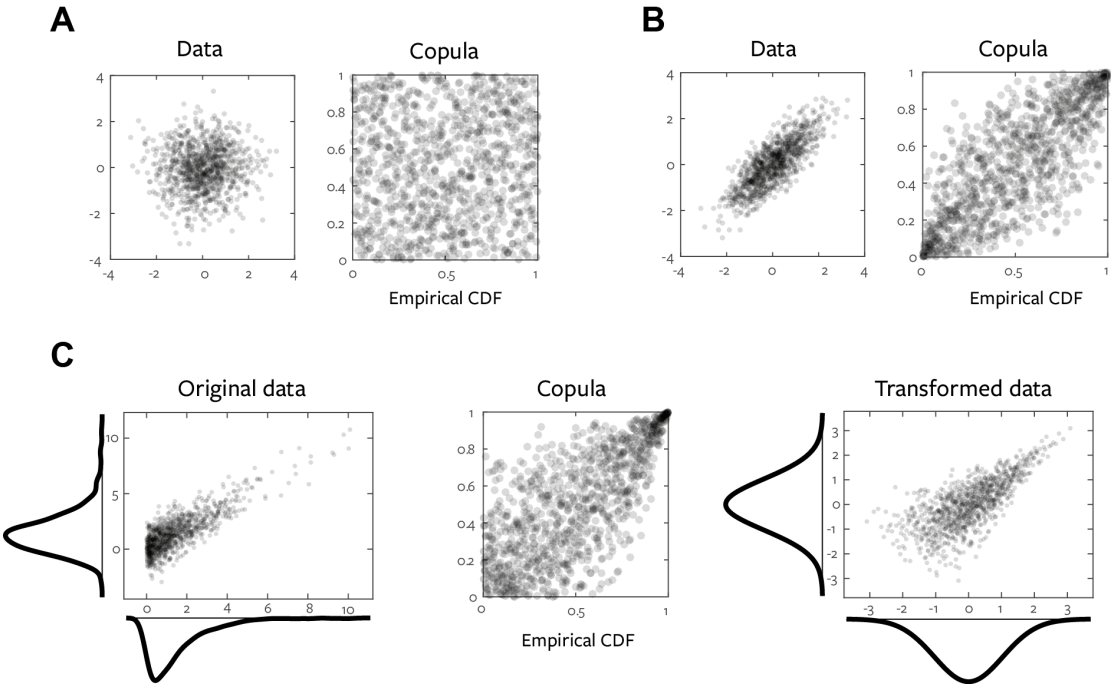

Figure 8 *Examples of Gaussian copulas and a copula-preserving Gaussian marginal transformation.* **A.** Scatter plots of simulated data from two independent standard normal variables, and their copula. **B.** Scatter plots of simulated data from two correlated standard normal variables (r=0.8), and their copula. **C.** Scatter plot and marginal densities of data simulated from the model x=exp(1.5); y=x+N(0,1) (left), the empirical copula (center) and the transformed data with Gaussian marginal but the same empirical copula as the original data (right).

This is useful because the copula linking two variables is directly related to the MI between them. The entropy of the joint distribution over two variables is equal to the marginal entropies plus the entropy of the copula. Plugging this into the

third form of Eq. (1), the marginal entropies cancel revealing that the MI between X and Y is equal to the negative entropy of their copula (Calsaverini and Vicente, 2009; Kumar, 2012; Ma and Sun, 2011; Zeng and Durrani, 2011); thus the copula fully encapsulates the relationship between the two variables. A corollary of this is that MI does not depend on the marginal distributions of the individual variables.

We can therefore estimate MI by applying continuous entropy estimates to the empirical copula (Ma and Sun, 2011); however these can still be computationally and data intensive. Instead we exploit the corollary mentioned above; since the copula entropy, and hence the MI, does not depend on the marginal distributions of the original variables, we can transform the marginals in any way we see fit. As long as we preserve the empirical copula linking the variables, the statistical relationship that is quantified by MI will be unchanged. We therefore transform the marginals to be standard Gaussian variables, to which we can apply the efficient parametric MI estimate described in the previous section.

Figure 8C illustrates this transformation; the variable plotted on the x-axis is drawn from an exponential distribution (rate=1.5); the variable on the y-axis is that value added to a standard normal. The left hand scatter plot of 1000 samples from this model shows the non-Gaussian marginal distributions of these data and the copula plot (center) illustrates the dependence between the variables. For each sample, the transformed value of each variable is obtained as the inverse standard normal CDF evaluated at the empirical CDF of that sample. This transformation produces a data set that has perfect standard normal marginals, and the same empirical copula as the original data (right). In other words, the rank-relationships between the variables are preserved. We can then calculate MI between the transformed variables using the parametric Gaussian model (Eqs 1,3).

The parametric MI estimation implicitly imposes the assumption of a Gaussian copula linking the two variables. If this assumption is violated the resulting MI value may not be accurate. However, since the Gaussian distribution has the maximum entropy for a given mean and covariance (Cover and Thomas, 1991) the Gaussian copula must also have the maximum entropy of possible copula models preserving second order statistics[12]. Since MI is the negative copula entropy other choices of parametric copula models (or direct estimation) could give higher, but not lower MI estimates: the Gaussian copula estimate is therefore a lower bound to the true MI value (Calsaverini and Vicente, 2009). This lower bound property is crucial for an estimator that is to be used for statistical testing, since it ensures that erroneous high values cannot occur due to mismatched assumptions between the statistic and the data; the measured value is always lower than the true value. In the multivariate case, the same Gaussian marginal transformation is applied independently to each constituent variable. This preserves the rank

[12] Otherwise the distribution of the same two Gaussian variables linked by this higher entropy copula would itself have a higher entropy, contradicting the proven maximum entropy property of the Gaussian.

relationships both within and between the two multivariate variables considered for the MI calculation (X and Y in Eq. 1). Zeng and Durrani (2011), propose to estimate MI between univariate variables via a Gaussian copula, the entropy of which is estimated via Kendall's tau. The advantage of the approach presented here is that it can be applied multidimensional variables (see Section 4) and to estimate MI between discrete and continuous variables (Section 3.2).

In summary, by transforming each univariate marginal to be a standard normal and applying a Gaussian parametric MI estimate, we obtain a lower-bound estimate of the MI. As the value of this estimate depends only on the empirical CDF of the data, it is in effect a rank statistic and so robust to outliers. Although the estimate derives from a parametric assumption on the copula linking the two variables, there is no assumption made on the marginal distributions and it can therefore be applied to any continuous valued data. Within neuroimaging, we propose use of this estimator as a statistical test to be employed with permutation based approaches that include correction for the problem of multiple comparisons (see Section 4). We therefore do not explicitly consider here the tightness of the lower bound, or the bias of our estimator. The lower bound property protects against false positive findings.

3.2. Estimating MI between a discrete and a continuous variable using a copula based approach

In many cases, the statistical inference of interest concerns a continuous valued neuroimaging signal recorded in response to a number of different discrete stimuli or under different experimental conditions. Here we extend the copula-based MI estimate introduced above to this problem – estimating MI between a (univariate) discrete and a (potentially multivariate) continuous variable. Despite the wide applicability of such a measure, the practical issue of computing MI between discrete and continuous variables has so far received little attention (Lefakis and Fleuret, 2014; Magri et al., 2009; Ross, 2014). One approach would be to discretize the continuous variable as described in Section 2.3 and use standard binned methods. However as discussed previously this suffers from the curse of dimensionality when considering multivariate spaces. Instead we develop an approach based on the Gaussian copula estimator we described above.

Given the lack of explicit parametric distributions over mixed continuous and discrete spaces it is convenient to start from the second form of Eq. (1). With *X* as the (possibly multivariate) continuous neuroimaging response and *Y* as the discrete stimulus:

$$I(X;Y) = H(X) - \sum_{y} P(y) H(X \mid Y = y) \tag{3}$$

It is straightforward to apply Gaussian parametric entropy estimation to the conditional entropy terms, using the samples available for each *y* value. However, when considering the unconditional entropy term *H(X)* there are two possible approaches. One option is to form the mixture model from the class conditional

parametric fits and numerically integrate this to obtain the entropy (since there is no closed form expression for the entropy of a Gaussian mixture). A second option is to fit a separate parametric model of the same form to the full data set and use that to estimate the unconditional entropy. Figure 9 illustrates these issues with an example of data generated under a Gaussian model with different means for each of two discrete stimulus conditions (Figure 9A). Figure 9B illustrates the two different models that can be used to estimate the unconditional entropy term, the actual mixture density (solid line) or a Gaussian fit (dotted line). Whichever strategy is used, MI is obtained as the difference between the entropy of the chosen unconditional model, and the average of the entropies of the class conditional distributions (Figure 9C). It is not clear if either of these is more appropriate than the other – each preserves a different interpretation of MI. The first leads to an MI estimate more consistent with the Kullback-Leibler divergence expression of MI; it is the property of a single distribution (the fitted parametric mixture distribution) and measures the deviation of that distribution from a surrogate independent model. The second leads to an MI estimate more consistent with a statistical testing viewpoint, comparing an unconditional Gaussian fit to class-conditional Gaussian fits (c.f. ANOVA).

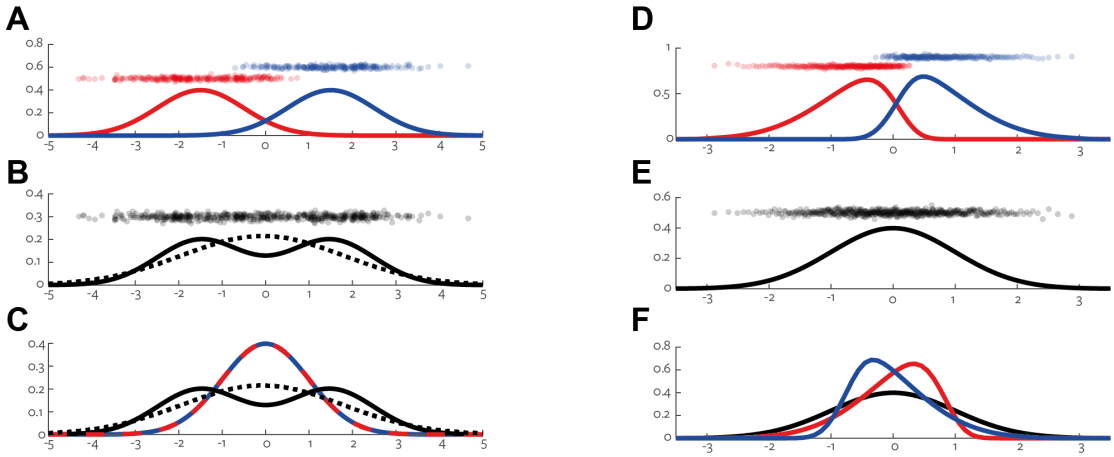

Figure 9 *Illustration of continuous-discrete MI calculation*. Left hand panels show data generated from a Gaussian mixture model with two classes. **A**. Sampled data points and PDF for each class. **B**. Sampled data points, unconditional PDF of the data (solid line), and maximum likelihood Gaussian fit (dotted line). **C**. True unconditional PDF (solid line), unconditional Gaussian fit (dotted line) and de-meaned class-conditional PDF's (red/blue). Right hand panels show the same data after copula Gaussian transformation. **D**. Transformed sampled data points and PDF (kernel density estimate) for each class. **E**. Transformed sampled data points and unconditional probability density (solid line). **F**. Unconditional PDF (black) and class-conditional PDF's (red/blue).

