An information theoretic approach to EEG–fMRI integration of visually evoked responses

Dirk Ostwald *, Camillo Porcaro, Andrew P. Bagshaw

School of Psychology, University of Birmingham, UK
Birmingham University Imaging Centre, University of Birmingham, UK

Abstract

The integration of signals from electro-encephalography (EEG) and functional magnetic resonance imaging (fMRI), acquired simultaneously from the same observer, holds great potential for the elucidation of the neurobiological underpinnings of human brain function. However, the most appropriate way in which to combine the data in order to achieve this goal is not clear. Here, we apply a novel route to the integration of simultaneously acquired multimodal brain imaging data. We adopt a theoretical framework developed in the study of neuronal population codes which explicitly takes into account the experimentally observed stimulus–response signal probability distributions using the concept of mutual information. We study the implications of this framework using simulated data sets generated from a set of linear Gaussian models, and apply the framework to EEG–fMRI data acquired during checkerboard stimulation of low and high contrast. We focus our evaluation on single-trial time-domain signal features from both modalities and provide evidence for the informativeness of a subset of these features with respect to the stimulus and each other. Specifically, the framework was able to identify the contrast dependency of the haemodynamic response and the P100 peak of the visual evoked potential, and showed that combining EEG and fMRI time-domain features by quantifying the information in their joint distribution was more informative than treating each one in isolation. In addition, the effect of different pre-processing strategies for EEG–fMRI data can be assessed quantitatively, indicating the improvements to be gained by more advanced methods. We conclude that the information theoretic framework is a promising methodology to quantify the relative importance of different response features in neural coding and neurovascular coupling, as well as the success of data pre-processing strategies.

© 2009 Elsevier Inc. All rights reserved.

Introduction

The integration of signals from electro-encephalography (EEG) and functional magnetic resonance imaging (fMRI), acquired simultaneously from the same observer, holds great potential for the elucidation of the neurobiological underpinnings of human brain function. However, the most appropriate way in which to combine the data in order to achieve this goal is not clear. The majority of previous EEG–fMRI studies have used aspects of the EEG time-series to inform fMRI data analysis within the framework of the general linear model (Debener et al., 2005; Eichele et al., 2005; Goldman et al., 2009; Philastides and Sajda, 2007) while others have used the opposite strategy of informing EEG source analysis using spatial constraints extracted from fMRI data (Dale and Halgren, 2001; Esposito et al., 2009, in press). Some recent developments have been directed towards the generation of common forward models for both modalities to be applied to the bimodal space–time-series (Brookings et al., 2009; Daunizeau et al., 2007; Kilner et al., 2005; Sotero and Trujillo-Barreto, 2008; Valdes-Sosa et al., in press).

Here, we apply an alternative route to the integration of simultaneously acquired multimodal brain imaging data. We adopt a theoretical framework developed in the study of neuronal population codes which explicitly takes into account the experimentally observed stimulus–response signal probability distributions (commonly referred to in the EEG/fMRI literature as ‘single-trial variability’ (Bagshaw and Warbrick, 2007; Debener et al., 2006; Herrmann and Debener, 2008; Valdes-Sosa et al., in press)) using the concept of mutual information developed in the mathematical theory of communication (Shannon, 1948). This framework allows the quantitative evaluation of various quantities which lend themselves to interpretations of stimulus–signal and signal–signal relationships, while also relaxing the linearity and Gaussian constraints embodied in more traditional quantities such as linear correlation and regression.

In the current study, we investigate the information content of bimodal time-domain signal features extracted from occipital cortex during visual checkerboard stimulation, aiming to obtain insights into the following questions (Panzeri et al., 2008):

• How much information do different unimodal brain imaging signal features convey about the stimulus?
• How much information does the joint observation of features from both modalities convey about the stimulus?
• How much information does one modality feature convey about features of the other modality?
• Lastly, how do different data pre-processing strategies affect the resulting information estimates?

This report comprises three parts: The first part and Appendices A and B review the underlying theory and some practical caveats from a theoretical perspective. In the second part and Appendix C we present the application of the information theoretic framework to artificial data sets to validate our implementation, gain insight into the behaviour of the various quantities under controlled response signal conditions and estimate systematic biases. In the third part we present the results of an application of the information theoretic framework to simultaneously acquired EEG–fMRI data in a visual stimulation (checkerboard) experiment. All custom written Matlab (The Mathworks, Natick, MA) code used in this report is available from http://www.buic.bham.ac.uk/downloads/EEG_FMRI_ITQ/EEG_FMRI_ITQ_Analysis.zip and the experimental data is available upon request from the corresponding author.

Materials and methods

Theoretical framework

We adopt an information theoretic framework that has its origins in the mathematical study of communication and has been previously applied to the problem of neural coding in the context of invasive electrophysiological recordings (Belitski et al., 2008; Panzeri et al., 2008; Schneidman et al., 2003) and fMRI data (Pessoa and Padmala, 2007; Fuhrmann Alpert et al., 2007; Rolls et al., 2009). A detailed review of the theory underlying the information theoretic framework is given in Appendices A and B.

Intuitively, the term ‘informative signal’ can be understood as ‘discriminative signal’: A signal is said to be informative if knowledge about it allows a reliable prediction of its cause (e.g. an external stimulus or a signal feature of another modality). The information theoretic framework allows the definition of the following quantities which are of interest in the context of multimodal brain imaging experiments (see also Table 1): (1) The information conveyed by unimodal signals about the stimulus \( I(S;R_i) \), (2) the information conveyed by signals of both modalities about the stimulus \( (I(S;R_1, R_2)) \) and its dependence on the respective unimodal information measures (Synergy), and (3) the information conveyed by a signal of one modality about a signal of the other modality, both in the presence \( (I(R_1; R_2 | S)) \) and the absence \( (I(R_1; R_2 | S)) \) of external stimulation.

From the practical perspective of estimating these quantities from a limited experimental data set, the question of bias estimation is of relevance. In brief, the fact that the joint stimulus–response signal probability distribution \( p(s,r_1,r_2) \) is not known, but can only be estimated from a finite data set as the observed frequency distribution \( p_d(s,r_1,r_2) \) by means of a histogram analysis, leads to a bias in the computed entropies resulting in an over-estimation of information content. A variety of methods to amend these biases has been devised in the application of the information theoretic framework to neurophysiological data (Panzeri et al., 2007). We are not aware of any theoretical or experimental developments that would favour one bias correction scheme over another for the case of multimodal brain imaging data, hence we chose bias correction methods of relatively low complexity (PT correction, shuffling correction), and according to the recommendations of Panzeri et al. (2007) supplement these by analyses of simulated data. Optimal bias correction for multimodal brain imaging experiments will need to be addressed in future research.

Simulations

To validate our implementation of the information theoretic framework, constrain the analysis of the experimental data with respect to free data analysis parameters, as well as to study the behaviour of the various information theoretic quantities under controlled response signal conditions, we performed a number of analyses on simulated data sets. These data sets were derived from four simple generative linear Gaussian models. We refer to these models as M0, M1, M2 and M3. Graphical representations of these models are given in Fig. 1A.

Each model consists of a ‘stimulus’ variable \( S \) and two ‘response’ variables, \( R_1 \) and \( R_2 \), simulating response signal features from both imaging modalities. As in the experimental context, the stimulus variable distribution is under the control of the experimenter and hence deterministic and uniform, while the response variables depend on the state of the stimulus variable in different ways, depending on the particular model. In this section we only provide an overview of the implemented models, the detailed specifications are given in Appendix C.

Model M0 is a ‘reference’ or Gaussian null model. For M0 the response variables are independent of the stimulus variable, and of each other. This refers to a case in which the observed EEG and fMRI signal features are not related and non-informative with respect to the stimulus. Model M1 captures a scenario in which both response variables depend on the stimulus, but are independent of one another. This refers to a case in which both EEG and fMRI signal features are modulated by the stimulus, but they do not co-vary with one another. Model M2 captures a scenario in which response variable \( R_1 \) (e.g. an EEG signal feature) co-varies with the stimulus, while the other response variable \( R_2 \) (e.g. an fMRI signal feature) is dependent on the stimulus only indirectly via the first variable. Finally, Model M3 represents an intermediate version of models M1 and M2 in which the second response variable \( R_2 \) (e.g. an fMRI signal feature) is sensitive to both changes in the ‘stimulus’ variable \( S \) and the first response variable \( R_1 \), which in terms of simultaneous EEG–fMRI experiments is possibly the most plausible assumption. It should be noted that models M1 and M2 represent special cases of model M3, which is implemented as a weighted linear interpolation between models M1 and M2.

In the simulations, we varied the following parameters for each response variable \( R_1 \) and \( R_2 \): a) multiplicative gain factors \( a_{g1} \) and \( a_{g2} \) which amplify the respective response variable input, b) deterministic biases \( b_{r1} \) and \( b_{r2} \) which are added to the respective amplified response variable input and c) random additive error terms which are governed by variance terms \( \sigma_{r1}^2 \) and \( \sigma_{r2}^2 \). Sampling from the models is
implemented in the respective .m files M0.m, M1.m, M2.m and M3.m. To sample from a bivariate Gaussian distribution we make use of the gsamp.m function of the netlab toolbox (Nabney, 2002).

