LETTER Communicated by Stefano Panzeri

Reliability of Information-Based Integration of EEG and fMRI Data: A Simulation Study

Sara Assecondi
sara.assecondi@me.com
School of Psychology, University of Birmingham, Birmingham, B17 2TT, U.K.

Dirk Ostwald
dirk.ostwald@mpib-berlin.mpg.de
Max Planck Institute for Human Development, 14195 Berlin, Germany

Andrew P. Bagshaw
a.p.bagshaw@bham.ac.uk
School of Psychology, University of Birmingham, Birmingham, B17 2TT, U.K.

Most studies involving simultaneous electroencephalographic (EEG) and functional magnetic resonance imaging (fMRI) data rely on the first-order, affine-linear correlation of EEG and fMRI features within the framework of the general linear model. An alternative is the use of information-based measures such as mutual information and entropy, which can also detect higher-order correlations present in the data. The estimate of information-theoretic quantities might be influenced by several parameters, such as the numerosity of the sample, the amount of correlation between variables, and the discretization (or binning) strategy of choice. While these issues have been investigated for invasive neurophysiological data and a number of bias-correction estimates have been developed, there has been no attempt to systematically examine the accuracy of information estimates for the multivariate distributions arising in the context of EEG-fMRI recordings. This is especially important given the differences between electrophysiological and EEG-fMRI recordings. In this study, we drew random samples from simulated bivariate and trivariate distributions, mimicking the statistical properties of EEG-fMRI data. We compared the estimated information shared by simulated random variables with its numerical value and found that the interaction between the binning strategy and the estimation method influences the accuracy of the estimate. Conditional on the simulation assumptions, we found that the equipopulated binning strategy yields the best and most consistent results across distributions and bias correction methods. We also found that within bias correction techniques, the asymptotically debiased (TPMC), the jackknife debiased (JD), and the best upper bound (BUB) approach give similar results, and those are consistent across distributions.

Neural Computation **27, 281–305** (2015)
doi:10.1162/NECO_a_00695

1 Introduction

Brain activity decodes and transfers information between brain and body structures to allow the individual to interact with the external world. The transmission of information through a given channel has been formalized by Shannon's theory of communication. In his seminal paper Shannon (1948) proposed a probabilistic framework that characterizes the information transmitted by a source through a communication channel. The framework is based on two quantities: entropy, a measure of the uncertainty or variability of a random variable, and mutual information, which quantifies how much the knowledge of one variable is improved by the knowledge of another variable.

The simultaneous recording of electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) allows for the joint monitoring of how the brain encodes external or internal stimuli into either the electrophysiological or hemodynamic signal (Herrmann & Debener, 2008; Ritter & Villringer, 2006; Vulliemoz, Lemieux, Daunizeau, Michel, & Duncan, 2010) Most EEG-fMRI studies rely on linear correlation of EEG and fMRI features within the framework of the general linear model (e.g., Debener, Ullsramon Siegel, Fiehler, von Cramon, & Engel, 2005; Eichele et al., 2005; Goldman et al., 2009). An alternative is the use of information-based measures such as mutual information and entropy (Ostwald & Bagshaw, 2011; Ostwald, Porcaro, & Bagshaw, 2011), which also incorporate higher-order correlations present in the data (Caballero-Gaudes et al., 2013; Ojemann, Ojemann, & Ramsey, 2013; Pouliot et al., 2012; Yešilyurt, Uğurbil, & Uludağ, 2008; Zhang, Zhu, & Chen, 2008). It is generally believed that exploiting higher-order correlations furthers our understanding of the stimulus-response and response-response relations (Herrmann & Debener, 2008), without the need of assumptions about the linearity of the underlying coupling. Information theory concepts may then unfold properties not detectable by more traditional linear approaches.

The reliability of the estimate of information quantities is highly dependent on the accuracy obtained when estimating the underlying probability density (Panzeri, Senatore, Montemurro, & Petersen, 2007). Information-based signal analysis requires a large number of samples to obtain an unbiased estimate of the information quantities of interest. In the context of neuroscience applications, this nonlinear approach has proven successful in the analysis of invasive electrophysiological recordings (Borst & Theunissen, 1999; Quian Quiroga & Panzeri, 2009; Victor, 2006). In the case of EEG-fMRI data, the number of samples, or trials, is limited by the duration of the experiment. This reduction in the number of samples leads to a bias in the estimate of entropy and information, which is more severe than in invasive electrophysiology For example, while a typical task-related EEG-fMRI experiment acquires, in the best-case scenario, a few hundred trials per condition (Bagshaw & Warbrick, 2007; Mulert & Lemieux, 2010; Ostwald,

Porcaro, & Bagshaw, 2010), in single-cell experiments with anesthetized animals, thousand of trials may be recorded (Panzeri, Magri, & Logothetis, 2008). The estimate might also be influenced by other parameters, such as the amount of correlation between variables and the discretization (or binning) strategy of choice. While these issues have been investigated for invasive neurophysiological data and bias-correction techniques have been developed (Panzeri et al., 2007), there has been no attempt to systematically examine the accuracy of information estimates for the multivariate distributions relevant to EEG and fMRI data. This is especially important given the differences between electrophysiological and EEG-fMRI recordings. The underlying properties (e.g., distribution, variance, amount of correlation between variables) are likely to be different for the two types of data. For example, while spike count in electrophysiological recordings is likely to be Poisson distributed (Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1996), for example, EEG or fMRI features are usually modeled as gaussian or gamma distributed (Hu et al., 2011; Van Zandt, 2000).

In this study, we drew random samples from simulated bivariate and trivariate distributions, mimicking the statistical properties of EEG-fMRI data. We estimated the information shared by simulated random variables. By comparing the estimated information with its true value, calculated numerically, we assessed how the number of samples, the amount of correlation, and the binning strategy interact with different bias correction techniques to affect the accuracy of information estimates.

2 Methods

2.1 Information Theory in Neuroscience. Originally developed with communication systems such as telegraphy in mind, the theory of communication (Shannon, 1948) formalizes the transmission of a message from a source to a destination through a given channel. During transmission, the encoded message can be affected by noise. Information theory quantifies the actual amount of transmitted information.

A given message is codified into a series of symbols. A symbol is the basic unit of the code. Symbols can be combined into words to uniquely identify certain messages. All the available symbols form the alphabet, while all the possible words form a dictionary. Knowing the alphabet and the dictionary allows one to properly encode and decode a message.

The theory of communication was transferred to neurophysiology (see Borst & Theunissen, 1999, and Rieke et al., 1996, for a review and references to early works), with the aim of modeling how the brain encodes and decodes external stimuli and translates them from sensory input to higher cognitive functions. A similar approach has also been taken with fMRI data (de Araujo et al., 2003; Fuhrmann Alpert, Hein, Tsai, Naumer, & Knight, 2008; Fuhrmann Alpert, Sun, Handwerker, D'Esposito, & Knight, 2007).

Neurons transmit information through trains of action potentials, or spikes, and the neural representation of information is called the neural code. The basic idea is "to derive a dictionary that, given some snapshot of spiking activity, tells us what sensory signal has occurred" (Panzeri et al., 2007; Rieke et al., 1996). In terms of noninvasive data, the neural information is encoded in hemodynamic or mass electrophysiological activity, adding a level of complexity to the decoding given the empirical uncertainties in understanding the relationship between spiking activity and noninvasive signals.

Irrespective of the specific data type, to create a dictionary to decode the neural code, first we need to define the observation period, that is, the time window necessary to fully encode one stimulus occurrence (e.g., the trial length). Second, we must decide which signal features encode the stimulus (e.g., signal amplitude, latency, spike count, spike frequency). Then we divide the observation period into time bins and extract the features within each time bin. The number of time bins represents the number of symbols in a word that encodes the stimulus. The number of different levels (different symbols) a feature can assume in a time bin depends on the discretization of that feature and forms the alphabet for that feature or variable—and it is related to the number of histogram bins used to estimate the probability density function (pdf) or probability mass function (pmf).