Fortunately, the use of the copula transform removes this dichotomy. Motivated by the previous section we apply the same copula transform to the unconditional data to obtain a surrogate data set with a standard normal distribution (Figure 9E), while preserving the rank-class relationships. Figure 9D shows the resulting class conditional distributions. Again, MI is calculated as the difference between the entropy of the unconditional entropy and the average entropy of the class conditional distributions (Figure 9F; Eq. (3)). Now, by design

the true unconditional distribution is Gaussian, so the unconditional Gaussian fit and the class-conditional mixture are equivalent.

It is clear that after the transformation, the class-conditional distributions are no longer Gaussian (Figure 9D), so our Gaussian conditional entropy estimate will be an approximation. However, again the maximum entropy property works in our favor: each class conditional Gaussian entropy estimate will necessarily be greater than or equal to the true entropy of the class. Since the conditional entropy terms are subtracted in Eq. (3) this ensures the estimated MI is again a lower bound on the true MI. As in the continuous case, this method can be applied whatever the original distribution of the data; it does not require Gaussian classes as in the example shown here. The key feature is that the copula transform preserves the rank-class relationships, and results in a data set to which the parametric Gaussian entropy estimates can be applied. Note that while the unconditional entropy $H(X)$ itself is not invariant to the copula normalization transform, as in the continuous case, the mutual information, as a difference of entropies, is invariant to marginal transformation.

3.3. Estimating MI in spectral data: phase and power

In many applications, analyzing M/EEG signals in the frequency domain is of particular interest because of the potential for understanding brain oscillations, which are increasingly thought to underlie many important cognitive processes (Schnitzler and Gross, 2005; Singer, 2013; Thut et al., 2012; Wang, 2010). The methodological issues surrounding how best to extract a frequency-based representation from M/EEG data, and perform statistical analysis on such data to determine the presence and nature of stimulus modulations have therefore received much attention (Gross, 2014). In this section, we emphasize how our new Gaussian copula MI estimate can be applied to spectral data. In fact, the approach described here can be applied to any vector quantity (e.g. magnetic field vectors) to separate stimulus modulations of amplitude and direction.

The most commonly employed frequency or time-frequency decompositions result in a complex valued spectral signal in each frequency band (Gross, 2014). The real and imaginary parts of this complex signal can be treated as a 2D response variable within the Gaussian-copula MI framework, where each is transformed separately prior to parametric MI estimation. This can be applied with either the continuous-discrete or continuous-continuous MI as described previously, to allow quantification of categorical experimental differences, as well as encoding of continuous valued stimulus features.

There is often additional interest in characterizing more specifically which aspects of the oscillatory activity – amplitude (the size of the oscillations) or phase (the temporal alignment of the oscillations) – are modulated by experimental conditions or activity in other brain signals (frequency bands or regions). Binned MI measured have already been applied to these sorts of problems (Belitski et al., 2010; Gross et al., 2013; Kayser et al., 2009; Schyns et al., 2011; Szymanski et al., 2011), and a number of other techniques have also been

developed (Kempter et al., 2012; Lachaux et al., 1999; Voytek et al., 2013). In such analyses, it is unclear how best to model or bin the data, because phase, the angle of the complex spectral signal, is a circular variable which 'wraps around' and has no clear ranking or extremal values (Berens, 2009; Lee, 2010). One example is whether the arbitrary cut-off in the numerical representation used should be taken as a bin edge (usually angles are returned in the range [-π, π] but this is implementation dependent).

With the Gaussian copula MI estimator, we suggest extracting phase and amplitude of a complex spectral signal as follows (Figure 10). We obtain amplitude in the normal way as the absolute value of the complex number. This is then transformed as a 1D variable for use with a Gaussian-copula MI estimate (Figure 10; center amplitude plots). Often the square of this amplitude is used and referred to as power, but we note that within the copula framework MI is a rank statistic and since amplitude is always positive and the square operation is monotonic, the choice of power or amplitude does not affect the MI value. To isolate phase, we normalize the complex number by its amplitude, resulting in a 2D variable where all points lie on the unit circle. We then apply the 2D Gaussian copula MI estimate on this 2D variable; transforming each dimension independently (Figure 10, right hand phase plots). Maintaining a 2D representation for phase avoids the technical issues surrounding circular variables, particularly modeling joint distributions of circular and linear variables. The Data Processing Inequality ensures that since only information about phase goes into the calculation (all amplitude variation is removed), whatever processing we apply (here copula transformation) cannot add information and so our estimate of MI carried by phase is valid. Figure 10 illustrates the approach with two simulated systems, one in which only phase is modulated by a discrete experimental condition (Figure 10A), and one in which only power is modulated (Figure 10B). The bar graphs of MI in the different signal representations (Figure 10; far right) demonstrate that this approach can correctly dissociate the two types of modulation.

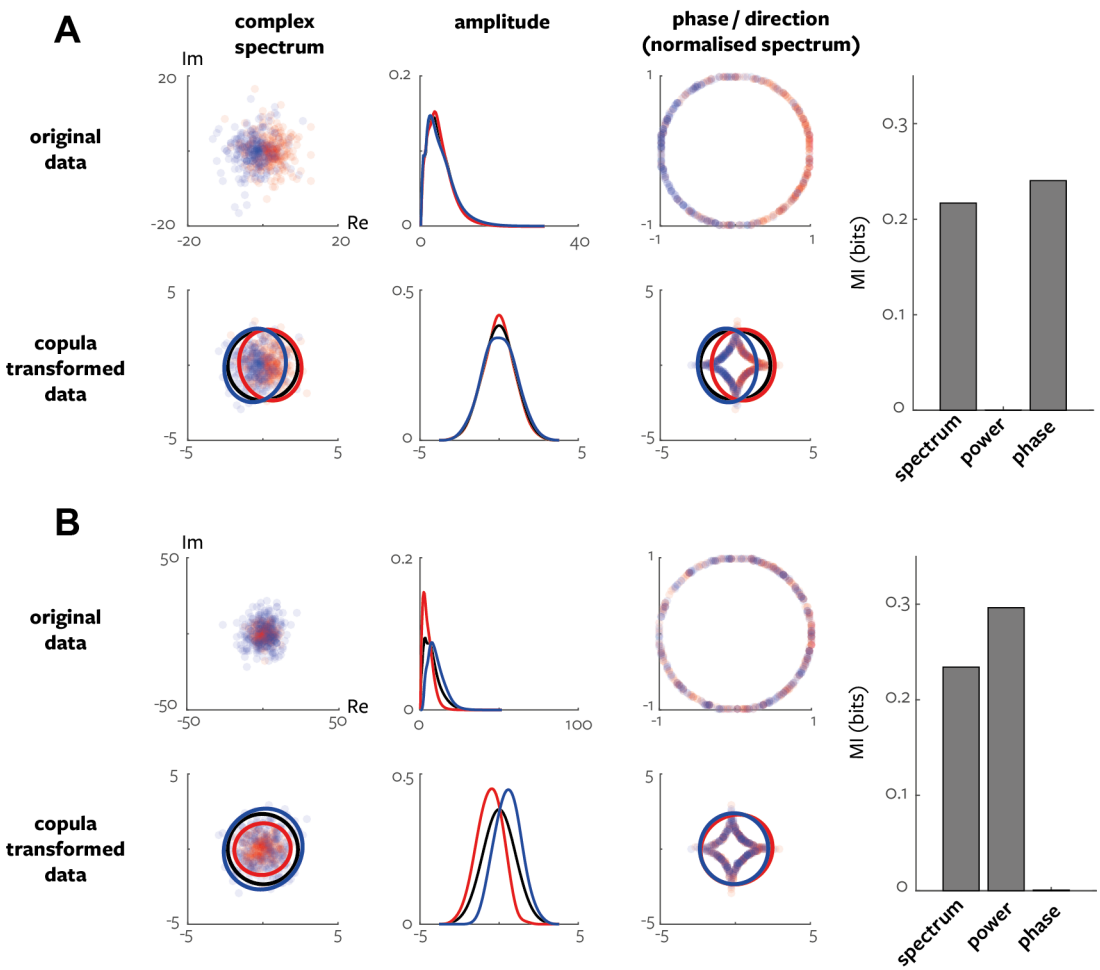

Figure 10 *Gaussian-copula MI applied to complex spectral data.* Spectral data were generated from two two-class models. **A**. Phase was sampled from von Mises distribution with class specific mean and amplitude was sampled from chi-square distribution (common across classes). **B**. Phase was drawn from von Mises distribution (common across classes) and amplitude was sampled from chi-square distribution with class-specific degrees of freedom. **A,B**. Left plots show generated complex data (top) and with marginal copula transformation (bottom). Solid lines show p=0.01 contours of multivariate Gaussian pdf. Centre plots show amplitude (top) and copula transformed amplitude (bottom). Right plots show amplitude-normalized spectrum (top) and copula transformed normalized spectrum (bottom). Far right bar graphs show the MI value in the different data representations.

We emphasize that our approach is equally applicable to the single-trial outputs of any frequency or time-frequency decomposition including Empirical Mode Decomposition, Hilbert-Huang transform and matching pursuit methods (Gross, 2014). It can also be applied to other vector quantities to determine the relative effects on amplitude and direction, for example planar magnetic field gradients (see Section 4.2).

4. Results

In this section we validate the performance of the new estimator as a statistical test and provide some example applications. The data used for the simulation and examples in this section are available in (Ince et al., 2016).