Visual stimulation experiment

Subjects

Fourteen subjects (6 male, 8 female, mean age 27.2, range 19–34 years) were recruited from the University of Birmingham campus and paid for their participation. All observers had normal or corrected to normal vision, no history of neurological disorders and gave written informed consent. The study was approved by the local ethics committee.

Stimuli

Left hemi-field reversing checkerboards were presented at a spatial frequency of 2 cycles per degree of visual angle at two different contrast levels (i.e. the size of the stimulus set |S| was 2), high (CMichelson = 1) and low (CMichelson = 0.25). Stimuli were presented with a central fixation cross.

Design and procedure

Individual trials (Fig. 1B) of the experiment consisted of a single presentation of the checkerboard stimulus for 1 s with phase reversal after 500 ms followed by a fixation period which was uniformly sampled from 16.5 to 21 s, discretised to 1.5 s (MR repetition time, see below). Individual sessions consisted of 17 trials per condition as well as fixation periods at the beginning (8 s) and at the end (variable), amounting to a total session length of 441 volumes × 1.5 s, i.e. 11 min. For each observer, we acquired data during experimental sessions, yielding Ns = 85 trials per condition and observer. The observer was asked to perform a simple task in order to maintain attention: on a random selection of half the trials of a given session, the fixation cross changed from a plus sign to an × during the fixation period at random time points, discretely (1.5 s) and uniformly sampled from the interval of 4.5–16.5 s after stimulus onset. The task was to report the change in fixation by a button press using the index finger of the right hand. Hit rate and number of false alarms were presented to the observer at the end of each session. Stimuli were presented and behavioural data collected using Psychotoolbox3 (Brainard, 1997; Pelli, 1997) for Matlab. Stimulus presentation timing was controlled by the MRI scanner volume trigger. For details on stimulus creation and presentation see CheckerBoardScanTrig.m.

MRI data acquisition

The experiment was conducted at the Birmingham University Imaging Centre using a 3 T Philips Achieva MRI scanner. An initial T1-weighted anatomical (1 mm isotropic voxel) and T2*-weighted functional data were collected with an eight channel phased array SENSE head coil. EPI data (gradient echo-pulse sequence) were acquired from 20 slices (2.5 × 2.5 × 3 mm resolution, TR 1500 ms, TE 35 ms, SENSE factor 2, flip angle 80°, with equidistant temporal slice spacing to facilitate synchronisation of the EEG clock (see below)), providing approximately half brain coverage in the dorsal–ventral direction. Slices were oriented parallel to the AC–PC axis of the observer’s brain and positioned to cover the entire occipital cortex.

EEG data acquisition

EEG data were recorded using a 64 channel MR compatible EEG system (BrainAmp MR Plus, Brain Products, Munich, Germany), which incorporates current limiting resistors of 5 kΩ at the amplifier input and in each electrode. The EEG cap consisted of 62 scalp electrodes, one of which was attached approximately 2 cm below the left collarbone for recording the ECG, while the other was attached below the left eye (on the lower orbital portion of the orbicularis oculi muscle) for detection of eyelid artefacts. Data were sampled at 5000 Hz. Impedance at all recording electrodes was less than 20 kΩ. The EEG data acquisition setup clock was synchronised with the MRI scanner clock using Brain Product’s SyncBox, resulting in exactly 7500 data points per EPI-TR interval.

Fig. 1. (A) Graphical models used in the simulations. In the framework of probabilistic graphical models, each node (circle) represents a random variable, while the links (arrows) express probabilistic relationships between these variables. aR1, aR2, aS1, aS2 and t refer to the model parameters detailed in Appendix C. For linear Gaussian models, the distributions of individual variables as well as the distribution over all variables are Gaussian. In the current application only the response variables R1 and R2 are Gaussian, while the stimulus distribution S is uniform. (B) Single-trial experimental design of the visual stimulation experiment. A hemi-field checkerboard of either low or high contrast was presented at time 0 and reversed after 500 ms. After another 500 ms the checkerboard was removed. At a subset of the trials, the fixation cross changed to an × at the onset of a randomly sampled EPI volume acquisition, and the observer was asked to indicate this change by a button press.
Data pre-processing

Raw EPI data (.par/.rec format) were converted to Niftii format using MRICroN’s dcmm2nii utility (Rorden and Brett, 2000). SPM5 (Friston, 2007) was used for fMRI data pre-processing. Data pre-processing involved anatomical realignment, slice scan time correction (reference slice 10), re-interpolation to 2 × 2 × 2 mm voxels, anatomical normalization and spatial smoothing (5 mm Gaussian kernel, for details see epi_preprocess_spm5.m). We next modelled the experimental data of each individual voxel using a standard GLM approach (Friston et al., 1994). Voxel time-courses were modelled in an event-related fashion using regressors obtained by convolving each stimulus onset impulse with a canonical haemodynamic response function and its first temporal derivative. Additional nuisance covariates included the realignment parameters to account for residual motion artefacts (for details, see epi_glm_spm5.m). After model estimation, we identified voxels displaying a stimulus (S1 and S2) correlated main effect, and extracted filtered and whitened data from a sphere of 5 mm radius centred on the global maximum of the resulting statistical parametric map. This was done individually for each observer and the data were averaged over the voxels included in the sphere. In the following, this data is referred to as ‘advanced pre-processing’. In addition, in order to examine the effect of different pre-processing strategies on information theoretic quantities, we extracted voxel average data from a 5 mm radius sphere centered at the reported MNI coordinates of right primary visual cortex on the MNI coordinates [6 − 82 − 4] (Hasnain et al., 1998; Wohlschlager et al., 2005) This data is referred to as ‘basic pre-processing’.

Raw EEG data were partitioned into data acquisition sessions and exported to .mat format using Brain Vision Analyzer (Brain Products, Munich, Germany). Subsequent pre-processing steps involved gradient artefact removal, pulse artefact correction (Optimal Basis Set method (Allen et al., 1998, 2000; Niazy et al., 2005)) as implemented as a plug-in to EEGLAB (Delorme and Makeig, 2004), low-pass filtering at 25 Hz, and down-sampling to 500 Hz.

Similarly to the advanced and basic pre-processing strategies adopted for the fMRI data, ‘advanced pre-processing’ EEG data underwent additional artefact removal. Specifically, after concatenation of the runs of an individual observer, a semiautomatic ICA-based procedure (Barbati et al., 2004; Porcaro et al., 2009) was used to identify artefactual non-cerebral components, i.e. eye-movements, residual BCG artefacts, and environmental noise. The final data were obtained by reconstruction from all non-artefactual components. Data were then extracted from the occipital electrode that showed the highest amplitude P100 in the average VEP (selected from electrodes POz, Oz, P04, P08 and O2). The ‘basic pre-processing’ EEG data were extracted from the same electrode omitting the ICA step (i.e. only gradient and BCG artefacts were removed).

Response signal feature extraction

For both modalities we focussed on a set of single-trial time-domain features (Table 2) which we extracted from the EEG electrode or fMRI sphere using both basic and advanced pre-processing strategies as described above. For both modalities, we extracted a signal feature value from each trial of each condition, yielding a matrix of 85 × 2 signal feature values for each modality, which then entered the information theoretic analysis as the sample.

For the EEG modality we focussed on three sets of features, which we group into amplitude features, latency features and root mean square (RMS) features. For the amplitude features we focussed on traditional visual evoked potential (VEP) peaks i.e. the N70, P100 and N140 for both stimulus onset and averaged over the two VEPs generated upon stimulus onset and reversal. In addition, we extracted the amplitude difference P100–N140 for stimulus onset and averaged over stimulus onset and reversal. For the latency features, we extracted the time of the N70, P100, and N140 peaks with respect to stimulus onset from windows of 40 ms centred on the respective peaks. Lastly, we computed as a measure of the overall power of the single-trial evoked potential the RMS from a post-stimulus onset window of 50–250 ms, from a post-stimulus reversal window of 50–250 ms, and across the entire stimulus presentation period of 0–1000 ms.

For the fMRI modality we focussed on a set of similar features of the evoked single-trial haemodynamic response functions, i.e., we extracted the maximum amplitude in BOLD percent signal change with respect to the pre-stimulus baseline (average of the three pre-stimulus onset TR values) and its latency with respect to stimulus onset in seconds. In addition we extracted the post-stimulus RMS as a measure of the overall magnitude of the evoked haemodynamic response.

For a complete overview of the extracted features as well as their definition, refer to Table 2. Feature extraction was implemented in the respective .m files extract_hrf_features.m and extract_evp_features.m.

Information theoretic analysis

We applied the same analysis scheme for both simulated and experimental data sets. The analysis is implemented in the functions ITQ_bivariate.m (simulations), and ITQ_bivariate_analysis.m (experimental data) and fully described in Appendices A and B.