In the case of simultaneous EEG and fMRI recordings, we may proceed as follows. The observation period can be the trial duration (in case of event-related designs) or another appropriate time window with a duration that depends on the dynamics of the process we would like to study (e.g., in case of resting data). The choice of features also depends on the hypothesis being tested. EEG, fMRI, and behavioral data can encode the stimulus independently or jointly. An appropriate feature can be, for example, the amplitude or latency of certain event-related components on one or more channels, the shape parameters of the BOLD response in one or more voxels, reaction times, or other behavioral measures. It is worth noticing that increasing the number of features increases exponentially the complexity of the problem.[1]

2.2 Information Theory: Background. Here we briefly introduce entropy and information, closely following Cover and Thomas (2006) and Rieke et al. (1996). We focused on discrete quantities because empirical data

[1]For the ith modality, let l_i be the quantization levels used to discretize the recorded response, n_i be the number of time bins used to segment the observation period, and c_i the number of recorded signal—EEG channels or fMRI voxels (ROIs) or behavioral variables. The number of symbols associated with the ith modality is $s_i = n_i c_i$ and the number of corresponding possible words is $w_i = l_i^{n_i c_i}$. If we consider M modalities, then the total number of possible words that codify the stimulus is $W = \prod_{i=1}^{M} w_i$. It is clear that the number of words corresponds to the complexity of the problem and increases exponentially, while the amount of data available stays the same.

are usually discretized during signal processing. Let X be a source signal that can be described as a random variable with associated probability mass function (pmf) $p(x)$ and alphabet $\mathcal{X}$. The entropy of X is a measure of the uncertainty associated with $p(x)$ and is calculated as follows:

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x). \tag{2.1}$$

Intuitively, the higher the variability associated with X, the more information X can encode. Consider another variable Y, with *pmf* $p(y)$ and alphabet $\mathcal{Y}$ observed at the same time as X. The joint entropy (i.e., the entropy of the joint distribution) is

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y). \tag{2.2}$$

The conditional entropy, that is, the remaining variability of X after the variability due to Y has been accounted for, is

$$H(X|Y) = -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x), \tag{2.3}$$

where $p(y|x)$ is the conditional probability, representing the probability of x given y.

Mutual information is a measure of the amount of information one random variable carries about another.[2] It is calculated as the distance between the joint distribution (corresponding to the case of dependent variable) and the product distribution (i.e, the joint distribution of two independent variables):

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(X,Y). \end{aligned} \tag{2.4}$$

The choice of the base of the logarithm is completely arbitrary. In this work, the logarithm to base 2 is used and information quantities are expressed in bits.

[2]The mutual information is a special case of relative entropy. The relative entropy (or Küllback-Leibler distance) is the distance between two probability mass functions $p(x)$ $q(x)$. It is defined as $D(p||q) = \sum_{x\in X} p(x)\log(p(x))/(q(x))$. The relative entropy is always nonnegative, and it is zero if and only if $p = q$.

2.3 Calculation of Information Quantities.

2.3.1 Where Does the Bias Come From?. The estimation of information quantities from experimental data requires the knowledge of the probability densities, usually estimated from the finite sample available.[3] These estimates are biased with respect to the true information, and the bias can be caused by the small sample size and the response quantization (or regularization) (Optican, Gawne, Richmond, & Joseph, 1991):

- *Bias due to sample size*. When the underlying distribution is estimated from a small sample, the variability (i.e., the entropy) associated with the *pmf* is underestimated because the small sample is unlikely to capture the full spectrum of the possible responses, resulting in an overestimation of the information.[4] This problem is more severe in EEG-fMRI data where the number of available trials is usually smaller than in invasive recordings.
- *Bias due to response quantization.* The estimation of information quantities requires the estimation of the underlying *pmf*. A common approach is the direct method (Strong, Koberle, de Ruyter van Steveninck, & Bialek, 1998), which is the normalized count of the data (histogram). Because of the variability of physiological responses to stimuli, multiple responses to the same stimulus are likely to be slightly different. This implies that for a given stimulus, each response will have a count of one, preventing the estimation of the underlying probability. The data therefore need to be regularized (i.e., binned) before counting. In this way, different responses are assigned to the same bin (if they fall within the range determined by the bin center and width). This regularization however, leads to an underestimation of the variability of the data (since data in the same bin are approximated by the bin center) and therefore an overestimation of the information.

2.3.2 Bias Correction Techniques. Several types of bias correction techniques have been introduced. In this study, we focus on techniques that have been successfully applied to neurophysiological data and for which an implementation is already available. We compare them with the PLUGIN estimate, for which the underlying probabilities are estimated as the histogram of the frequencies of the values in the available sample without

[3] Information-theoretic quantities are functionals, that is, functions of (probability mass/density) functions.

[4] $I(X;Y) = H(X) - H(X|Y)$. The entropy is underestimated, and the bias associated with $H(X)$ is smaller than the bias associated with $H(X|Y)$. This is due to the fact that, typically, $H(X|Y)$ is obtained as the average over conditional marginal entropies $H(X|Y = y)$, for each of which there are fewer samples than for the complete marginal of X. As a result, the information is overestimated because $H(X|Y)$ is more biased than $H(X)$and negative.

further correction. The bias due to sample size can be addressed by estimating the number of empty bins, and the bias due to response quantization can be addressed by not relying on a frequentist approach to estimate the underlying probability. We briefly describe the salient properties of each technique:

- *Jackknife debiased* (JD) (Efron & Tibshirani, 1993; Efron, 1979). Jackknife methods estimate the bias and variance of a statistic of interest by systematically recomputing that statistic leaving out one or more observations at a time from the sample set. The entropy is repeatedly estimated from the jackknife sample of the sample data and averaged over the number of repetitions. Then the bias is calculated as the unbiased average difference between the average jackknife estimate and the sample estimate.
- *Asymptotically debiased or Treves, Panzeri, Carlton, and Miller* (TPMC) (Carlton, 1969; Miller, 1955; Panzeri & Treves, 1996). The bias is approximated by a second-order expansion of the inverse power of the sample size. The free parameter of the leading term of the expansion in the asymptotic regime is estimated through an analytical approximation that takes into account the total number of relevant, or nonempty, bins and the number of relevant bins per stimulus. A Bayesian-like procedure is used to estimate the number of empty bins.
- *Debiased Ma bound* (MA) (Ma, 1981; Strong et al., 1998). As in the TPMC approach, the bias is approximated by a second-order expansion of the inverse power of the sample size. The free parameters of the expansion are estimated from the data.
- *Best upper bound* (BUB) (Paninski, 2003). The bias is reduced by estimating the entropy from the data, as the best approximating polynomial.
- *Coverage-adjusted* (CA) (Chao & Shen, 2003). The coverage-adjusted method takes into account the fact that a small sample may not fully cover the full range of the possible values. The method estimates the number of bins necessary to fully cover the range of possible values, thus correcting for the bias due to the small sample size.
- *Bayesian with a Dirichlet prior* (BD) (Wolpert & Wolf, 1995). Rather than using the frequentist approach (estimating the probabilities as the counts of the occurrences of a given sample, after discretization into bins), the entropy is estimated using a Bayesian approach.

2.3.3 Binning Strategies. In this work we used the direct method, as the normalized count of the binned data (histogram), for the estimation of probability mass functions associated with experimental data (Strong et al., 1998). A poor binning may affect the bias due to poor regularization (see section 2.3.1). It is therefore important to understand the effect of the

Table 1: The Multivatriate Distributions Considered in This Work.