4.1. Discrete experimental condition with continuous EEG response: Face detection

We performed a number of analyses in order to evaluate the performance of the Gaussian copula continuous-discrete MI estimator (Section 3.2) as a statistical test for neuroimaging applications. We consider EEG data collected from a single subject within an event-related design (Section 2.3), with presentation of two classes of images: faces and noise textures. Data were band-pass filtered between 1-30Hz and the Current Source Density transformation was applied (Rousselet et al., 2014a). We calculated MI independently for each time point and sensor using samples collected from the repeated presentations (Figure 4A). A common approach with neuroimaging studies is to apply a permutation test together with the method of maximum statistics in order to correct for multiple comparisons (Holmes et al., 1996; Nichols and Holmes, 2002). We calculated the two-sample Kolmogorov-Smirnov (KS) test statistic (Massey, 1951) for each time point and sensor from all the available 1000 trials and used this as the ground-truth for comparing the performance of tests with smaller numbers of trials (Figure 11A, top left, color plot). We performed the calculation over all sensors and time points 1000 times, randomly permuting the stimulus class labels each time. We took the maximum value over sensors and time points for each of these permutations, and used the 99th percentile of these maximum values as the threshold for significance (Figure 11A, top left, black and white plot). This procedure corrects for multiple comparisons and provides a Family-Wise Error Rate (FWER) of 0.01. We then repeatedly sub-sampled smaller sets of trials from the full data set and repeated the mass-univariate analysis for various statistics, together with the permutation approach. Figure 11A, middle and bottom show the results for a sub-sampled set of 100 trials. The significant sensors and time points were then compared to the full data KS-test ground truth to evaluate the performance of different statistics.

We considered four types of statistic, the new Gaussian copula MI, binned MI (with 2,4 or 8 bins), the t-test (unequal variances) and the KS-test. Figure 11B shows the time-courses obtained from a single sensor for each of these statistics. First we considered the performance of each of these statistics over the full space of time points and sensors, as a function of the number of trials available. For each sample size (25, 50, 100, 200, 500), we randomly selected that number of trials from the full data set, calculated all the statistics including 200 class-shuffled permutations, and determined the final multiple-comparison corrected inference for each statistic. We repeated this procedure 50 times for each sample size. For each statistic and sample size we then compared the result of the inference to the ground truth, considering sensitivity (Figure 11C, top left), specificity (Figure 11C, middle left) and MI with ground-truth (Figure 11C, bottom left). Sensitivity, or true positive rate, measures the proportion of significant ground-truth responses that are correctly detected by each test. Sensitivity increases with number of samples for all tests, the t-test and copula MI have the highest sensitivity over the full range of trials considered. Specificity, or true negative rate, measures the proportion of non-significant ground-truth responses that are correctly detected

as non-significant by each test. Specificity is high for all tests, due to the strong control on FWER provided by the permutation approach. Binned MI methods have the highest specificity, the t-test has the lowest, with the copula MI taking intermediate values. Finally, as an overall measure of the performance of the test we consider the MI in the contingency table (or confusion matrix) for each test (Quian Quiroga and Panzeri, 2009); i.e. for each repetition, the discrete MI between the binary ground-truth significance and the significance for that repetition, with time points and sensors providing the samples for the MI calculation. Similar in spirit to Matthew's Correlation Coefficient (Baldi et al., 2000; Matthews, 1975), this provides an overall accuracy measure, which incorporates possible asymmetries resulting from unbalanced classes. With this measure, all tests perform better with more samples; the t-test and copula MI stand out as more accurate than the other tests, with the t-test the most accurate.

Next, we fixed the number of trials at 100, and performed a similar analysis to evaluate the robustness of the different statistics to the presence of outliers. On each of 50 repetitions, a certain percentage of trials were chosen (0-50%) and for those trials the EEG signal at each sensor and time point was replaced with a random variable drawn from a Gaussian distribution with standard deviation (s.d.) 5 times greater than the actual s.d. of that response. We again computed sensitivity, specificity and MI against the same undistorted 1000 trial KS-test ground truth (Figure 11C, right). The sensitivity and MI plots illustrate the robustness of the copula MI estimator compared to the t-test. While the t-test has slightly higher sensitivity for the original data, 5% of trials corrupted by outliers is already enough to reduce the sensitivity of the t-test to half that of the copula MI.

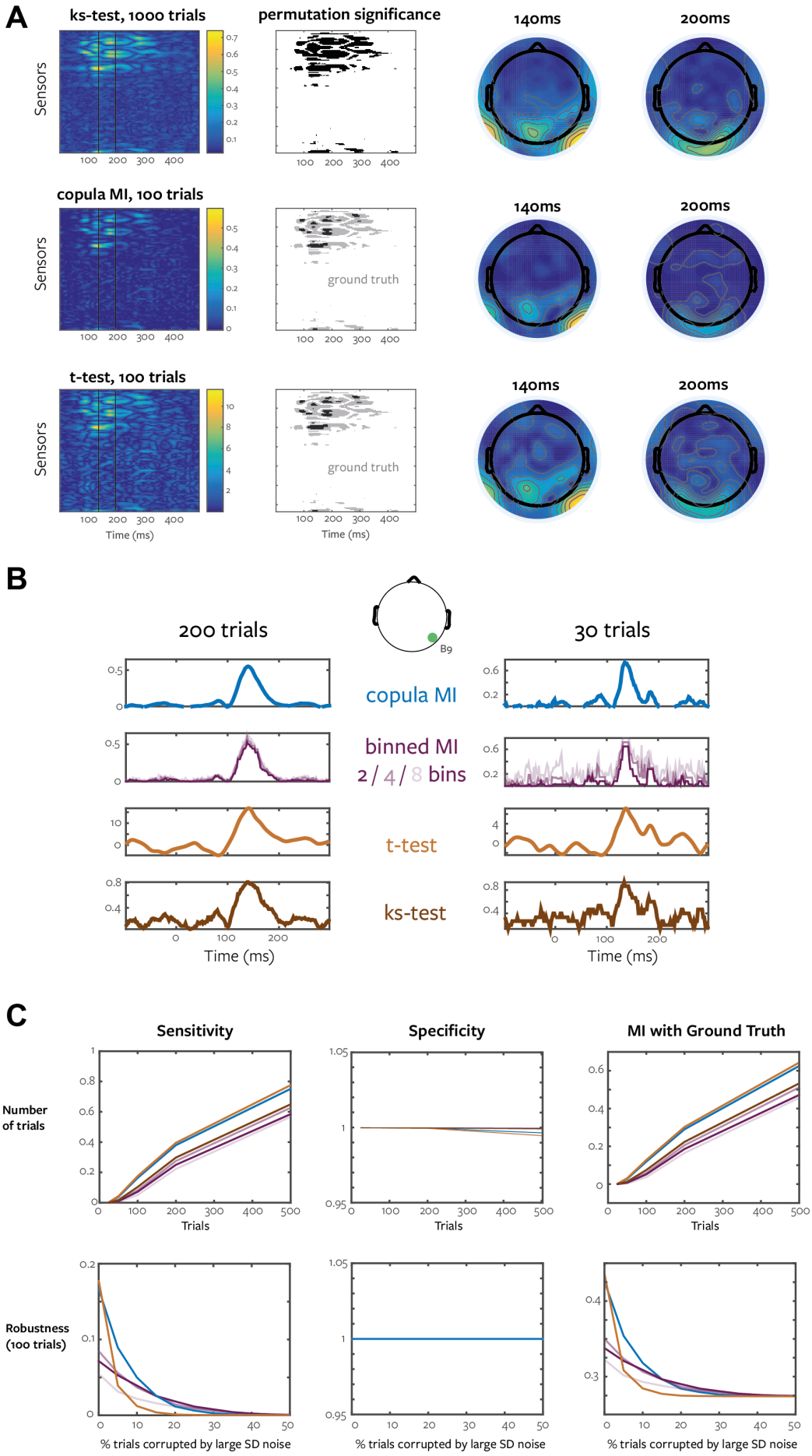
A
ks-test, 1000 trials
permutation significance
140ms
200ms
Sensors
copula MI, 100 trials
ground truth
t-test, 100 trials
Time (ms)
B
200 trials
30 trials
B9
copula MI
binned MI
2 / 4 / 8 bins
t-test
ks-test
Time (ms)
C
Sensitivity
Specificity
MI with Ground Truth
Number of trials
Trials
Robustness (100 trials)
% trials corrupted by large SD noise

Figure 11 *Performance of Gaussian copula MI estimator as a statistical test for EEG data with discrete stimuli in event related design.* **A.** Statistics are calculated for each sensor and time point (left colored image plots) and significance determined with permutation testing and the method of maximum statistics (black and white image plot). Topologies are shown for two indicated time points. The KS-test with all 1000 trials is used as the ground-truth (top); copula MI with 100 trials (middle) and t-test with the same 100 trials (bottom) are shown. The results of permutation significance for these statistics are shown (black) overlaid on the ground-truth significance (grey). **B.** Example time-courses of various effect sizes calculated with 200 (left) or 30 (trials). **C.** Results of numerical investigation of the performance of various statistics with permutation testing, as a function of the amount of data available (left column) and as a function of the amount of noise added to the data (right column).

In summary, when performing mass-univariate analyses with permutation testing and maximum statistics, the copula MI provides similar sensitivity to the more commonly employed t-test, but is considerably more robust to the presence of outliers. It also performs better than binned MI estimates, and can be applied to multi-variate responses.

4.2. Continuously varying experimental condition with continuous MEG response: listening to speech

To evaluate the performance of the Gaussian copula continuous-continuous MI estimator (Section 3.1) we consider MEG data collected from a single subject within a continuous design with an auditory speech stimulus (Gross et al., 2013; Park et al., 2015). For simplicity we focus here on a sensor-level time-domain analysis. Since previous work has shown speech entrainment mostly at lower frequencies, we extracted the wideband amplitude envelope of the speech stimulus (Chandrasekaran et al., 2009; Gross et al., 2013) and then low-pass filtered with a 12 Hz cutoff (3rd order non-causal Butterworth). The MEG signal was obtained from a 248-magnetometer whole-head MEG system (MAGNES 3600 WH, 4-D Neuroimaging). We band-pass filtered in the range 2-12 Hz (3rd order non-causal Butterworth), downsampled to 50Hz and then computed the planar gradient tangential to the head at each magnetometer (Bastiaansen and Knösche, 2000). This potentially simplifies interpretation of sensor-level data because it typically results in maximal signal directly above the corresponding source (Hämäläinen et al., 1993). We analyse 450s of recording (22,500 samples at 50Hz) during which the subject listened to a spoken story. Similar to a cross-correlation of two signals we calculate the relationship between the speech envelope and the MEG signal over a range of delays (0-350ms).