In brief, after obtaining a data sample, the first step towards an information theoretic analysis is to compute the observed joint frequency distribution, denoted as \( p_{O}(s_{1},r_{1},f_{2}) \), where the subscript \( N \) indicates the dependency on the number of samples. The observed joint frequency distribution is obtained by applying a grid of response value bins to the observed data, counting the samples that fall into each of the response bins and then normalizing by the total number of
response bins such that \( \sum_{s, r_1} \sum_{r_2} \sum_{t} p(\text{stim} s, r_1, r_2) = 1 \), where \( [S] \) indicates the (discrete) stimulus set, and \( R_1 \) and \( R_2 \) the (discrete) response sets, of each variable, spanning the (discrete) response set \( R_1 \times R_2 \).

Free parameters in the application of the response bin grid are the total number of response bins and the upper and lower limit of each response bin in each dimension.

To make efficient use of all sampled data points, the upper and lower limits of the response grid were set to the maximum and minimum value of each response variable. We chose the number of response bins to be approximately equal to half the number of trials per stimulus \( N_0 \) (Panzeri et al., 2007; Pola et al., 2003). Hence for each response dimension, the number of bins was set to 0.65*floor(\( \sqrt{N_0} \))≈6, resulting in 36 response bins for the bivariate response variable distributions. The choice of a specific bin size results in a trade-off between good bias control and sensitivity to information contained in the data. We validated our choice of the number of bins based on simulation results discussed in the results section. The histogram analysis is implemented in compute_joint_dist.m.

From the joint observed frequency distribution, the marginal distributions, \( p(\text{stim} s, r_1) \), \( p(\text{stim} s, r_2) \), and \( p(t) \), and their corresponding stimulus conditional distributions, \( p(r_1 | \text{stim} s) \), \( p(r_2 | \text{stim} s) \), \( p(t | \text{stim} s) \), can be obtained by summation and division, respectively. Based on these observed frequency distributions, the information theoretic quantities introduced above and detailed in Appendix A can be computed by applying the respective formulae. These raw values are referred to as ‘Plug-In’ quantities (Panzeri et al., 2007). To correct for systematic biases we then used the bayescount.m implementation of the PT-correction scheme (http://stefano.panzeri.googlepages.com/informationbiascorrections) for the case of the univariate information theoretic quantities \( I(S; R_1) \), \( I(S; R_2) \), \( I(R_1; R_2) \), and \( I(R_1; R_2 | S) \). To correct for the systematic bias in the case of the bivariate information theoretic quantity \( I(S; R_1, R_2) \) we used a shuffling approach implemented in the main analysis routine (1000 permutations). Note that in the case of the simulations we report un-normalised Synergy measures to obtain results at the same scale as the remaining quantities.

Results

Simulated data

To validate our implementation of the information theoretic framework, constrain the analysis of the experimental data with respect to free data analysis parameters, as well as to study the behaviour of the various information theoretic quantities under controlled response signal conditions, we performed a number of analyses on simulated data sets. Importantly, we performed all simulations using the parameters of the visual stimulation experiment, i.e. we used a stimulus set \( S \) of size \( |S| = 2 \), with \( s_1 = 1 \) (‘low contrast’) and \( s_2 = 2 \) (‘high contrast’), with \( N_0 = 85 \) trials per stimulus condition, unless otherwise stated.

Here we report the results of three of those simulations (see Supplementary material for additional simulations): 1) the analysis of a Gaussian null model (M0), which is of experimental relevance with respect to the implemented bias estimation and correction methods (M0_simulations.m), 2) the analysis of three linear Gaussian models implementing common assumptions about stimulus–EEG–fMRI interdependencies (M1, M2, M3) in order to demonstrate the differentiating power of the framework (M1_M2_M3_simulations.m), and 3) a validation of the selected number of response bins for the histogram analysis (BinNum_simulations.m).

We first note that for two equally distributed stimuli, which each occur with \( p(s_i) = 0.5 \), the entropy of the stimulus probability distribution is given by

\[
H(S) = \frac{1}{2} \log_2(2) + \frac{1}{2} \log_2(2) = 1.
\]

As the response distribution (both in the case of artificial and experimental data) will show larger variability and hence larger entropies, the mutual information between stimulus and responses is therefore bounded from above by 1 bit (Cover and Thomas, 2006; Schneidman et al., 2003).

MO simulations

In model M0, the response variables are independent of the stimulus and of each other. For an unbiased estimate of the respective information theoretic quantities one hence expects a value of zero, indicating no information content. Fig. 2A displays the results for 100 model realizations of M0 with parameters \( \sigma_{s1} = \sigma_{s2} = b_{l1} = b_{l2} = \sigma_{R1} = \sigma_{R2} = 1 \) for the ‘Plug-In’ (upper panel) and bias-corrected (lower panel) versions of the information theoretic quantities. All quantities are positively biased in the ‘Plug-In’ case, and except for the quantity \( I(N; S; R_1, R_2) \) are reduced to approximately zero after bias correction, with the largest variability displayed by \( I(N; R_1; R_2 | S) \). For \( I(N; S; R_1, R_2) \) and the closely related Synergy the shuffling bias correction results in a bias overcorrection of roughly −0.1 bits. Further simulations revealed that this pattern was independent of the model parameter settings (Fig. S1). However, we observed that the quality of the bias correction is a function of the experimental setting, i.e. the number of stimuli used and the number of repeats per stimulus (Fig. S2). Nonetheless, for independent Gaussian response variable distributions sampled

![Figure 2](image-url)

**Fig. 2.** (A) ‘Plug-In’ (upper panel) and bias-corrected (lower panel) estimates of the information theoretic quantities obtained from model M0 in 100 simulations with parameter settings \( \sigma_{s1} = \sigma_{s2} = b_{l1} = b_{l2} = \sigma_{R1} = \sigma_{R2} = 1 \). (B) Bias-corrected estimates of the information theoretic quantities obtained from 1000 simulations of model M0 with parameter settings as in (A). The error bars indicate the 0.999 confidence intervals for these point estimates.
twice at a sample size of 85, the bias correction for $I_S(S; R_1, R_2)$ was found to be too strong and improvable for $I(R_1; R_2)$ and $\langle I_S(R_1; R_2|S) \rangle_S$. To amend the observed (negative) biases for the Gaussian Null model in our experimental setting, we estimated the remaining bias magnitude after bias correction ($\Delta \text{Bias}$) for each information theoretic quantity from 1000 simulations of the model with $a_{R_1} = a_{R_2} = b_{R_1} = b_{R_2} = \sigma^2_{R_1} = \sigma^2_{R_2} = 1$. The resulting point estimates are shown in Fig. 2B together with their estimated 0.999 confidence intervals ($t_\alpha = 3.29$). These point estimates were added to the bias-corrected information theoretic quantities in all following analyses (simulated and experimental data).

![Variation of multiplicative gain factors $a_{R_1}$ and $a_{R_2}$ in models M1, M2 and M3. For each panel 100 simulations in the respective parameter space were carried out. (A) Variation of $a_{R_1}$ only. (B) Variation of $a_{R_2}$ only. (C) parallel variation of $a_{R_1}$ and $a_{R_2}$.](image)
M1–M3 coupling coefficient simulations

Fig. 3 shows the results of varying the multiplicative constants $a_{R1}$ and $a_{R2}$ individually and in parallel in the interval [1,10]. In model M1, an increase in $a_{R1}$ leads to an increase in both $I_{R}(S; R_1)$ and $I_{R}(S; R_1, R_2)$ to the boundary level of 1 bit, as the corresponding observed frequency distributions for stimulus 1 and stimulus 2 become non-overlapping. $I_{R}(S; R_2)$ and $I_{R}(R_1; R_2)$ are virtually zero, and not affected by variation of $a_{R2}$. As the multiplicative factor at the level of $R_1$ outweighs the level of variance at the level of $R_2$, the variables become slightly redundant, as indicated by the negative Synergy. In model M2, an increase of parameter $a_{R1}$ leads to an increase of the information about the stimulus contained in both response variables in isolation ($I_{R}(S; R_1)$ and $I_{R}(S; R_2)$), and combined ($I_{R}(S; R_1, R_2)$). Most notably the activity dependence $I_{R}(R_1; R_2)$ reaches a maximum of 1.1 bits, as the response variables become fully redundant. In model M3, variation of parameter $a_{R1}$ corresponds to a modulation of the connection from the stimulus variable $S$ to response variable $R_1$, while the connections between the stimulus variable and response variable $R_2$ and between both response variables remain constant. Correspondingly, $I_{R}(S; R_1)$ and $I_{R}(S; R_1, R_2)$ increase to the boundary level as in the simulation of M1, while in addition, $I_{R}(S; R_2)$ and $I_{R}(R_1; R_2)$ increase with larger parameter values of $a_{R1}$ reflecting the introduced link between the two response variables. It is worth noting, that the response variable co-variation is purely stimulus induced, i.e. $I_{R}(R_1; R_2) | S \rangle_{1/2}$ remains close to zero.