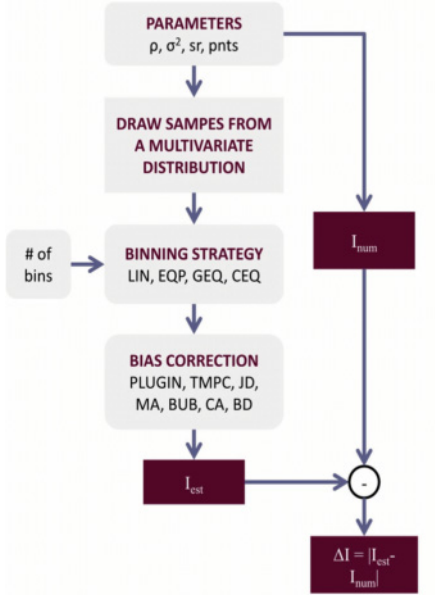
Number of Variables	Marginals	Abbreviation
2	Normal-normal	NN
2	Normal-uniform	NU
2	Normal-gamma	NG
3	Normal-normal-normal	NNN
3	Normal-normal-uniform	NNU
3	Normal-normal-gamma	NNG
3	Normal-uniform-gamma	NUG
3	Normal-uniform-uniform	NUU

binning strategy on the information estimate. We considered four binning strategies (Magri, Whittingstall, Singh, Logothetis, & Panzeri, 2009):

- *Linearly equispaced bins* (LIN). The sample range (i.e., the support of the histogram, between the smaller and the larger sampled values) is divided into N bins of equal width;
- *Equipopulated bins* (EQP). The sample range is divided into bins of variable width. The bins' widths are adjusted to obtain approximately the same number of elements in each bin. This renders the underlying distribution more uniform.
- *Gaussian equispaced bins* (GSE). The sample range is defined as the mean of the data plus or minus two standard deviations. The range is then divided into N equispaced bins.
- *Centered equispaced bins* (CEQ). The sample range is defined as the mean of the data plus or minus the largest absolute sampled value. The range is then divided into N equispaced bins.

2.4 Simulated Data. We simulate bivariate and trivariate distributions, with uniform, normal, and gamma marginals. These particular marginals were chosen because they may be representative of real-world scenarios: normal distributions correspond to common modeling assumptions in EEG or fMRI feature analysis; a uniform distribution may model an "uninformative data feature"; a gamma distribution may model nonnegative EEG frequency powers or reaction times (Hu et al., 2011; Van Zandt, 2000). The list of the distributions considered is shown in Table 1. This is a subset spanning the most likely cases.

2.4.1 Parameters. For each model, we varied the following parameters: variance (σ^2) of each marginal, correlation (ρ) between two marginals (kept constant across pairs in the case of trivariate distributions), and ratio of the standard deviation between marginals (SR). We tested four binning strategies and seven bias correction approaches in samples of different



PARAMETERS	
ρ	= [0.0, 0.3, 0.5, 0.7, 0.9]
σ^2	= [0.2, 0.4, 0.6, 0.8, 1]
sr	= [1 2 3 4]
pnts	= [32, 64, 128, 256, 512, 1024, 2048]

BINNING STRATEGY

- linearly spaced bins (LIN),
- equipopulated bins (EQP)
- gaussian equispaced bins (GSE)
- centered equispaced bins (CEQ)

BIAS CORRECTION (Statoolkit (Goldberg et al,2009))

- Plugin estimate (PLUGIN)
- Asymptotically debiased (TPMPC) (Treves & Panzeri, 1995)
- Jackknife debiased (JD) (Efron & Tibshirani, 1993)
- Debiased Ma bound (MA) (Ma, 1981)
- Best upper bound (BUB) (Paninski, 2003)
- Coverage-adjusted (CA) (Chao & Shen, 2003)
- Bayesian with a Dirichlet prior (BD) (Wolpert & Wolf, 1995)

Figure 1: (Left) Flowchart describing the method. (Right) Parameters and their values used.

numerosity (PNTS). The parameters and the allowed values are reported in Figure 1 (right). For clarity, in the case of the trivariate distribution, the σ^2 and SR were kept equal to 1, since they were found not to affect the information estimate in the case of the bivariate distribution. The specific correlation structure in the multivariate cases was modeled using copulas. A copula is a function that links univariate marginals to their multivariate distribution (see the appendix for more details on copulas). By using copulas we were able to vary the amount of correlation given a priori chosen marginal distributions. For generality, and because we had no reason to choose differently, a gaussian copula was used (Valdez, 1998). Examples of the simulated bivariate distributions are shown in Figure 2.

The bivariate gaussian covariance matrices were varied in terms of their first marginal standard deviation $s_1 > 0$, their correlation coefficient $\rho \in [-1, 1]$, and the ratio between first and second marginal standard deviation $sr \in Q$. Each simulated bivariate gaussian covariance matrix (Σ^{BV}) was then evaluated using

$$s_2 = sr \cdot s_1 \tag{2.5}$$

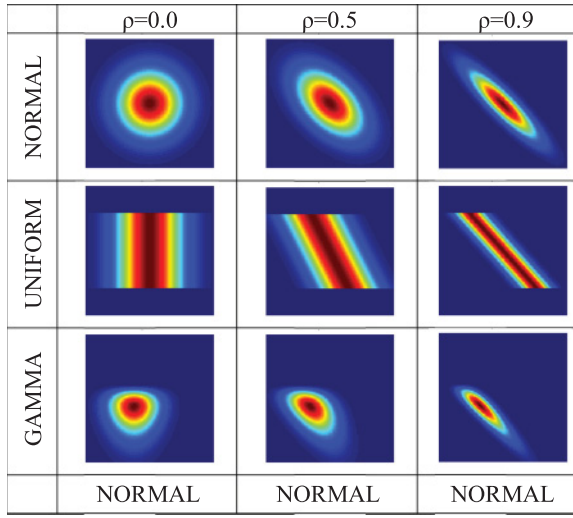


Figure 2: Example of the simulated underlying distributions for the bivariate case, with ρ equal to 0.0, 0.5, and 0.9, respectively.

as

$$\Sigma^{\mathrm{BV}} = \left(\sigma_{ij}^2\right)_{1\le i,\, j\le 2} := \begin{pmatrix} s_1^2 & \rho s_1 s_2 \\ \rho s_1 s_2 & s_2^2 \end{pmatrix}. \tag{2.6}$$

Specifically, in this manner, the covariance between two marginal variables X_i and X_j, where $i \neq j$, is given as

$$\sigma_{ij}^2 := \rho s_i s_j \Leftrightarrow \sigma_{ij}^2 = \rho\sqrt{\sigma_{ii}^2}\sqrt{\sigma_{jj}^2} \Leftrightarrow \rho = \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} \tag{2.7}$$

and allows recovering the standard definition of the correlation coefficient.

The trivariate gaussian covariance matrices were varied in terms of the pairwise correlation coefficients between marginals only. For $s > 0$ and $\rho \in [-1, 1]$, each simulated trivariate gaussian covariance matrix (Σ^{TV}) was evaluated as

$$\Sigma^{\mathrm{TV}} = \left(\sigma_{ij}^2\right)_{1\le i,\, j\le 3} := \begin{pmatrix} s^2 & \rho s^2 & \rho s^2 \\ \rho s^2 & s^2 & \rho s^2 \\ \rho s^2 & \rho s^2 & s^2 \end{pmatrix}. \tag{2.8}$$

2.4.2 Evaluation. For each model (varying PNTS, σ^2, SR, ρ) we drew 50 realizations. In the trivariate case, the 32 PNTS model was discarded because the number of samples was insufficient to obtain a meaningful binned distribution. The Infotoolbox (Magri et al., 2009) was used to bin the data, while Statoolkit (Goldberg, Victor, Gardner, & Gardner, 2009) was used to estimate the information with different bias corrections.