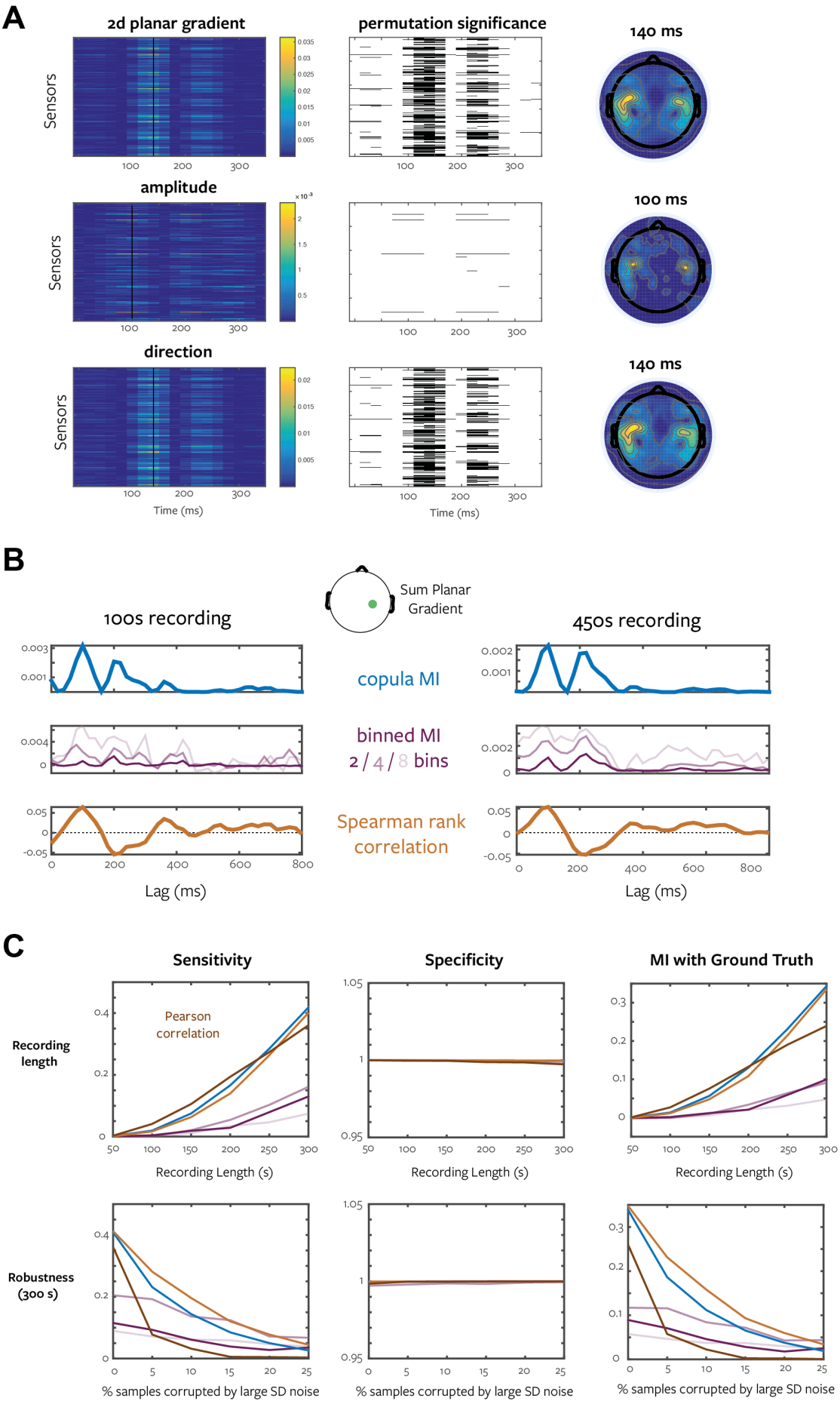
A
2d planar gradient
permutation significance
140 ms
Sensors
amplitude
100 ms
direction
140 ms
Time (ms)
B
Sum Planar Gradient
100s recording
450s recording
copula MI
binned MI
2 / 4 / 8 bins
Spearman rank correlation
Lag (ms)
C
Sensitivity
Specificity
MI with Ground Truth
Recording length
Pearson correlation
Recording Length (s)
Robustness (300 s)
% samples corrupted by large SD noise

Figure 12 *Performance of Gaussian copula MI estimator as a statistical test for MEG data with continuous stimuli in a continuous design.* **A.** Gaussian copula MI is calculated between the speech envelope and the full 2D planar gradient response (top), the amplitude (middle) or the direction (bottom) of the planar gradient vector, for each sensor and time point (left colored image plots). Significance is determined with block permutation testing and the method of maximum statistics (black and white image plots). Topologies are shown for the indicated time points. **B.** Example cross-correlation style delay plots of various effect sizes calculated with 100s or 450s of continuous stimulation. **C.** Results of numerical investigation of the performance of various statistics with block permutation testing, as a function of the amount of data available (left column) and as a function of the amount of noise added to the data (right column).

The resulting planar gradient signal consists of a 2d magnetic field vector (tangential to the head) at the position of each magnetometer. Typically, the amplitude of this 2d vector is used as the response signal of interest (Oostenveld et al., 2011). However, using the multivariate MI estimate we can quantify the modulation of the full 2d signal, as well as breaking down the stimulus effects on amplitude and direction separately (Figure 12A), as described for spectral data in Section 3.3. The top row of Figure 12A shows the MI between the speech envelope and the 2d planar gradient for each channel and speech-MEG delay lag. The black and white image plot shows the multiple comparison corrected permutation significance (p=0.01, 200 permutations, 10s block permutation scheme to preserve signal autocorrelation). The middle row shows the same for the 1d Pythagorean amplitude (Fieldtrip default), and the bottom row shows the MI in the planar gradient direction only, with the amplitude effects normalised out as described for phase in Section 3.3. This shows that, while there is a focal and statistically significant modulation of the planar gradient amplitude over the auditory cortices, in fact the direction of the planar gradient is modulated much more strongly by the speech envelope over a much wider area, with MI values an order of magnitude higher. In addition, the timing of the peak effect is different (earlier for the amplitude). This suggests that focusing on the amplitude of time-varying magnetic field vectors could result in reduced sensitivity, and provides an example of the potential advantage of using multivariate statistics that allow separate treatment of the direction and amplitude of vector values.

However, to investigate the statistical properties of the copula MI estimator we use the 1d amplitude signal, as there are not so many well-established statistical methods to compare for evaluating the relationships in the continuous multivariate case. Figure 12B shows the delay time courses for the sensor with the strongest amplitude modulation. To determine the performance of the copula MI as a statistical test we proceed as described in the previous section. First, we obtained ground-truth significance by applying Spearman's rank correlation between the speech envelope and the lagged MEG planar gradient amplitude using the full 450s of available data (1000 permutations, 10s block permutation scheme, maximum statistics corrected over all sensors and 18 delays considered). Then we subsampled (using the 10s block scheme) reduced amounts of data (50-300s, repeated 30 times each). For each subsampled repetition we computed a range of statistics (copula MI, binned MI, Pearson correlation, Spearman correlation) and determined the significance of each with block permutation and maximum statistics (p=0.05, 100 permutations, 10s blocks) (Figure 12C, left column). Sensitivity, specificity and MI with ground-truth show that the copula

method performs similarly to rank correlation (slightly lower sensitivity, slightly higher MI with ground-truth).

To investigate the robustness of the statistics we fixed the amount of data at 300s and investigated the effect of corrupting a fixed percentage of trials with Gaussian noise with standard deviation 5 times larger than that of the MEG signal (Figure 12C, right column). Here Spearman rank correlation performs best (measured both with sensitivity and MI), but the copula MI is close. Pearson correlation is very sensitive the addition of outliers with sensitivity reduced to ~30% of that of the copula MI with only 5% corruption. The use of Spearman's correlation to define the ground-truth, combined with the low number of significant responses may result in a bias towards the particular property of that measure.

In summary, when performing mass-univariate analyses with permutation testing and maximum statistics, the continuous copula MI provides similar performance as Spearman rank correlation. However, as an MI estimate it benefits from the useful properties of the MI effect size (e.g. additivity; see below), and we have shown that the ability to consider multivariate responses and separately quantify modulations of vector direction and amplitude have the potential to provide more detailed interpretations of MEG data.

4.3. Pairwise temporal interactions reveal modulation of gradient in EEG

To provide an example application of interaction information (Section 2.6; Figure 7) we consider temporal interactions within an event-related EEG experiment with a continuous valued stimulus feature. The task was face detection, with stimuli as in Figure 4A, but here the images were sampled with Bubbles: randomly positioned Gaussian apertures which selectively reveal different parts of the image on different trials (Gosselin and Schyns, 2001). Since it has been shown that in this paradigm it is primarily the visibility of the eye region which modulates the recorded EEG (Rousselet et al., 2014a), we focus here on the visibility of the left eye region (a continuous scalar value for each trial obtained by summing the bubble masks within the eye region) and its effects on a contra-lateral right occipital-temporal electrode. We consider 1092 trials from a single observer during which a bubbled face image was presented (Rousselet et al., 2014b). EEG data were band-pass filtered (1-30Hz) and the Current Source Density transformation was applied (Rousselet et al., 2014a).