In model M1 variation of parameter $a_{R2}$ evokes the complementary information theoretic signature with respect to variation of $a_{R1}$: here, $I_{R}(S; R_2)$ increases with $I_{R}(S; R_1, R_2)$ while the other quantities remain largely unaffected (only the Synergy again decreases slightly). Varying $a_{R2}$ in model M2 reveals a different picture: The information contained in the response variables with respect to the stimulus is unaffected (i.e. $I_{R}(S; R_1)$ and $I_{R}(S; R_2)$ and $I_{R}(R_1; R_2)$ do not vary with $a_{R2}$) while activity and conditional dependence rise in parallel, indicating the strong amplification of the random noise component of $R_1$ by $R_2$. Finally, varying $a_{R2}$ in model M3 corresponds to increasing both the information input from the stimulus variable $S$ as well as the noise input of $R_1$ to $R_2$. Consequently, all quantities show relatively little variation in this complementary scenario. Interestingly, due to the larger entropy of response variable $R_1$ compared to stimulus variable $S$, the activity and conditional independence ($I_{R}(S; R_1, R_2)$, $I_{R}(R_1; R_2) | S \rangle_{1/2}$) increase slightly over the stimulus related information in both response variables.

Finally, variation of $a_{R1}$ and $a_{R2}$ in parallel in the interval [1,10] reveals different information theoretic signatures for all models. In model M1 the information in the response variables about the stimulus reaches a maximum of 1 bit for all cases, $I_{R}(S; R_1)$, $I_{R}(S; R_2)$ and $I_{R}(S; R_1, R_2)$. Hence the variables become fully redundant with respect to the stimulus, resulting in a negative Synergy of $-1$. However, the conditional correlation $I_{R}(R_1; R_2) | S \rangle_{1/2}$ remains close to zero, as the two variables are not subjected to stimulus-independent co-variation. For model M2 all quantities are dependent on the two multiplicative coefficients and reach extreme values. The two response variables become fully redundant. The information theoretic signature of model M3 shows an intermediate pattern between the simulations of M1 and M2, as indicated by the behaviour of activity and conditional dependence. Superficially, the observed pattern is reminiscent of the variation of parameter $a_{R1}$ in model M2. However, the activity dependence goes to zero for small parameters values in M3, while it remains above zero for the simulation of model M2. This reflects the fact that $a_{R2}$ is constant in the $a_{R1}$ simulation of model M2, but not in the parallel $a_{R1}/a_{R2}$ simulation of model M3.

Validation of the selected number of response bins

To our knowledge there exists no investigation that demonstrates how to obtain an optimal trade-off between sensitive information estimation and bias control for combined EEG and fMRI data. While this is an important issue, it is also beyond the scope of the current study. As described above we have chosen reasonable precautions not to overestimate the information inherent in the experimental data based on established bias correction methods and the Gaussian null model simulations. We chose the number of response bins to obtain approximately twice as many trials per stimulus as response bins, which, based on simulations with respect to neurophysiological data (Panzeri et al., 2007; Pola et al., 2003) should yield conservative information estimates. In this section we discuss simulations of models M0, M1, M2 and M3 which further validate this choice.

Figs. 4A and B demonstrate the effects of varying the number of response bins per response variable between discrete values of 2 to 9, resulting in the total number of response bins varying between 4 and 81. As the effect of varying the number of response bins is dependent on the specific structure of the analyzed data, we simulated two scenarios, a low signal-to-noise scenario in which the ratio between parameters $a_{R1}$, $a_{R2}$ and $\sigma_S^2$ was set to 2 (Fig. 4A) and a high signal-to-noise scenario (Fig. 4B), with this parameter ratio set to 10. For models M1, M2, and M3 we applied the additional bias correction based on M0. As indicated by the M0 simulations in both Figs. 4A and B, increasing the number of response bins leads to diminishing control of the bias mainly for $I_{R}(S; R_1, R_2)$ and $I_{R}(R_1; R_2) | S \rangle_{1/2}$. However, simultaneously the sensitivity of the analysis to actual information in the data increases, as is particularly evident for activity and conditional dependence of model M2 and to a lesser degree model M3. Based on these results, we reasoned that choosing the number of response bins to equal 36 in total maximizes the sensitivity to information present in the data, while keeping the loss of bias control at bay. Note that this is specific to the experimental scenario of 85 trials per stimulus and two different stimuli. Future studies based on more reasonable assumptions about the underlying probability distributions in the case of combined EEG-fMRI data might provide further insight into the question of response space regularization.

In summary, these simulations indicate the following:

1) The bias correction methods work to reduce the estimated information theoretic quantities from a limited sample to zero in the case of no posited stimulus–response signal and response signal–stimulus signal dependencies (model M0). However, the shuffling correction of $I_{R}(S; R_1, R_2)$ (and to a lesser degree the bias correction of $I_{R}(S; R_1, R_2) | S \rangle_{1/2}$) ‘overcorrects’ the estimated quantity in the case of a stimulus set of size $|S| = 2$, and $N_S = 85$, which we ameliorate by estimating this overcorrection from repeated simulations, and then adding the estimated overcorrection to the shuffling corrected estimate of $I_{R}(S; R_1, R_2)$. Although the remaining bias is less severe for the other quantities, we apply the same procedure to the other quantities for consistency. The values estimated from the Gaussian null model are applied to the analysis of the experimental data below.

2) The assessment of the information theoretic quantities computed from data simulated from models M1, M2 and M3 demonstrate that our implementation yields results which are compatible with the interdependencies of the response variables posited in the models. In addition, these simulations demonstrate the differentiating power of the theoretical framework as for all parameters variations different behaviours of the information theoretic quantities for the three models were observed.

3) Finally, analyzing the models in our experimental scenario (85 trials per stimulus, two different stimuli) indicates that setting the number of response bins to approximately half the number of trials per stimulus yields an optimal trade-off between specificity and sensitivity for the information theoretic analysis.

Experimental data

Behavioural data

Due to a technical malfunction of the button response box, reliable behavioural data could only be obtained from 8 observers (8 of whom remained in the final sample, see below). For these, the hit rate in the
target (fixation cross change) detection task was high (0.95 ± 0.03 (SEM)) and the number of false alarms was low (3.1 ± 1.6 (SEM)) indicating good task performance, fixation and sustained attention.

Single subject and group average EEG and fMRI data

We acquired five reliable data sets (sessions) for 12 of the 14 observers. For the remaining two observers we were only able to obtain four reliable data sets each due to technical issues. These were excluded from further analysis to optimise the within-subject variance homogeneity in the group analyses presented. To evaluate the data quality of the recorded EEG we computed average visual evoked potentials (VEPs) for each subject individually (Fig. S3), as well as the grand average over subjects (Fig. 5A). Reliable P100 peaks could be identified in the reconstructed electrode data of at least one of the occipital electrodes in 10 of the 12 subjects (Fig. S3A). The subject-specific electrode which showed the maximal P100 peak amplitude was then chosen for subsequent single-trial feature extraction and the information theoretic analysis.

For the fMRI data, we computed average haemodynamic response functions (HRFs) in terms of percentage signal change with respect to pre-stimulus baseline, extracted from 5 mm radius spheres, as described above (Fig. S3). Reliable haemodynamic responses were identified in all subjects and the grand average calculated (Fig. 5B).

Information theoretic analyses

Using the theoretical framework introduced above, we aimed to obtain insights into the following questions: 1) Which of the EEG and fMRI time-domain features convey information about the stimulus and how do they compare (analysis of \( I(S; R) \) for both modalities)? 2) How much information do pair-wise combinations of the time-domain features from both modalities convey about the stimulus (analysis of \( I(S; R_1, R_2) \)), and how does this compare to the information they convey individually (analysis of \( \langle I(S; R_1) \rangle \) and \( \langle I(S; R_2) \rangle \))? 3) How much information do time-domain features of one modality convey about a feature from the other modality, both with respect to differential stimulation (analysis of \( I(R_1; R_2) \)) and in the absence of differential stimulation (analysis of \( \langle I(R_1; R_2|S) \rangle \)). 4) Finally, how do different data pre-processing strategies affect the estimation of information theoretic quantities? With the exception of question 4) regarding the effect of

---

**Fig. 4.** Variation of the number of response bins in the information theoretic analysis of artificial data generated by the set of linear Gaussian models M0–M3. (A) Low signal-to-noise scenario \((a_{R_1} = a_{R_2} = 2, \sigma_{R_1}^2 = \sigma_{R_2}^2 = 1)\), (B) high signal-to-noise scenario \((a_{R_1} = a_{R_2} = 10, \sigma_{R_1}^2 = \sigma_{R_2}^2 = 1)\).
pre-processing strategies, all information theoretic quantities were calculated using the data which had undergone advanced pre-processing.

Figs. 6–11 summarise the results of these analyses averaged over observers (mean ±SEM). For each information theoretic quantity we statistically evaluated whether it was significantly different from zero (one-sample t-test, two-tailed) and performed a number of statistical comparisons in the framework of the general linear model (repeated-measures ANOVA with Huynh–Feldt correction where appropriate) unless stated otherwise. The results relating to each of the research questions are discussed below.