For the simple cases reported in Table 1, a numerical solution for the information can be calculated. By comparing the estimation of the information (I_{est}) with its numerical counterpart (I_{num}), we assessed the amount of bias as a function of the different parameters. Numerical entropies were calculated with numerical integration implemented as Monte-Carlo integration with importance sampling. The random samples at which the integral was evaluated were generated according to a predefined probability distribution, matching the underlying data distribution. For each model, we tested the effect of the particular combination of bias correction and binning strategy.

Owing to the binning, the estimated entropy and information cannot be compared to their numerical counterparts, since they depend on the bin width. It has been shown that (Cover & Thomas, 2006)

$$\begin{aligned} H\left(X^{\Delta}\right) &= -\sum_{i=-\infty}^{+\infty} \Delta_i f\left(x_i\right) \log x_i - \sum_{i=-\infty}^{+\infty} \Delta_i f\left(x_i\right) \log \Delta_i \\ &= H\left(X\right) - \sum_{i=-\infty}^{+\infty} \Delta_i f\left(x_i\right) \log \Delta_i, \end{aligned} \tag{2.9}$$

where X^{Δ} is the discretized version of the continuous random variable X with *pdf* $f(x)$ and Δ is the bin width. In the multidimensional case Δ is also multidimensional (an area in the bivariate case and a volume in the trivariate case). The correction in equation 2.9 was applied to the estimated information before comparison. The number of bins used for discretization was calculated as $\sqrt[n]{PNTS/5}$, where n is the number of underlying variables. This was chosen so as to obtain approximately five elements per bin.

3 Results

In this section we first qualitatively describe how the difference between numerical and estimated information ($\Delta I = I_{est} - I_{num}$) varies as a function of the parameters considered. We then consider each modeled distribution separately and test the statistical significance of the overall main effects and interactions. We finally proceed with a follow-up analysis by breaking the data into subsets according to the amount of correlation and number of points used in each model. We concentrate on cases that better reflect the properties of EEG-fMRI or invasive electrophysiological data

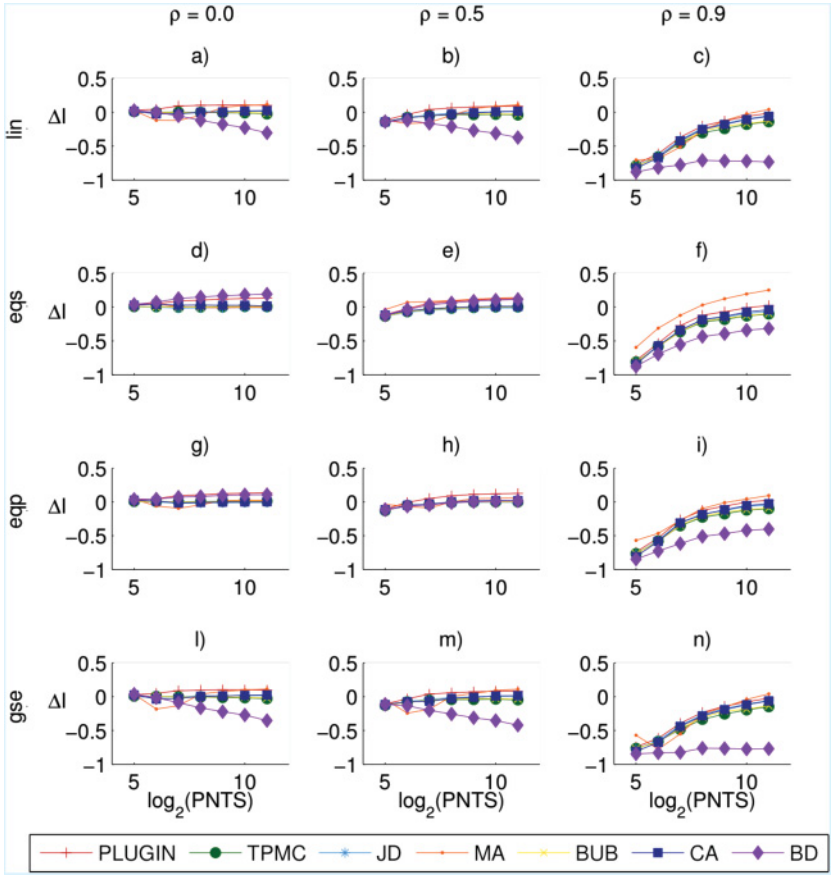


Figure 3: Bivariate normal-normal $\Delta I = I_{est} - I_{num}$ as a function of the number of sample (PNTS) for a subset of the possible correlation values (0.0, 0.5, 0.9) and for different binning strategies (LIN, EQP, GSE, CEQ) for each bias correction technique. When the correlation is weak ($\rho = 0.0$ and $\rho = 0.5$) the EQP and the GES binning strategies decrease the discrepancy between bias correction algorithms (see panels d, e, g, and h). When the correlation between variables is strong ($\rho = 0.9$), EQP give less stable estimates across algorithms (see panel f).

(32, 64, and 2048 samples) and a subset of the possible correlations (0.0, 0.5 and 0.9).

3.1 Qualitative Considerations. Figures 3 and 4 show the error $\Delta I = I_{est} - I_{num}$ as a function of the number of samples (PNTS) for different pairs of correlation and binning strategy (BS) for each bias correction (BC) technique

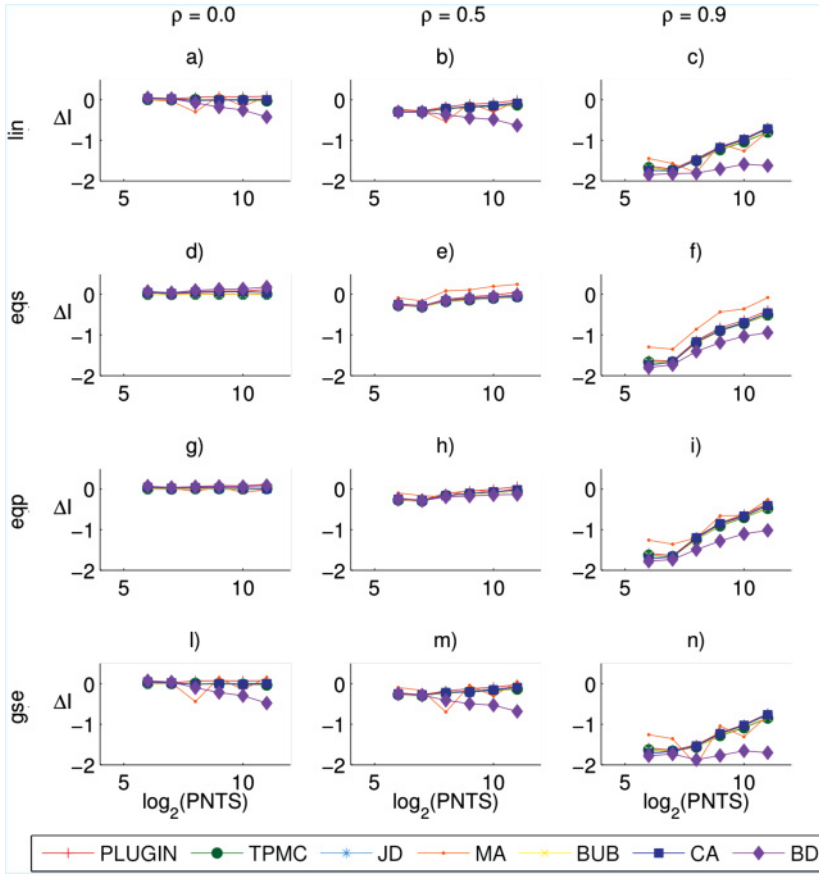


Figure 4: Trivariate normal-normal-normal $\Delta I = I_{est} - I_{num}$ as a function of the number of sample (PNTS) for a subset of the possible correlation values (0.0, 0.5, 0.9) and for different binning strategies (LIN, EQP, GSE, CEQ) for each bias correction technique. Note that the smallest number of samples used in the trivariate case is 64. When the correlation is weak ($\rho = 0.0$ and $\rho = 0.5$), the EQP and the GES binning strategies decrease the discrepancy between bias correction algorithms (see panels d, e, g, h). When the correlation between variables is strong ($\rho = 0.9$), EQP gives less stable estimates across algorithms (see panel f).