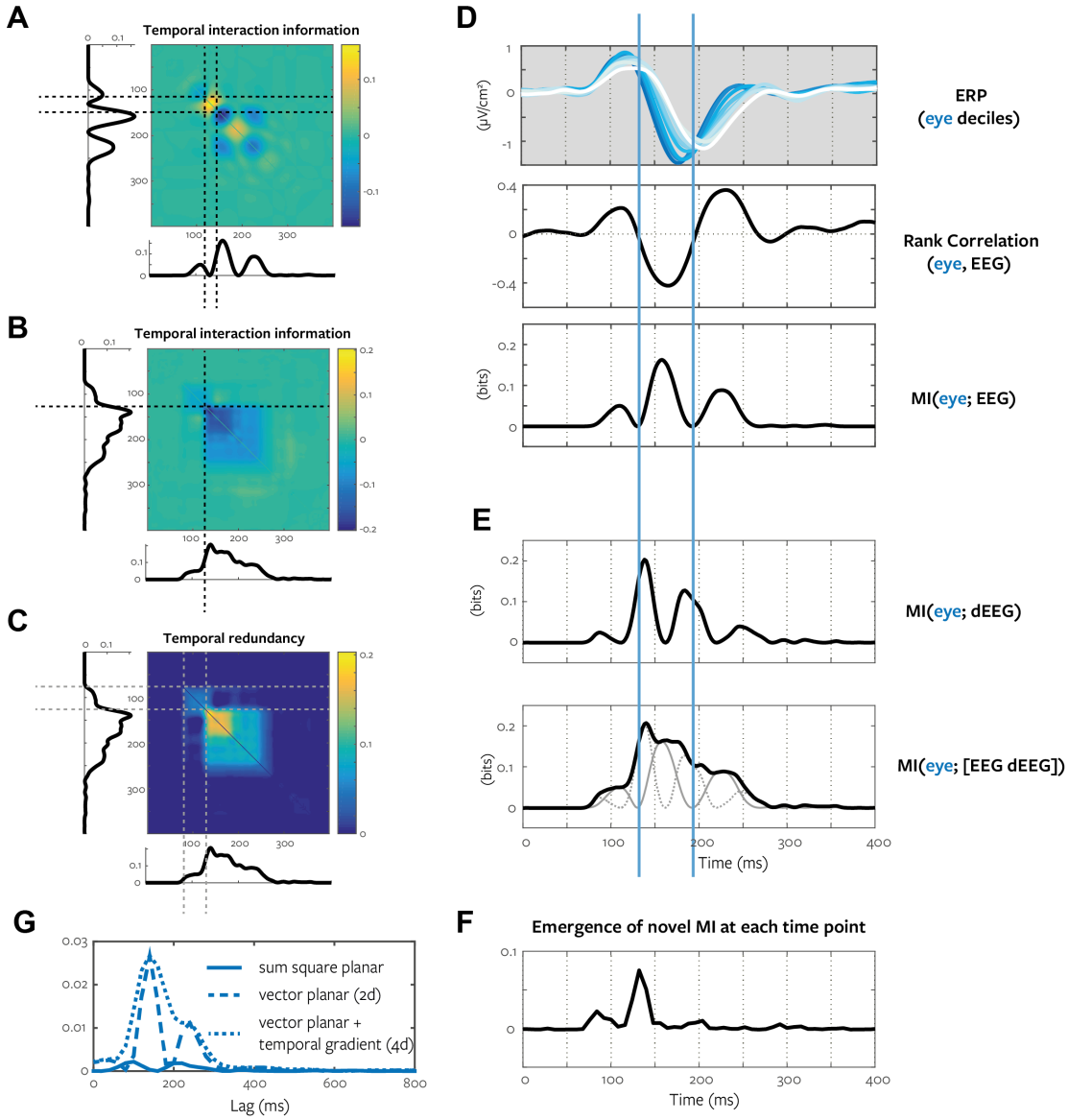

Figure 13 *Temporal interaction information reveals modulation of gradient*. **A.** Interaction information between EEG voltage at pairs of time points. Positive values correspond to synergy, negative values indicate redundancy. **B.** Interaction information between bivariate EEG voltage and temporal derivative at pairs of time points. **C.** As B, but only redundancy is shown. **D**. The mean ERP was calculated separately for each decile of the stimulus feature (white to blue increasing eye visibility). Spearman correlation and MI are calculated for the EEG voltage at each time point. **E.** The MI time course is calculated using the temporal derivative (upper) and a bivariate response consisting of the EEG voltage and temporal derivative at each time point (lower). This bivariate MI time course (black) is shown with the MI time courses of the constituent variables (EEG voltage, solid gray; temporal derivative, dotted gray lines). **F**. We down-sampled the data to 125 Hz, and calculated the new MI arriving at each time point (see text). **G**. Effect of including the temporal derivative with the 2d planar gradient response from Section 3.2, Figure 10.

Figure 13D shows how the modulation of the evoked time-course by the stimulus feature (eye visibility), and how this can be quantified by calculating Spearman's rank correlation or copula MI independently at each post-stimulus time point (Figure 4A). However, with MI we can calculate the Interaction Information (II) between pairs of time points (Figure 13A). This allows us to investigate the relationship between the modulation of the evoked signal at different times: positive II indicates a synergistic relationship between the responses; negative II means they are redundant. Here, the MI curve has 3 peaks;

the temporal interaction matrix reveals that the second and third peaks are mutually redundant, but the first peak appears to carry independent MI. Interestingly, there are striking patches of local synergy (indicated with dashed lines), equivalent in magnitude to the largest MI values over the time course and corresponding to time points where there is no MI in the raw EEG voltage. This indicates that in those regions, even though observing the recorded voltage at a single time point does not reveal anything about the value of the stimulus feature the relationship between nearby time points is highly informative.

The simplest quantification of the relationship between neighboring time points is the temporal gradient. To determine if this could account for the observed synergy we calculated the central difference temporal gradient of the EEG voltage for each trial and considered the MI in this response (Figure 13E, upper). Peaks in the gradient MI occur concurrently with zero points of the raw voltage MI. To incorporate the modulation of both response representations we combine them in a bivariate response consisting of the EEG voltage and the temporal gradient at each time point. We calculate the time course of MI about the stimulus feature in this bivariate response (Figure 13E, lower). Relating this to the voltage MI time course and the actual ERP modulation (Figure 13D) we can see that the zero points of the triple-peak MI profile result from zero-crossings where the sign of the correlation changes, due to the shape of the modulated bimodal ERP. However, by considering the conditional ERPs it is clear that these points fall within the time window where the overall shape of the evoked EEG response is modulated by the stimulus feature. We therefore take the view that considering the gradient together with the voltage (Figure 13E) provides a substantial advantage: these artifactual dips are smoothed out, providing a clearer picture of the time window over which the EEG signal is modulated by the changing stimulus.

We suggest that including the gradient of recorded neuroimaging signals could be a useful principle across a range of different analyses. For example, returning to the MEG data set under continuous speech stimulation (Section 4.2), including the temporal derivative of each planar gradient component (resulting in a 4d response) has the same effect of smoothing out the artifactual MI zero resulting from a change of sign in the effect (Figure 13G). Again, this gives a clearer, smoother picture of the range of delays over which the amplitude of the speech envelope modulates the MEG signal, which is made possible with the use of a multivariate statistic.

We can repeat the temporal interaction analysis on our bivariate responses (Figure 13B). This reveals that there is now little synergy, the main MI peak is mostly self-redundant, but the early part of the MI time-course appears to be independent from the main peak. To give a clearer picture we show the redundancy only (negative II; Figure 13C). A block structure is clearly apparent; the early MI appears to be independent from the bulk of the later MI (indicated with dashed lines). This suggests a functional differentiation between the initial P100, and the later N170: they appear to be modulated by the eye visibility in different ways and so possibly reflect different processing pathways. This would not be

apparent from inspecting the ERP modulation (Figure 13D) or the MI time course (Figure 13E) alone.

Another application of the copula MI framework allows us to directly quantify the emergence of novel MI over time. For each time point, we calculate the MI about the stimulus feature available in the (bivariate) EEG response at that time point. We then subtract the MI that is redundant with that at the previous time point, leaving only the amount of new MI about the stimulus arriving at that time. Mathematically, this is equivalent to calculating the CMI $I\left(\text{eye}; R_{t_i} \mid R_{t_i-1}\right)$. Figure 13F shows the result of this analysis. Two peaks of novel MI are clearly visible. In this analysis, for later time points the response at the time of the two peaks are also conditioned out to ensure only genuinely new MI is measured. The first peak corresponds to the early P100 modulation, the second to the stronger N170 modulation. This analysis corroborates the temporal interaction presented above, revealing that there appear to be two separate processes modulated by the stimulus feature – one beginning at 84 ms (P100), and one beginning at 132 ms (N170).

In summary, we have shown a few illustrative examples of the application of pairwise interaction information, focusing on interactions in the temporal domain with EEG data. We have shown how viewing pairwise interactions in terms of synergy and redundancy about a stimulus feature can provide useful insights – for example by revealing the importance of considering the temporal derivative when evaluating stimulus modulation of an evoked signal, and allowing us to directly quantify the emergence of new information over time. We emphasize that this approach is completely general and can be used across wide range of different responses (Figure 7A).

5. Discussion

Information theory provides a principled methodology for studying and quantifying statistical relationships between variables. As we have reviewed, the foundational quantities of information theory are entropy and mutual information. Here we have presented a novel approach to the practical estimation of these quantities, combining the statistical theory of copulas with the closed form solution for the entropy of Gaussian variables. This provides a computationally efficient and statistically robust lower bound estimate to MI with no specific assumptions on the marginal distribution of each variable. We have validated the use of this estimate as a statistical test within a neuroimaging context, considering both discrete and continuous experimental stimuli, and have shown that with 1D responses it performs as well as existing commonly used statistics. To accompany this article, we have released open-source code implementing the new methods for both Matlab and Python programming languages[13], together with tutorial examples covering the analyses presented here. The major advantage of our method over traditional statistical approaches is that it unifies a variety of applications (continuous, discrete and multidimensional variables) in a framework with a common effect size, and quantities like CMI and interaction information allow novel interpretations that are not available with other approaches.

Application to multi-dimensional spaces

A particular advantage of the proposed method is the ability to reliably estimate MI and other information theoretic quantities on multi-dimensional spaces. We suggest that there are many situations in neuroimaging where multivariate responses are interesting, but difficult to address with existing statistical methods. Our examples included considering complex MEG spectra (as well as separating effects on phase and amplitude), 2D planar magnetic field gradients and considering raw signal values together with the instantaneous temporal derivative. Similarly, we could consider 3D magnetic fields arising from MEG source localization techniques, higher order temporal derivatives or features describing single-trial ERP features (peak and latency) (Hu et al., 2010). The multivariate performance is also important for calculating higher order information theoretic quantities such as conditional mutual information and interaction information. This is challenging with existing methods due to the curse of dimensionality, which either results in excessive data requirements (binned methods) or high computational complexity (continuous methods), even for modest numbers of dimensions (i.e. pairwise interaction information on 2D response variables as in the example of Section 4.3 requires estimation of entropy over a 5D space). While a more thorough analysis of the data requirements of the proposed method is an important area for future work, our experience to date is

[13] See the "Gaussian-Copula Mutual Information (GCMI)" project at: https://github.com/robince/gcmi

Code for all simulations and figures is available at: https://github.com/robince/sensorcop

that with amounts of data that can reasonably be collected in a suitably designed neuroimaging experiment (i.e. hundreds of trials) spaces of dimension 5-10 can reliably be addressed, although of course this depends on the strength of the underlying effects. For much greater numbers of dimensions estimation of the required covariance matrices becomes problematic, even given the improved robustness resulting from the copula rank transformation.