I(R; S; R;): stimulus–response characteristics

The results for the information that time-domain features convey about the stimulus in isolation are depicted in Fig. 6. The first three panels depict EEG evoked potential features, grouped according to amplitude, latency and RMS. The final graph depicts the information properties of the fMRI features, namely the HRF amplitude, latency and RMS.

Within the EEG modality four features conveyed reliable information about the stimulus, namely the onset-reversal P100 amplitude, the onset-reversal P100 latency, the onset and the reversal RMS. Two additional features, the onset P100 amplitude and the combined onset-reversal P100–N140 amplitude showed a trend for reliable information estimates. A one-way ANOVA comparing all EEG feature means revealed significant differences (F(7.7,84.2) = 2.11, p = 0.04) between the features. Comparing the sub-features in each domain (amplitude, latency, and RMS) showed significant differences for the amplitude features (F(3.8, 42.2) = 3.12, p = 0.03), but not for the latency and RMS features (F(5, 55) = 1.72, p = 0.14, F(2, 22) = 2.69, p = 0.09), substantiating the higher information content of the onset-reversal P100 amplitude than the N70 and N140 amplitude features in the amplitude domain.

Within the fMRI modality, the information conveyed by all of the features was reliable and larger than for the EEG features. The differences between the information content of the fMRI modality features were not significant (one-way ANOVA: F(2,22) = 0.62, p = 0.55).

To evaluate the reliability of the differences between the fMRI and EEG features we performed a two-way ANOVA with factors ‘modality’ (EEG, fMRI) and ‘feature’ (amplitude, latency, and RMS), for which we collapsed the data over EEG sub-features. This ANOVA indeed revealed a significant main effect for ‘modality’ (F(1,11) = 12.41, p < 0.01), no significant main effect for ‘feature’ (F(2,22) = 0.42, p = 0.66) and no significant interaction between the two factors (F(2,22) = 0.67, p = 0.52).

It is worth noting that the smaller stimulus related information values for the EEG time-domain features compared to the fMRI time-domain features may be the consequence of the impoverished EEG data quality during simultaneous EEG–fMRI data acquisition and the necessary artefact subtraction methods. In fact, performing the same analysis on data collected from a subset of the included observers outside of the scanner just before the main experiment, suggested that the information content of EEG time-domain features is at least equal, if not larger than that of fMRI time-domain features (for details, see Fig. 6B).

I(R; S; R;):Synergy: feature combinations across modalities

An important question in the integration of simultaneously acquired EEG–fMRI data is whether the joint observation of both variables is in some way more (or less) informative than the observation of each variable in isolation. By estimating I(R; S; R;2) and Synergy for the pairwise combination of EEG and fMRI time-domain features, we attempted to gain some insight into this question. Fig. 7 displays the results for I(R; S; R;2) and all possible combinations of the time-domain features of interests. The data are grouped with different EEG feature domains (amplitude, latency, and RMS) as rows, and different fMRI feature domains (amplitude, latency, and RMS) as columns. All information values are significantly larger than zero, and appear slightly larger for the combination of EEG amplitude and RMS features with fMRI features than for the combination of EEG latency features.

To assess the reliability of differences between the feature pairings with respect to their information content, we performed two statistical analyses: first, a two-way ANOVA with factors ‘EEG feature’ (amplitude, latency, and RMS) and ‘fMRI feature’ (amplitude, latency, and RMS), where we collapsed the data over the respective EEG sub-features, and secondly a one-way ANOVA for each of the feature pairing sub-groups. The first of these comparisons yielded no significant main effect for ‘EEG feature’ (F(2,22) = 2.53, p = 0.10) and no significant main effect of ‘fMRI feature’ (F(2,22) = 1.17, p = 0.33). However, we observed a significant interaction between factors (F(3.5,39.0) = 3.73, p = 0.01), indicating a dependency of the information value on the specific EEG–fMRI feature pairing. Within the sub-groups, a significant main effect was detected in the pairings of EEG amplitude and fMRI latency features (F(7.1, 77.9) = 2.93, p = 0.01).

Comparing the results of I(R; S; R;2) to those of the information content of the modality features in isolation (I(R; S; R;)), Fig. 6 indicates higher information content in the joint observed frequency distribution. Specifically, the question arises whether the joint distribution of two modality features conveys more or less than the sum of each modality feature’s marginal distribution. To this end, the evaluation of the normalised Synergy of the two features is of interest.

As the information content of each marginal distribution is always equal to or less than the information content of the respective joint distribution (i.e. I(R; S; R;1) ≤ I(R; S; R;2) and I(R; S; R;2) ≤ I(R; S; R;1)) this ranges between –1 and 1. In one extreme case the response variable marginal distributions contain no information about the stimulus, and only the joint observation of both variables is informative, in which case the response variables are said to be fully synergistic (i.e. I(R; S; R;1) = I(R; S; R;2) = 0 ⇔ Synergy = 1). In the other extreme case each marginal distribution contains all information present in the joint distribution, and the variables are said to be fully redundant (i.e. I(R; S; R;1) = I(R; S; R;2) = 1 ⇔ Synergy = –1).
The normalised Synergy for all feature pairings is shown in Fig. 8. The data are again grouped with different EEG feature domains (amplitude, latency, and RMS) as rows, and different fMRI feature domains (amplitude, latency, and RMS) as columns. The normalised Synergy measure is very sensitive to estimation errors (e.g. for information values close to zero, or negative values due to non-optimal bias correction). For this reason, in order to avoid artefactual Synergy values resulting from the inclusion of noisy information values, it was computed only from those observers for whom the respective values of $I_{N}(S;R_1)$ and $I_{N}(S;R_2)$ were larger than 0.01 and $I_{N}(S;R_1,R_2)$ larger than 0.1, reflecting roughly the average values observed in Figs. 6 and 7. As this resulted in different degrees of freedom and the pairing of different observers across quantities, we did not perform explicit statistical comparisons in this case. Inspecting the data nonetheless shows that the Synergy varies widely between different feature pairings. With approximately the same number of negative and positive results, the overall pattern of results indicates that for the current experimental paradigm, the response signal marginal distributions contain the majority of the stimulus-related information and the observed joint distributions do not add much to this. The most reliable synergistic effects were observed for the pairings of VEP RMS and HRF latency features, while the most reliable redundancy effects appear for pairings of VEP latency and HRF latency/RMS features.

$IN(R_1;R_2)$ and $IN(R_1;R_2|S)$: inter-modality comparisons

Next, we asked how much information time-domain features of each modality convey about each other. Fig. 9 depicts the mutual information $I_{N}(R_1;R_2)$ for all possible pairwise combinations of EEG and fMRI features. The first column depicts all pairings of EEG features with HRF amplitude, the second with HRF latency, and the last with HRF RMS. For all feature pairings, the mutual information is small, and only in a few cases significantly different from zero. Specifically, only the combined P100 onset/reversal latency and HRF latency, as well as...
the VEP onset/VEP total RMS and HRF RMS showed reliable activity dependence.

We evaluated the differences between EEG feature domains and fMRI features using a two-way repeated-measures ANOVA with factors ‘EEG feature’ (amplitude, latency, and RMS) and ‘fMRI feature’ (amplitude, latency, and RMS), collapsing the data over EEG sub-features. This ANOVA yielded no significant main effect of ‘EEG feature’ ($F(2,22) = 0.30, p = 0.74$), a significant main effect of ‘fMRI feature’ ($F(2,22) = 5.20, p = 0.01$) and no significant interaction ($F(4,44) = 1.17, p = 0.33$). Further, we detected no significant differences between different EEG features in each of the EEG feature domains. Hence, we fail to reject the null hypothesis that the different EEG and fMRI features are equally informative about each other.

In sum, the observed mutual information values are low and mostly not significantly different from zero. They do not therefore indicate a clear functional relationship between the EEG and fMRI features.

The mutual information $I_N(S;R_1,R_2)$ between two time-domain features of each modality was determined in the presence of differential stimulation. In principle it is possible that the probabilistic dependence that exists between two features is accounted for by this differential stimulation alone, or, in the other extreme, totally independent of differential stimulation, e.g. due to another source of co-variation (also called noise correlation). To disentangle these possibilities, the (stimulus-averaged) stimulus conditional mutual information $I_N(S;R_1;R_2)$ can be computed and compared to $I_N(S;R_1)$ and $I_N(S;R_2)$. However, from a practical perspective, it has to be noted that the number of trials available for the estimation of the stimulus conditional observed frequency distributions reduces proportionally to the inverse of the stimulus set cardinality, and hence is likely to be more error prone than the other measures discussed here.