for the bivariate (NN) and the trivariate (NNN) case, respectively. Increasing the amount of correlation increases the error. At the same time, increasing the number of samples decreases the error, particularly when the correlation is strong. The choice of the binning strategy also affects the results: when the correlation is weak ($\rho = 0.0$ and $\rho = 0.5$), the EQP and the GES binning

strategies decrease the discrepancy between bias correction algorithms (see Figures 3d, 3e, 3g, and 3h and Figures 4d, 4e, 4g, and 4h).

When the correlation between variables is strong ($\rho = 0.9$), EQP gives less stable estimates across algorithms (see Figures 3f and 4f). Similar considerations apply to both the bivariate and the trivariate cases, when the underlying distribution is prevalently normal (NNN, NNU, NNG). When the underlying distribution is not prevalently normal (NGU, NUU), LIN and CEQ give more stable estimates across algorithms (data not shown). A stable estimate across algorithms would be preferable in cases where the underlying parameters cannot accurately be estimated, for example, with very few samples, allowing only a poor estimate of the statistics during the exploratory analysis.

3.2 Statistical Analysis. In the following, rather than report all possible comparisons, we have concentrated on the most informative and relevant lower-order interactions. A $\sigma^2(4) \times \text{SR}(4) \times \text{PNTS}(5) \times \rho(5) \times \text{BC}(7) \times \text{BS}(4)$ repeated measurements ANOVA (with BC and BS as repeated factors) revealed a significant ($p < 0.001$) main effect and interaction of correlation, binning strategy, bias correction, and number of samples, regardless of the underlying distribution, in the bivariate case. The main effects and interactions related to σ^2 and SR were not significant in the NN and NU bivariate case, but were significant for NG. However, in the latter case, the distribution became very distorted and nonphysiological for high SR values. A similar test ($\text{PNTS}(5) \times \rho\ (5) \times \text{BC}(7) \times \text{BS}(4)$ repeated measurements ANOVA, with BC and BS repeated factors) also revealed a significant ($p < 0.001$) main effect and interaction of correlation, binning strategy, bias correction, and number of samples again, regardless of the underlying distribution, in the trivariate case.

3.3 Follow-Up Analysis and Pairwise Comparison. Since the main effects and interactions related to σ^2 and SR were found not significant (when their values do not distort the shape of the underlying distribution to the point that it became physiologically meaningless, as in the NG case), we subsequently focused on the case of $\sigma^2 = 1$ and $\text{SR} = 1$. We considered nine submodels for each distribution—for every possible combination of PNTS (32, 64, and 2048) and ρ (0.0, 0.5, 0.9) in the bivariate case, and PNTS (64, 2048) and ρ (0.0, 0.5, and 0.9) in the trivariate case. For clarity, we refer to the different model as $\text{XX}_{(\text{PNTS},\ \rho)}$ where XX indicates the underlying marginals, PNTS the number of samples, and ρ the amount of correlation for that specific model.

A $\text{BC}(7) \times \text{BS}(4)$ repeated measurements ANOVA, with BC and BS repeated factors, revealed a significant main effect for BC ($p < 0.001$), regardless of the model, for both bivariate and trivariate distributions. We also found a main effect of BS in most of the models (except for $\text{NN}_{(32,0.0)}$, $\text{NU}_{(32,0.0)}$, $\text{NG}_{(32,0.0)}$, $\text{NN}_{(32,0.9)}$, and $\text{NU}_{(32,0.0.9)}$and $\text{NNN}_{(64,0.0)}$, $\text{NNN}_{(64,0.0)}$, $\text{NNU}_{(64,0.0)}$, $\text{NUU}_{(64,0.0)}$, $\text{NUU}_{(64,0.5)}$, and $\text{NUU}_{(64,0.9)}$). The ANOVA also

revealed a significant interaction between BS and BC ($p < 0.001$) for every model, regardless of the number of samples, the amount of correlation, or the number of underlying variables.

To gain a better understanding of how a particular bias correction technique interacts with a given binning strategy, we restrict the following analysis to TPMC, JD, and BUB, the best-performing bias correction techniques according to Figures 3 and 4, and the PLUGIN estimate for reference. Figure 5 shows the absolute error of the information estimate for different combinations of BS (LIN, EQP, GES, CEQ) and BC (PLUGIN, TMPC, JD, and BUB) for different models (PNTS = 32, 64 or 2048; $\rho = 0.0, 0.5$, or 0.9) in the case of a bivariate NN and a trivariate NNN. Similar results were found with the other marginal distributions tested (data not shown).

In both the bivariate and the trivariate case, with no correlation (see Figures 5a, 5d, and 5g), the error associated with the PLUGIN estimate is significantly larger ($p < 0.001$) than the one associated with other bias correction techniques, regardless of the underlying distribution (data not shown) or the number of points.

Increasing the correlation reduces the difference between the PLUGIN and the other bias correction techniques (one-way ANOVA with BC as a factor and grouping all BS together). In the bivariate case, the PLUGIN estimate significantly outperforms the bias-corrected estimate (always $p < 0.001$), when few points are considered (see Figures 5b, 5c, 5e, and 5f) or the correlation is very high (see Figures 5c, 5f, and 5i). When we increase the number of samples, the bias-corrected estimate significantly ($p < 0.001$) outperforms the PLUGIN estimate in the medium-correlation range (see Figure 5h). In the trivariate case, in the low- and medium-correlation range, the bias-corrected estimate significantly ($p < 0.001$) outperforms the PLUGIN estimate in the low-correlation range (see Figures 5a and 5d), while the opposite result is found in the medium-correlation range (see Figures 5b and 5e), regardless of the underlying distribution (data not shown). In the high-correlation range, the difference between the PLUGIN and the bias-corrected estimates loses significance in most of the underlying distributions (see Figures 5c and 5f), especially when 64 samples are used.

A one-way ANOVA with BS as factor and grouping all BC together revealed that increasing the correlation and the number of samples significantly increases the differences between binning strategies, with EQP and GES being on average better than LIN and CEQ. Similar considerations apply to both bivariate and trivariate distribution, regardless of the underlying marginals.