The application of regularization through Bayesian priors or other means might provide a way to extend the measure to even higher dimensional spaces. Alternatively, we propose that one way to extend our estimator to very high-dimensional spaces is to combine it with decoding approaches based on supervised learning algorithms. For example, by first using a decoding algorithm (e.g. a linear discriminant) as a dimensionality reduction step, we can calculate the MI of the low-dimensional predictor signal, within a cross-validation framework. We suggest that MI has some advantages as a statistic to evaluate the performance of a decoder (Quian Quiroga and Panzeri, 2009) compared to commonly used measures (such as mean performance, area under ROC curve, etc.). Again it uses a common scale, provides the ability to relate MI in different signals (e.g. between EEG sensor array linear discriminant output and single-trial fMRI voxel beta activations, see below) and allows us to condition out correlated features.

Second level analyses on MI values

MI provides a high contrast statistic that can be used as input for second level analyses. MI reveals a functional property of the system that might change with different experimental conditions, for example with the rhythmic structure of speech stimuli (Kayser et al., 2015) or spatial attention (Guggenmos et al., 2015; Saproo and Serences, 2010). Functional connectivity, measured with directed information, has been shown to be affected by the intelligibility of speech (Park et al., 2015). When considering MI computed in different experimental conditions accurate bias correction is important because bias may not be equal in each condition (for example due to differing numbers of samples, or different degrees of signal autocorrelation). Mass-univariate MI calculations can also provide a rich and descriptive input for subsequent dimensionality reduction. For example Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999) is a dimensionality reduction technique that is well matched for application to MI results (since they are non-negative, and the high signal-to-noise contrast of MI complements the mean squared-error objective of the NMF algorithm). This can be used to extract task relevant features from a high-dimensional naturalistic stimulus (Ince et al., 2015); the same approach could also be applied to reduce the dimensionality of a high-dimensional neuroimaging response (e.g. an EEG sensor array) into specific spatio-temporal task or stimulus MI components.

Quantifying pairwise interactions

We believe that understanding brain function from brain activity requires a focus on the particular information processing functions performed under different tasks and conditions (Kriegeskorte et al., 2006; Schyns et al., 2009). To

fully exploit the potential of this information processing perspective requires methods that not only identify the systematic modulations of brain responses, but the relationship between such modulations, or representations, across different times, regions and signals (Figure 7A). The use of such approaches within neuroimaging is growing, with development and application of techniques such as Representation Similarity Analysis (RSA) (Kriegeskorte and Kievit, 2013), and the use of supervised classification algorithms together with cross-validation (Hastie et al., 2001), often referred to as a decoding (Haxby et al., 2014; King and Dehaene, 2014; Quian Quiroga and Panzeri, 2009). Our new MI estimator allows us to address these issues within the unified framework of information theory. Particular advantages include the common, meaningful scale that allows direct comparison of redundancy with the MI in different signals, experiments, or behavior. With information theory we can normalize redundancy as a percentage, which provides a more intuitive measure for the degree of overlap than is available with other methods, and we can perform all analyses conditioning out multiple correlated stimulus features if necessary. Detailed comparison of our information theoretic framework with methods such as RSA will be the subject of future work.

Our novel estimator can be applied to calculate measures of functional connectivity such as directed information (transfer entropy). In combination with the information perspective described above, by adding the concept of redundancy within the Granger causal framework we have developed a measure of functional connectivity that quantifies the communication of specific information content (Ince et al., 2015). The copula MI estimate is crucial to allow practical computation of this measure, which requires conditioning directed information on the particular stimulus features considered. This measure allows for dynamic network analyses that are based not on general relationships between areas, but on communication of specific information about the stimulus or task.

Broader applications

We have focused here on application to M/EEG data, but we emphasize that the copula MI estimator can be applied to any signal, facilitating the comparative study of neural information coding across experimental methodologies and scales of brain measurement (Panzeri et al., 2015). For example, it could be applied to single trial fMRI GLM beta activations directly, or in combination with a multi-voxel decoding approach as described above. The common scale allows direct comparison of the strength of the modulation between different neuroimaging responses, and interaction information opens up the promising possibility of directly relating the information content in different signals (Cichy et al., 2014). For example, calculating the redundancy over time between MI in a (multivariate) EEG response and individual voxel single-trial beta activations, would allow mapping the spatial region that is redundant with the EEG at each time point.

Similarly, while we have focused here on neuroimaging, MI has broad applicability as a general statistical framework. It can be used for analyzing

behavioral data, where many of the properties we have highlighted could be useful (e.g. CMI, interactions). It has been used for feature selection in general classification problems (Lefakis and Fleuret, 2014; Peng et al., 2005; Torkkola, 2003) and we hope our estimator would also provide practical advantages in many such applications. We further suggest that the copula normalization could be used as a general preprocessing step that would convert any covariance based statistic or algorithm into a robust rank-based version (e.g. common spatial patterns, canonical correlation analysis, linear/quadratic discriminant analysis).

Conclusion

We have presented a novel approach to estimate MI and associated quantities. This provides a general, computationally efficient, flexible, robust, multivariate statistical framework based on information theory. This framework provides effect sizes on a common meaningful scale and allows for unified treatment of discrete and continuous variables. Beyond measuring the strength of direct, possibly multivariate, relationships, quantities like CMI and interaction information have the potential to provide transformative interpretations of neuroimaging data, for example by relating information content in different brain responses. This framework allows investigators to take full advantage of the properties of each neuroimaging signal and their experimental designs to develop a better understanding of the information processing functions of brain networks.

Acknowledgements

We thank Arno Onken, Daniel Chicharro and Stefano Panzeri for useful discussions. JG and PGS are supported by the Wellcome Trust [098433, 107802]. GAR, PGS, BG and CK are supported by the UK Biotechnology and Biological Sciences Research Council (BBSRC) [BB/J018929/1, BB/D01400X1, BB/M009742/1, BB/L027534/1]. CK is supported by the European Research Council (ERC-2014-CoG; grant No 646657).

References

Abásolo, D., Hornero, R., Espino, P., Álvarez, D., Poza, J., 2006. Entropy analysis of the EEG background activity in Alzheimer's disease patients. Physiol. Meas. 27, 241. doi:10.1088/0967-3334/27/3/003

Adolf, D., Baecke, S., Kahle, W., Bernarding, J., Kropf, S., 2011. Applying multivariate techniques to high-dimensional temporally correlated fMRI data. J. Stat. Plan. Inference 141, 3760–3770. doi:10.1016/j.jspi.2011.06.012

Akaike, H., 1992. Information Theory and an Extension of the Maximum Likelihood Principle, in: Kotz, S., Johnson, N.L. (Eds.), Breakthroughs in Statistics, Springer Series in Statistics. Springer New York, pp. 610–624.

Archer, E., Park, I., Pillow, J., 2013. Bayesian and Quasi-Bayesian Estimators for Mutual Information from Discrete Data. Entropy 15, 1738–1755. doi:10.3390/e15051738

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424. doi:10.1093/bioinformatics/16.5.412

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

Bastiaansen, M.C.M., Knösche, T.R., 2000. Tangential derivative mapping of axial MEG applied to event-related desynchronization research. Clin. Neurophysiol. 111, 1300–1305. doi:10.1016/S1388-2457(00)00272-8

Belitski, A., Panzeri, S., Magri, C., Logothetis, N.K., Kayser, C., 2010. Sensory information in local field potentials and spikes from visual and auditory cortices: time scales and frequency bands. J. Comput. Neurosci. 29, 533–545. doi:10.1007/s10827-010-0230-y

Berens, P., 2009. CircStat: A MATLAB Toolbox for Circular Statistics. J. Stat. Softw. 31.

Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., Ay, N., 2014. Quantifying Unique Information. Entropy 16, 2161–2183. doi:10.3390/e16042161

Borst, A., Theunissen, F.E., 1999. Information theory and neural coding. Nat. Neurosci. 2, 947–957.

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

Caballero-Gaudes, C., Van de Ville, D., Grouiller, F., Thornton, R., Lemieux, L., Seeck, M., Lazeyras, F., Vulliemoz, S., 2013. Mapping interictal epileptic discharges using mutual information between concurrent EEG and fMRI. NeuroImage 68, 248–262. doi:10.1016/j.neuroimage.2012.12.011

Calsaverini, R.S., Vicente, R., 2009. An information-theoretic approach to statistical dependence: Copula information. EPL Europhys. Lett. 88, 68003. doi:10.1209/0295-5075/88/68003

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A.A., 2009. The Natural Statistics of Audiovisual Speech. PLoS Comput Biol 5, e1000436. doi:10.1371/journal.pcbi.1000436

Chicharro, D., 2014. A Causal Perspective on the Analysis of Signal and Noise Correlations and Their Role in Population Coding. Neural Comput. 1–56. doi:10.1162/NECO_a_00588

Chicharro, D., Ledberg, A., 2012. When Two Become One: The Limits of Causality Analysis of Brain Dynamics. PLoS ONE 7, e32466. doi:10.1371/journal.pone.0032466

Chicharro, D., Panzeri, S., 2014. Algorithms of causal inference for the analysis of effective connectivity among brain regions. Front. Neuroinformatics 8. doi:10.3389/fninf.2014.00064

Cichy, R.M., Pantazis, D., Oliva, A., 2014. Resolving human object recognition in space and time. Nat. Neurosci. 17, 455–462. doi:10.1038/nn.3635

Cover, T.M., Thomas, J.A., 1991. Elements of information theory. Wiley New York.

Darbellay, G.A., Vajda, I., 1999. Estimation of the information by an adaptive partitioning of the observation space. Inf. Theory IEEE Trans. On 45, 1315–1321.

Eckhorn, R., Pöpel, B., 1974. Rigorous and extended application of information theory to the afferent visual system of the cat. I. Basic concepts. Kybernetik 16, 191–200. doi:10.1007/BF00288979

Endres, D., Foldiak, P., 2005. Bayesian bin distribution inference and mutual information. IEEE Trans. Inf. Theory 51, 3766–3779. doi:10.1109/TIT.2005.856954

Fairhall, A., Shea-Brown, E., Barreiro, A., 2012. Information theoretic approaches to understanding circuit function. Curr. Opin. Neurobiol. 22, 1–7. doi:10.1016/j.conb.2012.06.005

Faivishevsky, L., Goldberger, J., 2009. ICA based on a Smooth Estimation of the Differential Entropy, in: Proceedings of the 22 Annual Conference on Neural Information Processing Systems (NIPS22).