We evaluated the stimulus conditional mutual information between fMRI and EEG features for all possible comparisons (Fig. 10). Evaluation of differences between the feature pairings in a two-way ANOVA with factors ‘EEG feature’ (amplitude, latency, and RMS, collapsed over sub-features) and ‘fMRI feature’ (amplitude, latency, and RMS) yielded a significant main effect of ‘EEG feature’ ($F(1,19,9) = 12.79, p < 0.01$) indicating the larger variability of $I_N(S;R_1)$ values compared to $I_N(S;R_2)$, where this was not the case. Further, no significant main effect of ‘fMRI feature’ ($F(2,22) = 0.82, p = 0.45$), and no significant interaction ($F(4,44) = 1.26, p = 0.30$) were detected. Within each EEG feature domain, no significant differences were detected.

To compare $I_N(S;R_1)$ and $I_N(S;R_2)$, we first carried out a three-way repeated-measures ANOVA with factors ‘information theoretic quantity’ ($I_N(S;R_1)$, $I_N(S;R_2)$), ‘EEG feature’ (amplitude, latency, and RMS, collapsed over sub-features) and ‘fMRI feature’ (amplitude, latency, and RMS). This ANOVA revealed a significant main effect of information theoretic quantity ($F(1,11) = 27.5, p < 0.01$), indicating some overall differences between $I_N(S;R_1)$ and $I_N(S;R_2)$. For a more detailed assessment of the differences between the unconditional and stimulus conditional mutual information, we next focussed only on those feature pairings for which we detected significantly non-zero unconditional mutual information. The results of this assessment (paired sample t-tests) are given in Table 3. This analysis shows that the differences between the estimated values of $I_N(S;R_1)$ and $I_N(S;R_2)$ are significant for the RMS features. However, in no case is the activity dependence significantly larger than
the conditional dependence indicating, if any, stimulus-independent response signal co-variation.

To summarize, the observed mutual information between EEG and fMRI time-domain features was quite low and mostly not significantly different from zero, with a few exceptions.

Influence of data pre-processing methods
As a final analysis, we aimed to quantify the effect of the data pre-processing methods on the estimated information theoretic quantities. To this end, we computed the observed information theoretic quantities for the most informative EEG and fMRI time-domain features in isolation, the onset-reversal P100 and HRF amplitude for both feature extraction strategies (see Materials and methods). The results of this analysis are shown in Fig. 11. Clearly, the advanced pre-processing methods lead to an increase in the estimated information values, which is most evident for those information theoretic quantities that refer to the stimulus. Since both $I_{\text{IN}}(S; R_1, R_2)$ and $I_{\text{IN}}(S; R_1, R_2)$ are fairly low and unreliable for the basic processing strategies, we refrained from including Synergy in this assessment. To assess the reliability of the observed results, we carried out a two-way repeated-measures ANOVA with factors ‘method’ (ICA and GLM, raw data and literature V1 coordinates) and ‘information theoretic quantity’. This ANOVA revealed a significant main effect of ‘method’ ($F(1,11) = 17.8, p < 0.01$), a significant main effect of ‘information theoretic quantity’ ($F(2.2, 24.0) = 5.70, p < 0.01$), as well as a significant interaction between both factors ($F(2.19, 24.1) = 4.88, p = 0.02$), substantiating the larger information values observed for the advanced pre-processing strategies.

Discussion
The aims of the study reported here were threefold. Firstly, to adopt a theoretical framework consisting of a set of information theoretic quantities established in the electrophysiological study of neuronal population codes to multimodal (EEG–fMRI) brain imaging experiments. Secondly, to study this framework under controlled response signal conditions using artificial data sets generated by a set of linear Gaussian models and thirdly, to practically apply the framework to a combined EEG–fMRI data set obtained with robust visual stimulation. We will discuss the results of each point in turn.

Theoretical framework
First and foremost, the application of an information theoretic framework to experimental EEG–fMRI data allows a precise definition and differentiation of the colloquial terms ‘information’ or ‘informative signal’. Most notably, the framework allows the differentiation between signal features which are informative about the external stimulus they represent (in isolation and combination) and those which are informative about other signal features, both in the presence and absence of external stimulation. Accordingly, signals from the EEG and fMRI modalities can be integrated with respect to the stimulus-related information they provide ($I(S; R_1, R_2)$) or the information they provide with respect to each other (activity and conditional dependence).

The theoretical application of the framework proposed by (Panzeri et al., 2008; Schneidman et al., 2003) to multimodal brain imaging
experiments is straightforward, as it is based on the quantification of signal features which is readily achieved in the case of neurophysiological and multimodal brain imaging data. However, it should be noted that the selection of signal features is necessarily based on prior knowledge about the signals. For our experimental application, we selected traditional time-domain features of evoked potentials (EEG) and haemodynamic response functions (fMRI). It is obvious that this is not the only possible (and potentially not even the most relevant) choice for these signal modalities. Potential other applications, especially in the case of the EEG, are the quantification of frequency-domain features, both stimulus-related and stimulus-independent, which might turn out to be informative with respect to the stimulus and signal features of the fMRI modality (Logothetis et al., 2001; Niessing et al., 2005; Mukamel et al., 2005). We also based the analysis on the selection of signal features from pre-specified anatomical locations. Another potential application of the framework could include a spatial maximization approach, i.e. to determine those locations both in anatomical (MRI) and electrode space, for which a specific information theoretic quantity based on a specific signal feature is maximised.

Another theoretically straightforward extension of the framework is to account for response variables of higher dimensionality than two. Specifically, the question of how much information higher order combinations of \( n \geq 2 \) features from both modalities can provide about the stimulus and each other is of considerable interest. In addition, for cognitive neuroimaging experiments, the integration of another response variable type, namely behaviour (with its common quantifications such as performance accuracy and reaction times) is of interest.

Further, it should be noted that we deliberately chose to present a basic information theoretic framework. As the application of information theory to electrophysiological data has a long history, more elaborate implementations then the one we presented exist, e.g. (Panzeri et al., 2008). Another aspect for the successful application of information theoretic approaches to the study of multimodal imaging data will be the theoretical development of bias correction techniques (either from available or novel ones) which are tailored to the particulars of EEG–fMRI experiments.

**Simulations**

The aim of the simulations was to provide an intuitive understanding of the various quantities in low-complexity response signal generation settings. Moreover, they were useful for the practical development and validation of the framework. It can be argued that given the simplicity of the models, an analytical treatment with respect to the information theoretic quantities is readily achieved. However, we believe that for an intuitive understanding even in analytical treatments, the examination of the behaviour of the various quantities as functions of the model parameters is necessary. In addition the simulations of a Gaussian null model allowed an evaluation of the quality of the bias correction techniques of our experimental settings. We discuss the implications of this below.

**Experimental results**

Practically applying the information theoretic framework to a set of EEG–fMRI single-trial time-domain features obtained during
Checkerboard stimulation has provided several interesting results. Firstly, a subset of the time-domain features we chose was found to be informative with respect to the stimulus in isolation. For the EEG data, the most informative measures were related to the amplitude and latency of the P100 component (Fig. 6), which has previously been shown to be reliably modulated by stimulus contrast (Shawkat and Kriss, 2000; Odom et al., 2004). Similarly, the information theoretic framework also identified the effect of stimulus contrast on the haemodynamic response (Bagshaw and Warbrick, 2007; Logothetis et al., 2001; Wan et al., 2006). The fact that the methodology proposed here is able to correctly identify the primary features that are affected by stimulus contrast is a useful validation.

It is of note that fMRI features were generally more informative than EEG features for simultaneously acquired data. However, as discussed above, this is probably due to the diminished data quality of artefact affected EEG data in the concurrent acquisition of both modalities. Although a full analysis is beyond the scope of this study, the comparison of data acquired inside and outside the scanner suggests that there is some loss of stimulus related information specifically for the lower amplitude peaks of the evoked potential, N70 and N140. These effects are likely to be reduced as the quality of concurrently recorded EEG data improves. As demonstrated by the comparison of basic and advanced pre-processing strategies (Fig. 11), the information theoretic framework provides a quantitative methodology to assess the effect of different artefact removal techniques.

Secondly, combining EEG and fMRI time-domain features by quantifying the information in their joint distribution was more informative than treating each one in isolation. However, the overall pattern of the Synergy results remains inconclusive, as both synergistic

---

**Fig. 10.** $\langle N(R_1; R_2|S) \rangle$: Stimulus conditional dependence of EEG and fMRI features, averaged over observers±SEM (n = 12). VEP: visual evoked potential, HRF: Haemodynamic response function. Onset and Onset & Reversal refer to the EEG features averaged over stimulus onset or stimulus onset and reversal. Asterisks indicate means significantly larger than zero (p < 0.05, two-tailed one-sample t-test).

**Fig. 11.** Influence of data pre-processing methods on the estimated information theoretic quantities for the EEG signal feature ‘onset-reversal averaged P100 amplitude’ (P100) and the fMRI signal feature ‘haemodynamic response function amplitude’, averaged over observers (n = 12) ± SEM.