4 Discussion

The application of information theory to EEG-fMRI data sets may reveal stimulus-response and response-response relationships that up to now have been overlooked because of the assumptions the general linear model relies

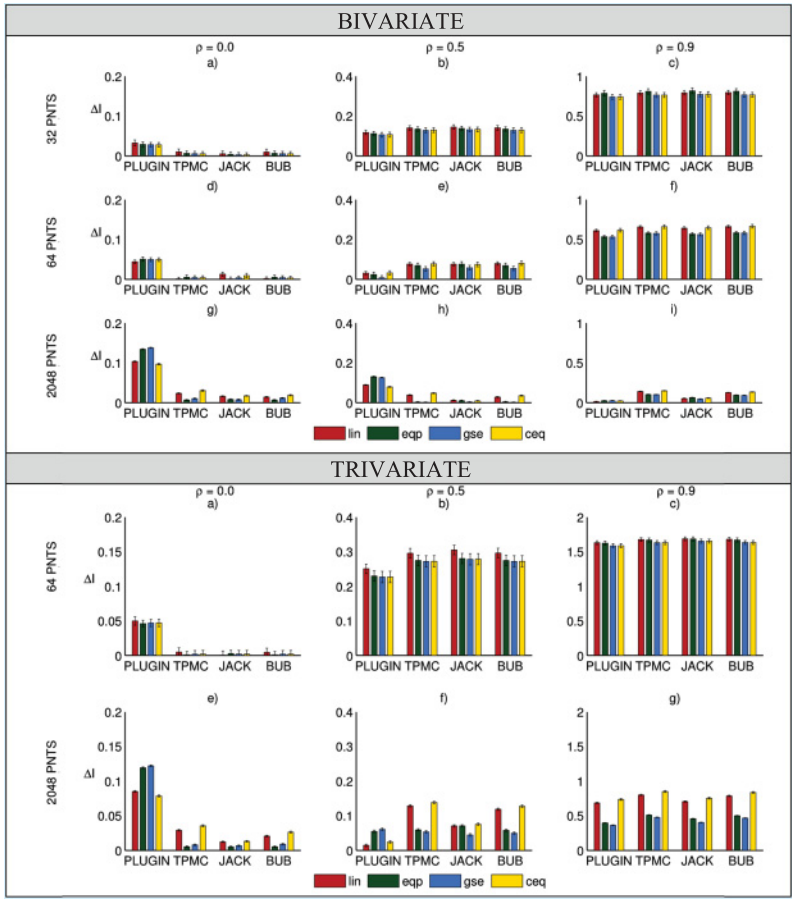


Figure 5: Bivariate (NN) and trivariate (NNN) $\Delta I = I_{est} - I_{num}$ for different combinations of BS (LIN, EQP, GES, CEQ) and BC (PLUGIN, TMPC, JD, and BUB) for different models (PNTS = 32, 64, or 2048; $\rho = 0.0$, 0.5, or 0.9). With no correlation, the error associated with the PLUGIN estimate is significantly larger than the one associated with other bias correction techniques. Increasing the correlation reduces the difference between the PLUGIN and the other bias correction techniques.

on. However, a thorough investigation of the effect of properties typical of EEG-fMRI data on the accuracy of the information estimate has not previously been done. Hence, with simulations, we systematically examined the accuracy of information estimates for the multivariate distributions relevant to EEG and fMRI experiments.

4.1 Summary of Main Results. The performance of bias correction techniques depends on the underlying data statistics and the experimental design (how many feature we assume to encode the stimulus) (Panzeri et al., 2007). It is therefore important to test these approaches on simulated data with characteristics similar to the experimental data of interest. The scope of this work was to test information-based methods on simulations representing EEG-fMRI experimental designs (i.e., characterized by limited number of samples, correlation between features and different statistical properties, in terms of variance and SNR).

We found that the number of available samples, the amount of correlation in the model, and the underlying statistics of the random variables parameters would all interact with the binning and bias correction techniques. These parameters can vary depending on the EEG or fMRI feature considered. In the practice of applying information-based methods to experimental EEG-fMRI recordings, one could, during a preliminary exploratory analysis, estimate the parameters (e.g., underlying statistics or correlation) and, based on the findings, choose the most appropriate bias correction and binning strategy.

The variance of the underlying marginal distributions does not affect the error of the estimate, regardless of the BC used. This is due to the fact that changing the ratio of the variances does not affect the shape of the multivariate distribution when we are in the range of variances representative of EEG-fMRI data.

The numerosity of the sample clearly affects the estimates of the information quantities. In the low-sampling regime, the performance of BC techniques is worse when the correlation increases. The bias due to small sample size is corrected for by BC when there is no correlation. With a highly correlated small sample of data, the BC methods leave a considerable bias in the low-sampling regime because the sample cannot fully capture the underlying statistics. We also found that increasing the sample size improves the performance of the bias corrections, even when the correlation is high, which is in line with previous results on simulations reproducing invasive recordings (Panzeri et al., 2007) The numerosity may be an issue not only in task-related design but also in cases where only a limited number of events is available, for example, when dealing with epileptic spikes (Caballero-Gaudes et al., 2013).

We found an interaction between binning strategies (BS) and bias correction techniques (BC). This is explained by the fact that most BCs are sensitive to the number of empty bins, which in turn is dependent on the binning strategy of choice. The LIN binning strategy samples the full range of data as does the CEQ, but this may result in a higher number of empty bins (because with normal or gamma distributions, data at the tails are less likely to be represented, especially if the sample is small). EQP and GES are less sensitive to empty bins because they either sample the center of the data range, where data are more likely to be present even in a small sample, as

in GES, or adjust the bin width to guarantee approximately the same number of samples is counted in each bin. Although our simulations showed a consistent behavior of GES across distributions, this binning strategy may not be suited for distributions that are not gaussian, since it might capture only part of the underlying data statistics.

4.2 Summary of Limitations. We found that the binning strategy mainly affects MA and BD, generating a wider range of errors. MA needs to be in the asymptotic regime; therefore, increasing the number of points reduces the error, while in the low-sampling regime, where it is more likely to have empty bins, the error increases. In the case of BD, it might be that the parameters were not optimal for these simulated data (e.g., prior-related parameters). However, optimizing each specific method was out of the scope of this research.

In the context of information theory, correlation can be either detrimental to or beneficial for the transmission of information. It is therefore important to take the correlation into account when information quantities are estimated from the data (Abbott, 1999; Panzeri et al., 2007). Recordings from a single neuron or a neural population can show two types of correlation: a noise correlation and a signal correlation (Gawne & Richmond, 1993). The signal correlation is the correlation between two average responses to the same stimulus; noise correlation is the correlation of the trial-to-trial variability of the two signals in response to the same stimulus. In this work, we simulated correlated marginals, and this kind of correlation can be interpreted as noise correlation. We found that the error of the information estimate increases in highly correlated distributions, and this effect is worse when only a limited sample of data is available (as is usually the case in EEG-fMRI experiments). Solutions to the problem of correlation have been proposed (Panzeri et al., 2007). If the responses to the same stimulus from different neurons (modalities) are shuffled, the single trials lose their correspondence. In this way, the average stimulus response (signal) is unaffected, while the noise correlation (trial-to-trial variability) is destroyed. This step further reduces the bias, and it can be applied before any other bias correction, further decreasing the difference between the estimated and the "true" information. However, this test was out of the scope of this work, since we decided to concentrate on different bias corrections and preferred not to introduce an additional variable. Not taking into account the additional shuffling correction might explain the better performance of the PLUGIN estimate with respect to the bias-corrected estimates when the correlation is high (0.5 or 0.9) and the number of samples is small (32 or 64). A representative set of parameters for EEG-fMRI data would be, for example, 64 points and $\rho \sim 0.5$. It is therefore important to consider shuffling as an additional bias correction when dealing with EEG-fMRI recordings.

4.3 Recommendations for EEG-fMRI IT Analyses. Our results show that the bias cannot be completely removed in the parameter range typical of EEG-fMRI experiments. It is therefore important to understand the effect of the bias on the interpretation of the results.

The amount of correlation between features in simultaneous EEG-fMRI experiments is highly dependent on the features chosen (see Wong, Olafsson, Tal, & Liu, 2013, for an example of correlation between EEG and fMRI features). However, a medium correlation scenario is a sensible expectation in simultaneous EEG-fMRI, considering that the signal of interest should reflect the same underlying process but with different temporal dynamics and independent superimposed noise. A preliminary exploratory analysis of the underlying statistics of the features chosen is necessary to guide the researcher toward the most appropriate choice of binning strategy and bias correction.