Fraser, A.M., Swinney, H.L., 1986. Independent coordinates for strange attractors from mutual information. Phys. Rev. A 33, 1134–1140. doi:10.1103/PhysRevA.33.1134

Friston, K., 2012. Ten ironic rules for non-statistical reviewers. NeuroImage 61, 1300–1310. doi:10.1016/j.neuroimage.2012.04.018

Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J., 1997. Psychophysiological and Modulatory Interactions in Neuroimaging. NeuroImage 6, 218–229. doi:10.1006/nimg.1997.0291

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

Garner, W.R., McGill, W.J., 1956. The relation between information and variance analyses. Psychometrika 21, 219–228.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural Networks and the Bias/Variance Dilemma. Neural Comput. 4, 1–58. doi:10.1162/neco.1992.4.1.1

Genest, C., Gendron, M., Bourdeau-Brien, M., 2009. The Advent of Copulas in Finance. Eur. J. Finance 15, 609–618. doi:10.1080/13518470802604457

Goodman, N.R., 1963. The Distribution of the Determinant of a Complex Wishart Distributed Matrix. Ann. Math. Stat. 34, 178–180. doi:10.2307/2991291

Gosselin, F., Schyns, P.G., 2001. Bubbles: a technique to reveal the use of information in recognition tasks. Vision Res. 41, 2261–2271. doi:10.1016/S0042-6989(01)00097-9

Granger, C.W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econom. J. Econom. Soc. 424–438.

Griffith, V., Koch, C., 2012. Quantifying synergistic mutual information. arXiv:1205.4265.

Groppe, D.M., Urbach, T.P., Kutas, M., 2011. Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. Psychophysiology 48, 1711–1725. doi:10.1111/j.1469-8986.2011.01273.x

Gross, J., 2014. Analytical methods and experimental approaches for electrophysiological studies of brain oscillations. J. Neurosci. Methods 228, 57–66. doi:10.1016/j.jneumeth.2014.03.007

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., Garrod, S., 2013. Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. PLoS Biol 11, e1001752. doi:10.1371/journal.pbio.1001752

Guggenmos, M., Thoma, V., Haynes, J.-D., Richardson-Klavehn, A., Cichy, R.M., Sterzer, P., 2015. Spatial attention enhances object coding in local and distributed representations of the lateral occipital complex. NeuroImage 116, 149–157. doi:10.1016/j.neuroimage.2015.04.004

Hämäläinen, M., Hari, R., Ilmoniemi, R.J., Knuutila, J., Lounasmaa, O.V., 1993. Magnetoencephalography - theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev. Mod. Phys. 65, 413–497. doi:10.1103/RevModPhys.65.413

Harder, M., Salge, C., Polani, D., 2012. A Bivariate Measure of Redundant Information. arXiv:1207.2080.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning. Springer Series in Statistics.

Haxby, J.V., Connolly, A.C., Guntupalli, J.S., 2014. Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. Annu. Rev. Neurosci. 37, 435–456. doi:10.1146/annurev-neuro-062012-170325

Holmes, A.P., Blair, R.C., Watson, G., Ford, I., 1996. Nonparametric Analysis of Statistic Images from Functional Mapping Experiments. J. Cereb. Blood Flow Metab. 16, 7–22. doi:10.1097/00004647-199601000-00002

Hu, L., Mouraux, A., Hu, Y., Iannetti, G.D., 2010. A novel approach for enhancing the signal-to-noise ratio and detecting automatically event-related potentials (ERPs) in single trials. NeuroImage 50, 99–111. doi:10.1016/j.neuroimage.2009.12.010

Hu, M., Liang, H., 2014. A copula approach to assessing Granger causality. NeuroImage 100, 125–134. doi:10.1016/j.neuroimage.2014.06.013

Ince, R.A.A., Giordano, B.L., Kayser, C., Rousselet, G.A., Gross, J., Schyns, P.G., 2016. Data from: A statistical framework based on a novel mutual information estimator utilizing a Gaussian copula. Dryad Digit. Repos. doi:10.5061/dryad.8b146

Ince, R.A.A., Mazzoni, A., Bartels, A., Logothetis, N.K., Panzeri, S., 2012. A novel test to determine the significance of neural selectivity to single and

multiple potentially correlated stimulus features. J. Neurosci. Methods 210, 49–65. doi:10.1016/j.jneumeth.2011.11.013

Ince, R.A.A., Montani, F., Arabzadeh, E., Diamond, M.E., Panzeri, S., 2009. On the presence of high-order interactions among somatosensory neurons and their effect on information transmission. J. Phys. Conf. Ser. 197, 012013 (11pp).

Ince, R.A.A., Panzeri, S., Kayser, C., 2013. Neural Codes Formed by Small and Temporally Precise Populations in Auditory Cortex. J. Neurosci. 33, 18277–18287. doi:10.1523/JNEUROSCI.2631-13.2013

Ince, R.A.A., Senatore, R., Arabzadeh, E., Montani, F., Diamond, M.E., Panzeri, S., 2010. Information-theoretic methods for studying population codes. Neural Netw. 23, 713–727.

Ince, R.A.A., van Rijsbergen, N., Thut, G., Rousselet, G.A., Gross, J., Panzeri, S., Schyns, P.G., 2015. Tracing the Flow of Perceptual Features in an Algorithmic Brain Network. Sci. Rep. 5, 17681. doi:10.1038/srep17681

Inouye, T., Shinosaki, K., Sakamoto, H., Toi, S., Ukai, S., Iyama, A., Katsuda, Y., Hirano, M., 1991. Quantification of EEG irregularity by use of the entropy of the power spectrum. Electroencephalogr. Clin. Neurophysiol. 79, 204–210. doi:10.1016/0013-4694(91)90138-T

Kayser, C., Ince, R.A.A., Panzeri, S., 2012. Analysis of Slow (Theta) Oscillations as a Potential Temporal Reference Frame for Information Coding in Sensory Cortices. PLoS Comput Biol 8, e1002717. doi:10.1371/journal.pcbi.1002717

Kayser, C., Logothetis, N.K., Panzeri, S., 2010. Millisecond encoding precision of auditory cortex neurons. Proc. Natl. Acad. Sci. 107, 16976 –16981. doi:10.1073/pnas.1012656107

Kayser, C., Montemurro, M.A., Logothetis, N.K., Panzeri, S., 2009. Spike-Phase Coding Boosts and Stabilizes Information Carried by Spatial and Temporal Spike Patterns. Neuron 61, 597–608.

Kayser, S.J., Ince, R.A.A., Gross, J., Kayser, C., 2015. Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. J. Neurosci. 35, 14691–14701. doi:10.1523/JNEUROSCI.2243-15.2015

Kempter, R., Leibold, C., Buzsáki, G., Diba, K., Schmidt, R., 2012. Quantifying circular–linear associations: Hippocampal phase precession. J. Neurosci. Methods 207, 113–124. doi:10.1016/j.jneumeth.2012.03.007

Kennel, M.B., Shlens, J., Abarbanel, H.D.I., Chichilnisky, E., 2005. Estimating entropy rates with Bayesian confidence intervals. Neural Comput. 17, 1531–1576.

King, J.-R., Dehaene, S., 2014. Characterizing the dynamics of mental representations: the temporal generalization method. Trends Cogn. Sci. 18, 203–210. doi:10.1016/j.tics.2014.01.002

Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Phys. Rev. E 69, 66138.

Kriegeskorte, N., Bandettini, P., 2007. Analyzing for information, not activation, to exploit high-resolution fMRI. NeuroImage 38, 649–662.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868. doi:10.1073/pnas.0600244103

Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. 17, 401–412. doi:10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. Front. Syst. Neurosci. 2. doi:10.3389/neuro.06.004.2008

Kumar, P., 2012. Statistical Dependence: Copula functions and mutual information based measures. J. Stat. Appl. Probab. Int. J. 1, 1–14.

Lachaux, J.-P., Rodriguez, E., Martinerie, J., Varela, F.J., 1999. Measuring phase synchrony in brain signals. Hum. Brain Mapp. 8, 194–208. doi:10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C

Latham, P., Roudi, Y., 2009. Mutual information. Scholarpedia 4, 1658. doi:10.4249/scholarpedia.1658

Lee, A., 2010. Circular data. Wiley Interdiscip. Rev. Comput. Stat. 2, 477–486. doi:10.1002/wics.98

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791. doi:10.1038/44565

Lefakis, L., Fleuret, F., 2014. Jointly Informative Feature Selection, in: International Conference on Artificial Intelligence and Statistics.

Magri, C., Whittingstall, K., Singh, V., Logothetis, N.K., Panzeri, S., 2009. A toolbox for the fast information analysis of multiple-site LFP, EEG and spike train recordings. BMC Neurosci. 10, 81.

Ma, J., Sun, Z., 2011. Mutual Information Is Copula Entropy. Tsinghua Sci. Technol. 16, 51–54. doi:10.1016/S1007-0214(11)70008-6

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190. doi:10.1016/j.jneumeth.2007.03.024

Massey, F.J., 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. J. Am. Stat. Assoc. 46, 68–78. doi:10.1080/01621459.1951.10500769

Massey, J., 1990. Causality, feedback and directed information, in: Proc. Int. Symp. Inf. Theory Applic.(ISITA-90). Citeseer, pp. 303–305.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta BBA - Protein Struct. 405, 442–451. doi:10.1016/0005-2795(75)90109-9

McGill, W.J., 1954. Multivariate information transmission. Psychometrika 19, 97–116. doi:10.1007/BF02289159

Miller, G., 1955. Note on the bias of information estimates. Inf. Theory Psychol. Probl. Methods 95–100.

Misra, N., Singh, H., Demchuk, E., 2005. Estimation of the entropy of a multivariate normal distribution. J. Multivar. Anal. 92, 324–342.

Montemurro, M.A., Senatore, R., Panzeri, S., 2007. Tight Data-Robust Bounds to Mutual Information Combining Shuffling and Model Selection Techniques. Neural Comput. 19, 2913–2957.

Moon, Y.I., Rajagopalan, B., Lall, U., 1995. Estimation of mutual information using kernel density estimators. Phys. Rev. E 52, 2318.

Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., Pouget, A., 2014. Information-limiting correlations. Nat. Neurosci. 17, 1410–1417. doi:10.1038/nn.3807

Murray, R.F., 2011. Classification images: A review. J. Vis. 11, 2–2. doi:10.1167/11.5.2

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. NeuroImage, Multivariate Decoding and Brain Reading 56, 400–410. doi:10.1016/j.neuroimage.2010.07.073

Nelken, I., Chechik, G., 2007. Information theory in auditory research. Hear. Res., Auditory Cortex 2006 - The Listening Brain 229, 94–105. doi:10.1016/j.heares.2007.01.012

Nelsen, R.B., 2007. An introduction to copulas. Springer.