---

**Table 3**

<table>
<thead>
<tr>
<th>Feature combination</th>
<th>Difference</th>
<th>T-Value</th>
<th>DF</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>VEP latency/HRF latency</td>
<td>0.019</td>
<td>1.30</td>
<td>11</td>
<td>0.21</td>
</tr>
<tr>
<td>(VEP: Onset &amp; reversal P100)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VEP RMS/HRF RMS (VEP: Onset RMS)</td>
<td>−0.05</td>
<td>−4.10</td>
<td>11</td>
<td>0.00</td>
</tr>
<tr>
<td>VEP RMS/HRF RMS (VEP: Total RMS)</td>
<td>−0.05</td>
<td>−3.90</td>
<td>11</td>
<td>0.00</td>
</tr>
</tbody>
</table>
(Synergy \textgreek{>0}) and redundant (Synergy \textless 0) effects were observed. Further, the absolute Synergy results were relatively small (mainly \textless -0.4), implying that most of the information about the stimulus is actually conveyed by the marginal response distributions, rather than their joint distribution. The fact that the exact relation between neuronal activity, surface-electrode EEG phenomena and BOLD signal changes remains debatable complicates the interpretation of the obtained results. For example, if one assumes that EEG and fMRI phenomena represent the same underlying neuronal activity, only redundancy should be observed. The fact that this is not the case can be argued biologically, i.e. that there might be differences in the neuronal activity that the modality features reflect, and only the joint observation from both modalities yields an estimation of the information that is conveyed at the neuronal level. On the other hand, assuming that EEG and fMRI measures reflect the same underlying neuronal information content, the fact that we amend the estimation of $R_2(S; R_1, R_2)$ using a Gaussian null model and observe synergistic effects may indicate that the bias correction procedure we chose is not optimal. Clearly, a better understanding of neuronal information content and its relation to information estimated from non-invasive brain imaging techniques is necessary for conclusive interpretation of synergistic effects.

Finally, we found that the mutual information between time-domain features of the two modalities is relatively low. This finding is in line with previous findings on the correlation of single-trial EEG and fMRI data (Bagshaw and Warbrick, 2007) and posits some interesting questions for the ‘integration through prediction’ approach to EEG–fMRI integration, which capitalizes on the linear co-variation of EEG and fMRI signal features. Possibly, the prediction of the fMRI time-series could be improved if a set of more informative EEG features than the one studied here and used previously to predict the fMRI time-series (Debener et al., 2005; Eichele et al., 2005; Philiastides and Sajda, 2007) could be identified.

Finally, comparing the overall pattern of results obtained from the experimental data to the linear Gaussian model simulations indicates the following: firstly, the experimentally observed mutual information values do not assume their theoretically possible maximum values, probably reflecting the low signal-to-noise ratio of the data. In terms of the simulations displayed in Fig. 3, this would correspond to low ratios between values of the ‘signal gain factors’ $a_{1G}/a_{2G}$ and the ‘noise factors’ $\sigma_{G1}/\sigma_{G2}$. Furthermore, the experimentally observed pattern of stimulus related information in both the response marginal distributions and the response joint distribution, no prominent Synergy, and relatively low activity and conditional independence is most reminiscent of the low signal-to-noise ratio evaluations of model M1 in the parallel $a_{1G}/a_{2G}$ simulation. Arguing indirectly, this would support a view in which both EEG and fMRI signal features contain stimulus related information, but show little additional single-trial co-variation. It should however be noted that this interpretation assumes that the EEG and fMRI feature distributions are actually Gaussian and is of course confined to the current experimental paradigm. Also, comparing the information signatures of observed data with the ones created artificially based on very simple Gaussian models is an indirect approach to model evaluation at best. In the current study, the main purpose of the simulations was to constrain the analysis of experimental data, and to demonstrate the potential power of the framework. Future work based on generative forward models for EEG and fMRI and their inversion based on machine learning based approaches (Valdes-Sosa et al., in press) appear more promising for a direct assessment of the underlying nature of stimulus-dependent and inde-pendent neural activity, and resulting EEG and fMRI signal changes.

Three challenges that we encountered in the practical application of the information theoretic framework to real data are worth noting. Firstly, the simulations on the Gaussian null model showed that for our experimental setting (i.e. low number of different stimuli, relatively low number of trials per stimulus), the shuffling bias correction technique resulted in overcorrection of the respective quantities which we amended as described in the Methods section. For the practical application of information theoretic concepts to multimodal brain imaging data it therefore appears necessary to develop specifically tailored bias correction techniques. The approach we chose in the current study has an obvious drawback: the estimated information theoretic quantity values are unbiased only if assumed to be generated by a Gaussian null model. This indirectly reintroduces a Gaussian assumption into the experimental results, which is clearly not desirable (although of course common in many brain imaging studies).

The second challenge we identified was the computation of the normalised Synergy measure from noisy, bias-corrected data, as this data can have the detrimental property of being close to zero and non-positive. We chose an ad-hoc amendment for this, restricting our analysis of Synergy to those subjects whose data did not have these detrimental properties. However, for future studies a more principled approach in this regard is desirable.

Finally, the estimated quantities depend on the experimental settings (as shown for the EEG features extracted from data recorded inside and outside MRI scanner) as well as the data pre-processing techniques. As this is quite natural for empirical results, we do not regard this as a drawback, but rather as a potential application of the methods proposed: the estimated quantities allow a meaningful evaluation of the quality of an experimental setting and hence can be used directly for the optimization of experimental techniques and designs.

Conclusion

The question arises of what our treatment of multimodal brain imaging data adds with respect to previous integrative approaches. While a full treatment of this question is beyond the scope of this report (e.g. it would have to include an in depth consideration of the differences between linear correlation and mutual information), we would like to point to some similarities and dissimilarities with other approaches.

(Klinner et al., 2005) categorised previous signal integration approaches as follows: 1) Integration through prediction (using a property of the EEG signal to predict changes in the BOLD response in the statistical framework of the general linear model), 2) integration through constraints (e.g. constraining EEG source localization based on fMRI results), 3) and integration through common forward models (a comprehensive model or set of models which describes the mechanism generating the two signals and the links between them).

With respect to approach 1) and response signal–signal relationships, the framework provided here is more tailored to explicitly address the question of which signal features are mutually dependent, initially free of additional assumptions. The success of approach 1) is based on the assumption that extracted EEG features linearly co-vary with a pre-defined haemodynamic response function under additive Gaussian noise. This assumption is validated by obtaining ‘meaningful’ statistical parametric maps. The framework proposed here explicitly states quantitative (effect sizes) degrees of dependency, while in the current implementation it is based on some prior, spatially constrained feature extraction method (in our example, the definition of regions of interest and particular electrodes). In this vein, the information theoretic framework also allows the explicit differentiation of event-related and non-event-related signal co-variation (Debener et al., 2006; Herrmann and Debener, 2008) by means of the stimulus unconditional and stimulus conditional mutual information measures, which is usually not addressed in approach 1).

Lastly, also with respect to stimulus–response signal relationships, the framework proposed here provides a unified approach using the same mutual information as for the response signal–signal relationships.

With respect to approach 2) of data integration, again the success of approach 2) is based on ‘meaningful’ co-localization results, while
the evaluation of the framework presented here is more explicit. A potential application of the measures proposed here within approach 2) might be the evaluation of obtained co-localization results with respect to the various mutual information measures, or indeed the use of mutual information measures explicitly in the constraint of the source localisation.

Finally approach 3), the integration of both modalities through common forward models fitted to empirical data in a full Bayesian treatment appears to be the most principled approach to real data integration. In our opinion the framework presented here can add to this line of research by explicitly taking into account the response signal probability distributions. Studying the signal feature distributions of real data sets as well as their relationships can potentially be useful to the more realistic specification of prior distributions in a Bayesian generative model context.

Acknowledgments

This work was supported by grant number EP/F023057/1 from the UK Engineering and Physical Sciences Research Council (EPSRC). The authors would like to thank Nina Salman for help with data acquisition and Therese Lennert for comments on an earlier version of this manuscript.

Appendix A. Theoretical framework

Understanding neural stimulus coding and response signal relationships in a multimodal data acquisition context requires a quantitative description of the respective stimulus–response signal and response signal–response signal interdependencies. In a typical experiment, the observer is presented with a stimulus while an EEG signal and an fMRI signal are concurrently recorded. For clarity, we differentiate between response signals, response signal modalities and response signal features: By ‘signal modality’ we understand the type of data acquisition modality (i.e. EEG, fMRI, behaviour), by ‘signal’ the entirety of the acquired signal modality in its raw form, while by ‘signal feature’ we understand a given quantified aspect of a signal after data pre-processing. Many different signal features are conceivable, for example the amplitude of evoked haemodynamic response functions in a given brain area for the fMRI signal, or the power of a given frequency band of the EEG signal extracted from specific electrodes. In this very general context of cognitive neuroimaging experiments, a set of questions follows naturally (Panzeri et al., 2008): Which response signal features convey information about the stimulus or cognitive task? Do different signal features within and across response signal modalities convey redundant or complementary information about the stimulus? Which signal features are most closely related within and across response modalities? Is this relationship stimulus-dependent or not?