As we pointed out in section 2.1, the definition of the alphabet is an important step in calculating information-based quantities. The alphabet reflects the assumptions we make on how the message (in this case the stimulus) is encoded. Deciding how many features (and modalities) represent the stimulus has an impact on the alphabet and the dictionary. One might think that increasing the number of features or modalities, or both, would increase the chances to decode the stimuli. However, this will also increase the complexity of the problem while the amount of data is limited by time constraints and the experimental design, especially in EEG-fMRI experiments. A careful choice of few informative features, based on prior knowledge or previous literature, might be more beneficial than the use of several features chosen without strong hypotheses.

When applying information theory concepts to EEG-fMRI recordings, one might want to use a fully balanced design (where each condition has the same number of trials) instead of an unbalanced design (where different conditions may have a different number of trials). Since the number of samples (i.e., trials) has an impact on the estimate of the probability distribution and the amount of bias (see Figures 3 and 4), comparing the information estimates between conditions in an unbalanced design with different amounts of bias may result in false positives or false negatives. Moreover, it might be worth considering a "sham" or neutral condition to have a baseline for the bias. If this is not possible because of constraints on the number of trials or the experiment duration, alternative ways to obtain a baseline distribution for the bias should be explored. Possibilities include simulations with simple (Ostwald et al., 2010) or more realistic models for the generation of EEG-fMRI data (Sotero & Trujillo-Barreto, 2008).

Overall our results suggest that with the range of parameters typical of EEG-fMRI experiments (i.e., small numerosity, medium correlation, and the bivariate or trivariate distributions considered), bias correction is necessary when estimating information-related quantities, and additional shuffling correction may be beneficial when the correlation is high. We found that

within binning strategies, the EQP approach gives better and more consistent results across bias correction methods and distributions, even when nongaussian marginals are considered. We also found that within bias correction techniques, TPMC, JD, and BUB give similar results, and those are consistent across distributions.

5 Conclusion

In this work, we investigated the interaction between binning strategies and bias correction techniques in a simulated framework representative of simultaneous EEG-fMRI recordings. Several choices affect the estimation of information-based quantities from experimental data. We assessed how the number of available samples, the amount of correlation in the model, and the underlying statistics of the random variables interact with the binning and bias correction techniques. We found that the interaction between the binning strategy and the estimation method influences the accuracy of the estimate. We discussed the implication of using information-based measures to analyze EEG-fMRI data and proposed how to modify the experimental design to minimize the effect of residual bias on the results of the analysis. Conditional on the assumptions of the simulations, we can now transfer our finding into the practice of experimental EEG-fMRI recordings (usually a medium correlation scenario, with only a limited sample available). However, further investigation using more realistic underlying models that better represent electrophysiological and hemodynamic data is necessary to assess the reliability of information-based analysis of EEG-fMRI data.

Historically, the application of information-theoretic approaches to study information coding and transfer within the brain has been primarily performed with invasive electrophysiological recordings. Mutual information has been used previously for fMRI experiments (de Araujo et al., 2003; Fuhrmann Alpert et al., 2008, 2007), but it is a particularly attractive proposition for integration of EEG-fMRI data given the uncertainty surrounding the linearity of EEG and fMRI responses and their interaction. Previous studies applying information-theoretic quantities to EEG-fMRI data have not estimated the accuracy of the entropy and information estimates (Caballero-Gaudes et al., 2013; Ostwald et al., 2010, 2011; Ostwald, Porcaro, Mayhew, & Bagshaw, 2012), introducing uncertainty into their interpretation. Here we have shown how to obtain, with the appropriate choice of parameters, a meaningful representation of information from noninvasive human recordings. While previous studies (Panzeri et al., 2007) have optimized the accuracy of information estimates from single-neuron recordings, such invasive data are of limited availability in human subjects, meaning that improvements in the accuracy of noninvasive techniques are needed if the links between massed neuronal activity and human behavior are to be uncovered.

Appendix: Copulas

A copula is a function that describes the dependencies between univariate variables, regardless of their respective marginals. Copulas thus allow us to model multivariate correlated data with arbitrary marginal distributions with a given correlation structure. The theory of copulas is based on two concepts: Sklar's theorem and the method of inversion. The following description closely follows Meucci (2011).

Let X be an arbitrary univariate random variable with probability density function (pdf) f_X and its corresponding cumulative density function (cdf) F_X, such that

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(z)dz. \tag{A.1}$$

By transforming X thorugh F_X, we obtain the grade of X, which is uniformly distributed on the unit interval, regardless the original pdf f_X (see Meucci, 2011, for a proof):

$$U \equiv F_X(x) \text{ with } U \sim U(0,1). \tag{A.2}$$

Vice versa, given $U \sim U(0,1)$, we have

$$X \equiv F_X^{-1}(U) \text{ with } X \sim f_X, \tag{A.3}$$

where F_X^{-1} is the inverse cdf of X. Equations A.2 and A.3 form the so-called method of inversion, which states that a random variable with an arbitrary pdf f_X can be transformed into a uniform random variable by passing it into its cdf) F_X. Vice versa, we can transform a uniform random variable into a variable with an arbitrary pdf by feeding it into its corresponding inverse cdf F_X^{-1}.

These results are readily extended to the multivariate case. Let $X = \{X_1, \ldots, X_n, \ldots, X_N\}$ be an N-dimensional random variable with joint pdf f_X and marginals f_{X_n}. Let F_X and F_X^{-1}be the cdf and inverse cdf of X, respectively, and F_{X_n} and $F_{X_n}^{-1}$ be the marginal cdf and inverse cdf of X_n, respectively.

By feeding each marginal into its corresponding cdf F_{X_n}, we obtain a set of uniformly distributed variables U_n, that is, the grades of X:

$$U_n \equiv F_{X_n}(x_n) \text{ with } U_n \sim U(0,1). \tag{A.4}$$

It is important to note that U_n are not independent, but they have the same degree of dependency as the original variables X_n. Therefore the joint distribution f_U associated with U is not uniformly distributed, and it is called

copula C. The copula can be seen as the missing information necessary to go from the marginals to the joint distribution.

Sklar's theorem states that the pdf of the copula can be obtained from the joint and the marginals distributions using the inversion method:

$$f_X\left(F_{X_1}^{-1}(u_1),\ldots F_{X_N}^{-1}(u_N)\right) = f_U(u_1,\ldots u_N) \times f_{X_1}\left(F_{X_1}^{-1}(u_1)\right) \times \ldots \times f_{X_N}\left(F_{X_N}^{-1}(u_N)\right). \quad \text{(A.5)}$$

In practice, we generated samples drawn from multivariate distributions in three steps:

1. We simulated multivariate (bi- and trivariate) gaussian distributions with a given correlation structure.
2. We fed the generated data into a gaussian cdf, obtaining a set of nonindependent uniformly distributed variables (the grades).
3. We fed the grades into the inverse cdf of the desired marginals (uniform, normal, gamma) to obtain multivariate distributions with given marginals and correlation structure.

It is worth noticing that the choice of multivariate gaussian as the starting point for the data generation defines the type of copula (i.e., joint dependence) that we assumed between variables. Other choices are available (e.g., the Student t-distribution); however, the main difference between copulas resides in the degree of dependency in the tails of the copula (i.e., the most extreme cases) (Schmidt, 2007; Venter, 2002). In the case of EEG-fMRI data, we believe the strongest association to be in the most likely cases (the center of the copula) and therefore decided to use a gaussian copula.

Acknowledgments

Support for this research was provided by the Dr. Hadwen Trust for Humane Research, the leading U.K. medical research charity that funds and promotes exclusively human-relevant research that encourages the progress of medicine with the replacement of the use of animals in research.

References

Abbott, L. F. (1999). The effect of correlated variability on the accuracy of a population code, *101*, 91–101.