Nemenman, I., Bialek, W., de Ruyter van Steveninck, R., 2004. Entropy and information in neural spike trains: Progress on the sampling problem. Phys. Rev. E 69, 56111.

Nemenman, I., Lewen, G.D., Bialek, W., de Ruyter van Steveninck, R.R., 2008. Neural Coding of Natural Stimuli: Information at Sub-Millisecond Resolution. PLoS Comput Biol 4, e1000025. doi:10.1371/journal.pcbi.1000025

Neyman, J., Pearson, E., 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. Philos. Trans. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character 231, 289–337.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.

Olbrich, E., Bertschinger, N., Rauh, J., 2015. Information Decomposition and Synergy. Entropy 17, 3501–3517. doi:10.3390/e17053501

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M., 2011. FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Comput. Intell. Neurosci. 2011, 1–9. doi:10.1155/2011/156869

O'Reilly, J.X., Woolrich, M.W., Behrens, T.E.J., Smith, S.M., Johansen-Berg, H., 2012. Tools of the trade: psychophysiological interactions and functional connectivity. Soc. Cogn. Affect. Neurosci. 7, 604–609. doi:10.1093/scan/nss055

Ostwald, D., Bagshaw, A.P., 2011. Information theoretic approaches to functional neuroimaging. Magn. Reson. Imaging.

Overath, T., Cusack, R., Kumar, S., von Kriegstein, K., Warren, J.D., Grube, M., Carlyon, R.P., Griffiths, T.D., 2007. An information theoretic characterisation of auditory encoding. PLoS Biol. 5, e288. doi:10.1371/journal.pbio.0050288

Paninski, L., 2003. Estimation of entropy and mutual information. Neural Comput. 15, 1191–1253.

Panzeri, S., Macke, J.H., Gross, J., Kayser, C., 2015. Neural population coding: combining insights from microscopic and mass signals. Trends Cogn. Sci. 19, 162–172. doi:10.1016/j.tics.2015.01.002

Panzeri, S., Magri, C., Logothetis, N.K., 2008. On the use of information theory for the analysis of the relationship between neural and imaging signals. Magn. Reson. Imaging 26, 1015–1025.

Panzeri, S., Petersen, R.S., Schultz, S.R., Lebedev, M., Diamond, M.E., 2001. The Role of Spike Timing in the Coding of Stimulus Location in Rat Somatosensory Cortex. Neuron 29, 769–777.

Panzeri, S., Senatore, R., Montemurro, M.A., Petersen, R.S., 2007. Correcting for the Sampling Bias Problem in Spike Train Information Measures. J. Neurophysiol. 98, 1064–1072.

Panzeri, S., Treves, A., 1996. Analytical estimates of limited sampling biases in different information measures. Netw. Comput. Neural Syst. 7, 87–107.

Park, H., Ince, R.A.A., Schyns, P.G., Thut, G., Gross, J., 2015. Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners. Curr. Biol. 25, 1649–1653. doi:10.1016/j.cub.2015.04.049

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238. doi:10.1109/TPAMI.2005.159

Pernet, C.R., Latinus, M., Nichols, T.E., Rousselet, G.A., 2015. Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. J. Neurosci. Methods, Cutting-edge EEG Methods 250, 85–93. doi:10.1016/j.jneumeth.2014.08.003

Quian Quiroga, R., Panzeri, S., 2009. Extracting information from neuronal populations: information theory and decoding approaches. Nat. Rev. Neurosci. 10, 173–185.

Quinn, C.J., Coleman, T.P., Kiyavash, N., Hatsopoulos, N.G., 2011. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. J. Comput. Neurosci. 30, 17–44. doi:10.1007/s10827-010-0247-2

Reich, D.S., Mechler, F., Victor, J.D., 2001. Independent and redundant information in nearby cortical neurons. Science 294, 2566–2568.

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C., 2011. Detecting Novel Associations in Large Data Sets. Science 334, 1518 –1524. doi:10.1126/science.1205438

Rolls, E.T., Treves, A., 2011. The neuronal encoding of information in the brain. Prog. Neurobiol. 95, 448–490. doi:10.1016/j.pneurobio.2011.08.002

Ross, B.C., 2014. Mutual Information between Discrete and Continuous Data Sets. PLoS ONE 9, e87357. doi:10.1371/journal.pone.0087357

Rousselet, G.A., Ince, R.A.A., Rijsbergen, N.J. van, Schyns, P.G., 2014a. Eye coding mechanisms in early human face event-related potentials. J. Vis. 14, 7. doi:10.1167/14.13.7

Rousselet, G.A., Ince, R.A.A., van Rijsbergen, N.J., Schyns, P.G., 2014b. Data from: Eye coding mechanisms in early human face event-related potentials. Dryad Digit. Repos. doi:10.5061/dryad.8m2g3

Salvador, R., Martínez, A., Pomarol-Clotet, E., Sarró, S., Suckling, J., Bullmore, E., 2007. Frequency based mutual information measures between clusters of brain regions in functional magnetic resonance imaging. NeuroImage 35, 83–88. doi:10.1016/j.neuroimage.2006.12.001

Saproo, S., Serences, J.T., 2010. Spatial Attention Improves the Quality of Population Codes in Human Visual Cortex. J. Neurophysiol. 104, 885–895. doi:10.1152/jn.00369.2010

Schneidman, E., Bialek, W., Berry, M.J., 2003. Synergy, Redundancy, and Independence in Population Codes. J. Neurosci. 23, 11539–11553.

Schnitzler, A., Gross, J., 2005. Normal and pathological oscillatory communication in the brain. Nat. Rev. Neurosci. 6, 285–296. doi:10.1038/nrn1650

Schreiber, T., 2000. Measuring Information Transfer. Phys. Rev. Lett. 85, 461–464. doi:10.1103/PhysRevLett.85.461

Schyns, P.G., Gosselin, F., Smith, M.L., 2009. Information processing algorithms in the brain. Trends Cogn. Sci. 13, 20–26. doi:10.1016/j.tics.2008.09.008

Schyns, P.G., Thut, G., Gross, J., 2011. Cracking the Code of Oscillatory Activity. PLoS Biol 9, e1001064. doi:10.1371/journal.pbio.1001064

Serences, J.T., Saproo, S., Scolari, M., Ho, T., Muftuler, L.T., 2009. Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. NeuroImage 44, 223–231. doi:10.1016/j.neuroimage.2008.07.043

Shannon, C.E., 1948. A mathematical theory of communication, Bell Syst. Bell Syst. Tech. J. 27, 379–423.

Sharpee, T.O., 2014. Toward Functional Classification of Neuronal Types. Neuron 83, 1329–1334. doi:10.1016/j.neuron.2014.08.040

Shlens, J., Kennel, M.B., Abarbanel, H.D.I., Chichilnisky, E., 2007. Estimating information rates with confidence intervals in neural spike trains. Neural Comput. 19, 1683–1719.

Singer, W., 2013. Cortical dynamics revisited. Trends Cogn. Sci., Special Issue: The Connectome 17, 616–626. doi:10.1016/j.tics.2013.09.006

Sklar, M., 1959. Fonctions de répartition à n dimensions et leurs marges. Université Paris 8.

Smith, M.L., Gosselin, F., Schyns, P.G., 2004. Receptive Fields for Flexible Face Categorizations. Psychol. Sci. 15, 753–761. doi:10.1111/j.0956-7976.2004.00752.x

Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage 44, 83–98. doi:10.1016/j.neuroimage.2008.03.061

Sokal, R.R., Rohlf, F.J., 1981. Biometry. New York: WH Freeman and Company.

Steuer, R., Kurths, J., Daub, C.O., Weise, J., Selbig, J., 2002. The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics 18, S231–S240. doi:10.1093/bioinformatics/18.suppl_2.S231

Strong, S.P., Koberle, R., de Ruyter van Steveninck, R.R., Bialek, W., 1998. Entropy and Information in Neural Spike Trains. Phys. Rev. Lett. 80, 197–200.

Szymanski, F.D., Rabinowitz, N.C., Magri, C., Panzeri, S., Schnupp, J.W.H., 2011. The Laminar and Temporal Structure of Stimulus Information in the Phase of Field Potentials of Auditory Cortex. J. Neurosci. 31, 15787 –15801. doi:10.1523/JNEUROSCI.1416-11.2011

Thut, G., Miniussi, C., Gross, J., 2012. The Functional Importance of Rhythmic Activity in the Brain. Curr. Biol. 22, R658–R663. doi:10.1016/j.cub.2012.06.061

Timme, N., Alford, W., Flecker, B., Beggs, J.M., 2013. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. J. Comput. Neurosci. 36, 119–140. doi:10.1007/s10827-013-0458-4

Torkkola, K., 2003. Feature extraction by non parametric mutual information maximization. J. Mach. Learn. Res. 3, 1415–1438.

Victor, J., 2006. Approaches to information-theoretic analysis of neural activity. Biol. Theory 1, 302–316.

Victor, J.D., 2002. Binless strategies for estimation of information from neural data. Exp Brain Res Phys Rev E 66, 051903.

Voytek, B., D'Esposito, M., Crone, N., Knight, R.T., 2013. A method for event-related phase/amplitude coupling. NeuroImage 64, 416–424. doi:10.1016/j.neuroimage.2012.09.023

Wang, X.-J., 2010. Neurophysiological and Computational Principles of Cortical Rhythms in Cognition. Physiol. Rev. 90, 1195–1268. doi:10.1152/physrev.00035.2008

Wibral, M., Vicente, R., Lindner, M., 2014. Transfer Entropy in Neuroscience, in: Wibral, M., Vicente, R., Lizier, J.T. (Eds.), Directed Information Measures in Neuroscience, Understanding Complex Systems. Springer Berlin Heidelberg, pp. 3–36.

Wiener, N., 1956. The theory of prediction. McGraw-Hill: New York, NY, USA.

Wilcox, L.D., Niles, L.T., 1995. Maximum mutual information vector quantization, in: , 1995 IEEE International Symposium on Information Theory, 1995. Proceedings. Presented at the , 1995 IEEE International Symposium on Information Theory, 1995. Proceedings, p. 434–. doi:10.1109/ISIT.1995.550421

Williams, P.L., Beer, R.D., 2010. Nonnegative Decomposition of Multivariate Information. arXiv:1004.2515.

Zeng, X., Durrani, T.S., 2011. Estimation of mutual information using copula density function. Electron. Lett. 47, 493–494. doi:10.1049/el.2011.0778