Bagshaw, A. P., & Warbrick, T. (2007). Single trial variability of EEG and fMRI responses to visual stimuli. *NeuroImage*, *38*, 280–292.

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

Caballero-Gaudes, C., Van de Ville, D., Grouiller, F., Thornton, R., Lemieux, L., Seeck, M., . . . Vulliemoz, S. (2013). Mapping interictal epileptic discharges using mutual information between concurrent EEG and fMRI. *NeuroImage*, *68*, 248–262.

Carlton, A. G. (1969). On the bias of information estimates. *Psychological Bulletin*, *71*, 108–109.

Chao, A., & Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, *10*, 429–443.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. New York: Wiley Interscience.

de Araujo, D. B., Tedeschi, W., Santos, A. C., Elias, J., Neves, U. P. C., & Baffa, O. (2003). Shannon entropy applied to the analysis of event-related fMRI time series. *NeuroImage*, *20*, 311–317.

Debener, S., Ullsramon Siegel, M., Fiehler, K., von Cramon, D. Y., & Engel, A. K. (2005). Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *J. Neurosci.*, *25*, 11730–11737.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1–26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Eichele, T., Specht, K., Moosmann, M., Jongsma, M. L. a, Quiroga, R. Q., Nordby, H., & Hugdahl, K. (2005). Assessing the spatiotemporal evolution of neuronal activation with single-trial event-related potentials and functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 17798–17803.

Fuhrmann Alpert, G., Hein, G., Tsai, N., Naumer, M. J., & Knight, R. T. (2008). Temporal characteristics of audiovisual information processing. *Journal of neuroscience*, *28*, 5344–5349.

Fuhrmann Alpert, G., Sun, F. T., Handwerker, D., D'Esposito, M., & Knight, R. T. (2007). Spatio-temporal information analysis of event-related BOLD responses. *NeuroImage*, *34*, 1545–1561.

Gawne, T. J., & Richmond, B. J. (1993). How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.*, *13*, 2758–2771.

Goldberg, D. H., Victor, J. D., Gardner, E. P., & Gardner, D. (2009). Spike train analysis toolkit: Enabling wider application of information-theoretic techniques to neurophysiology. *Neuroinformatics*, *7*, 165–178.

Goldman, R. I., Wei, C.-Y., Philiastides, M. G., Gerson, A. D., Friedman, D., Brown, T. R., & Sajda, P. (2009). Single-trial discrimination for integrating simultaneous EEG and fMRI: Identifying cortical areas contributing to trial-to-trial variability in the auditory oddball task. *NeuroImage*, *47*, 136–147.

Herrmann, C. S., & Debener, S. (2008). Simultaneous recording of EEG and BOLD responses: A historical perspective. *International Journal of Psychophysiology*, *67*, 161–168.

Hu, L., Liang, M., Mouraux, A., Wise, R. G., Hu, Y., & Iannetti, G. D. (2011). Taking into account latency, amplitude, and morphology: Improved estimation of

single-trial ERPs by wavelet filtering and multiple linear regression. *Journal of Neurophysiology*, *106*, 3216–3229.

Ma, S. (1981). Calculation of entropy from data of motion. *Journal of Statistical Physics*, *26*, 221–240.

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 Neuroscience*, *10*, 81.

Meucci, A. (2011). A short, comprehensive, practical guide to copulas. Social Science Research Network working paper series.

Miller, G. A. (1955). Note on the bias of information estimates. In H. Quastler (Ed.), *Information theory in psychology: Problems and methods* (pp. 95–100). Glencoe, IL: Free Press.

Mulert, C., & Lemieux, L. (Eds.). (2010). *EEG-fMRI physiological basis, technique, and application*. Berlin: Springer-Verlag.

Ojemann, G. A., Ojemann, J., & Ramsey, N. F. (2013). Relation between functional magnetic resonance imaging (fMRI) and single neuron, local field potential (LFP) and electrocorticography (ECoG) activity in human cortex. *Frontiers in Human Neuroscience*, *7*, 34.

Optican, L. M., Gawne, T. J., Richmond, B. J., & Joseph, P. J. (1991). Unbiased measures of transmitted information and channel capacity from multivariate neuronal data. *Biol. Cybern.*, *65*, 305–310.

Ostwald, D., & Bagshaw, A. P. (2011). Information theoretic approaches to functional neuroimaging. *Magnetic Resonance Imaging*, *29*, 1417–1428.

Ostwald, D., Porcaro, C., & Bagshaw, A. P. (2010). An information theoretic approach to EEG-fMRI integration of visually evoked responses. *NeuroImage*, *49*, 498–516.

Ostwald, D., Porcaro, C., & Bagshaw, A. P. (2011). Voxel-wise information theoretic EEG-fMRI feature integration. *NeuroImage*, *55*, 1270–1286.

Ostwald, D., Porcaro, C., Mayhew, S. D., & Bagshaw, A. P. (2012). EEG-fMRI based information theoretic characterization of the human perceptual decision system. *PloS one*, *7*, e33896.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, *15*, 1191–1253.

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. *Magnetic Resonance Imaging*, *26*, 1015–1025.

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. *Neural Computation in Neural Systems*, *7*, 87–107.

Pouliot, P., Tremblay, J., Robert, M., Vannasing, P., Lepore, F., Lassonde, M., ... Lesage, F. (2012). Nonlinear hemodynamic responses in human epilepsy: A multimodal analysis with fNIRS-EEG and fMRI-EEG. *Journal of Neuroscience Methods*, *204*, 326–340.

Quian Quiroga, R., & Panzeri, S. (2009). Extracting information from neuronal populations: Information theory and decoding approaches. *Nature reviews. Neuroscience*, *10*, 173–185.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1996). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Ritter, P., & Villringer, A. (2006). Simultaneous {EEG}-f{MRI}. *Neuroscience and Biobehavioral Reviews*, *30*, 823–838.

Schmidt. (2007). Coping with copulas. In J. Rank (Ed.), *Copulas: From theory to application in finance*. London: Risk Books.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Sotero, R. C., & Trujillo-Barreto, N. J. (2008). Biophysical model for integrating neuronal activity, EEG, fMRI and metabolism. *NeuroImage*, *39*, 290–309.

Strong, S. P., Koberle, R., de Ruyter van Steveninck, R., & Bialek, W. (1998). *Physical Review Letters*, *80*, 197–200.

Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Computation*, *7*, 399–407.

Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, *2*, 1–25.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review*, *7*, 424–465.

Venter, G. G. (2002). Tails of copulas. *Proceedings of the Casualty Actuarial Society*, *89*, 68–113.

Victor, J. D. (2006). Approaches to information-theoretic analysis of neural activity. *Biological Theory*, *1*, 302–316.

Vulliemoz, S., Lemieux, L., Daunizeau, J., Michel, C. M., & Duncan, J. S. (2010). The combination of EEG source imaging and EEG-correlated functional MRI to map epileptic networks. *Epilepsia*, *51*, 491–505.

Wolpert, D. H., & Wolf, D. R. (1995). Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E Stat. Phys.: Plasmas Fluids Relat Interdiscip Topics*, *52*, 6841–6854.

Wong, C. W., Olafsson, V., Tal, O., & Liu, T. T. (2013). The amplitude of the resting-state fMRI global signal is related to EEG vigilance measures. *NeuroImage*, *83*, 983–90.

Yeśsilyurt, B., Uğurbil, K., & Uludazğ, K. (2008). Dynamics and nonlinearities of the BOLD response at very short stimulus durations. *Magn. Reson. Imaging*, *26*, 853–862.

Zhang, N., Zhu, X.-H., & Chen, W. (2008). Investigating the source of BOLD nonlinearity in human visual cortex in response to paired visual stimuli. *Neuroimage*, *43*, 204–212.

Received October 18, 2013; accepted June 27, 2014.