The information theoretic framework we adopted for multimodal brain imaging experiments has its roots in the study of communication (Shannon, 1948) and has previously been applied to the problem of neural coding in invasive electrophysiological studies (Belitski et al., 2008; Panzeri et al., 2008; Schneidman et al., 2003). An exhaustive introduction to information theory is provided by (Cover and Thomas, 2006) and (MacKay, 2003), amongst others. Presented below are the concepts most relevant to the current application.

Mutual information

A central notion of the framework is the concept of mutual information, i.e. the information that a signal feature conveys about the stimulus or another response signal feature. Consider a neuroimaging experiment with one response signal modality (e.g. EEG). For each stimulus of a given stimulus set \( S = \{s_1, \ldots, s_N\} \) (by \(|X|\) we denote the cardinality of a set, i.e. the number \( n=|X| \) of elements contained in the finite and countable set \( X \)), a response signal feature is determined for each repeat of the stimulus \( s_i \in S \). From the observed responses \( r_j \in R \), the joint probability distribution \( p(s, r) \) can be constructed from the respective entries \( p(s=s_i, r=r_j) \), where we denote the response set as \( R = \{r_1, \ldots, r_M\} \). From this joint distribution the marginal distributions \( p(s) \) and \( p(r) \) can be obtained by summation, and the conditional probability distribution \( p(r | s) \) by division. It should be noted that \( p(s) \), the stimulus probability distribution, is usually under the control of the experimenter.

(Shannon, 1948) proposed entropy as a measure of the variability (or uncertainty) of a given random variable, as it satisfies a set of intuitive assumptions (e.g. it is maximised in the case of a uniform distribution indicating highest uncertainty). Here we consider two entropy measures that can be computed from the distributions introduced above: the response entropy and the noise entropy (Panzeri et al., 2008). For the equations given below it is assumed that both the stimulus set \( S \) and the response set \( R \) are appropriately discretised.

The response entropy, \( H(R) \), captures the overall variability of a given response signal feature across all stimuli, and is obtained from the marginal distribution \( p(r) \) according to

\[
H(R) = \sum_{r \in R} p(r) \log_2 \left( \frac{1}{p(r)} \right). \tag{A1}
\]

On the other hand, the noise entropy, \( H(R|S) \), captures the variability of the signal feature that remains after the signal feature variability due to the stimulus is accounted for. The noise entropy therefore is the signal feature variability that exists even in the absence of differential stimulation, and is given by

\[
H(R|S) = \sum_{r \in R, s \in S} p(s)p(r|s) \log_2 \left( \frac{1}{p(r|s)} \right). \tag{A2}
\]

The information that the response signal conveys about the stimulus distribution is denoted as the mutual information, \( I(S; R) \), between the stimulus and signal feature distribution, and is given as

\[
I(S; R) = H(R) - H(R|S) = \sum_{s \in S} \sum_{r \in R} p(s, r) \log_2 \frac{p(s, r)}{p(s)p(r)}. \tag{A3}
\]

The mutual information as defined above has many properties that correspond well with intuitive connotations of ‘information’, for example that information is additive with respect to statistically independent variables. Its unit is the ‘bit’, if a logarithm of base 2 is used. One bit of information means that, on average, the observation of the signal feature on a given trial reduces the observer’s stimulus uncertainty by a factor of 2 (Cover and Thomas, 2006).

In the current context of EEG–fMRI experiments, the notion of mutual information allows us to gain insight into the stimulus–response signal relationship for each signal modality individually, i.e. we can compute the \( I(S; R_1) \) for an EEG response signal feature and \( I(S; R_2) \) for an fMRI response signal feature. However, for the case of two response signal features as in the case of a neuroimaging experiment we can define additional quantities of interest, which we adopt from Schneidman et al. (2003). The basic idea is to quantify the dependencies (or independencies) of the response signals on the stimulus and on each other. Three concepts of dependence are central: information dependence, activity dependence and conditional dependence ((Schneidman et al., 2003) use the term ‘correlation’ for ‘dependence’ which we try to avoid here in order to prevent confusion with the well-known concept of linear correlation. In fact in can be shown that the concept of mutual information encompasses correlations at any order).

Information dependence

Firstly, we note that the representation of the mutual information between response signal feature and stimulus as given by Eq. (A3)
naturally extends to response signal feature distributions of higher dimensionality, i.e. the case of more than one type of response signal feature. In the current application we are primarily interested in the in the case of two response signal features, one from each modality. Substituting the joint distributions \( p(s, r) \) by \( p(s, r_1, r_2) \) and \( p(r) \) by \( p(r_1, r_2) \) in Eq. (A3), where \( r_1 \) and \( r_2 \) indicate response signal features from each of the two modalities, yields

\[
I(S; R_1, R_2) = \sum_{s \in S} \sum_{r_1 \in R_1} \sum_{r_2 \in R_2} p(s, r_1, r_2) \log_2 \left( \frac{p(s, r_1, r_2)}{p(s)p(r_1, r_2)} \right).
\]

(Eq. A4)

Eq. (A4) represents the information about the stimulus that is conveyed by the joint distribution of response signal features \( R_1 \) (e.g. an EEG signal feature) and \( R_2 \) (e.g. an fMRI signal feature).

Next, the concept of 'information dependence' relates to the question of whether particular response features convey redundant or complementary information about the stimulus. Consider for example the case in which the two response signal features are sensitive to orthogonal stimulus features (a case of particular importance in invasive neurophysiological single cell studies). In this case, the information that the combined response signals convey about the stimulus should be the sum of their individual stimulus mutual information

\[
I(S; R_1, R_2) = I(S; R_1) + I(S; R_2).
\]

(A5)

In addition, two scenarios are possible: the two response signals can be synergistic (i.e. their joint distribution conveys more information about the stimulus than their individual distributions), or redundant (i.e. they jointly convey less information about the stimulus than their individual distributions). These scenarios can be readily captured in the equation for the normalised signal feature Synergy:

\[
\text{Synergy} = (I(S; R_1, R_2) - (I(S; R_1) + I(S; R_2))) / I(S; R_1, R_2).
\]

(A6)

Synergy is a quantity that ranges between \(-1\) for fully redundant response signal features, and \(+1\) for fully synergistic response signal features. In summary, we have now introduced three measures of information that two response signal features convey about the stimulus: \( I(S; R_1) \), \( I(S; R_1) \), and \( I(S; R_1, R_2) \), as well as a standardised measure of their mutual relationship, Synergy. Next we turn to the information that different response signal features convey about one another.

Activity dependence and conditional dependence

'Activity dependence' relates to the question of how much the activity of one response signal feature \( R_i \) depends on the activity of another response signal feature \( R_j \). In the context of EEG–fMRI experiments, this concept is mainly related to the question of which EEG features are most tightly coupled to BOLD signal features (or vice versa). It should be noted that response signal feature dependencies can occur both in the absence as well as the presence of external stimulation (Herrmann and Debener, 2008). Formally, activity dependence is assessed by studying the experimentally obtained joint distribution of stimulus and response signals, \( p(s, r_1, r_2) \), and marginalizing appropriately. An obvious measure for the degree of activity dependence between a pair of response signals is the mutual information that one signal feature conveys about the other, given by

\[
I(R_i; R_j) = H(R_i) - H(R_i | R_j) = \sum_{r_1 \in R_1} \sum_{r_2 \in R_2} p(r_1, r_2) \log_2 \left( \frac{p(r_1, r_2)}{p(r_1)p(r_2)} \right).
\]

(A7)

As an example, \( r_1 \) could represent the power at a given frequency band of the EEG signal at a given electrode, while \( r_2 \) could relate to the integral of the haemodynamic response function in a corresponding brain area. If the two response signals are independent, by definition the joint response distribution factors to \( p(r_1, r_2) = p(r_1)p(r_2) \) and the mutual information \( I(R_i; R_j) \) is zero.

Finally, 'conditional dependence' refers to the notion of noise correlation, that is the question of whether the two response signals are interdependent in the absence of external stimulation. In the multimodal imaging context one would primarily assume that the strength of the neurovascular coupling is independent of the presence or absence of external stimulation, although this might not necessarily be the case. Intuitively, this concept can be understood by considering the conditional distribution of the two response signal features at a given stimulus level \( s_0 \) and \( p(r_1, r_2 | s = s_0) \). If the two response signals are dependent on each other due to some factor other than the stimulus (e.g. the 'hard-wired' neurovascular coupling), then \( p(r_1, r_2 | s = s_0) \) will reveal some dependency between the response signal features. If the two response signals are independent or only dependent on each other under stimulation, the conditional distribution will factor to \( p(r_1, r_2 | s = s_0) = p(r_1 | s = s_0)p(r_2 | s = s_0) \). From these considerations, an obvious measure of response signal conditional dependence is the mutual information between the response signals given the stimulus:

\[
I(R_i; R_j | S) = H(R_i | S) - H(R_i | R_j, S) \quad \text{(A8)}
\]

and more compact, its average across stimulus levels, i.e.

\[
\langle I(R_i; R_j | S) \rangle = \sum_{s \in S} I(R_i; R_j | S = s) / |S| \quad \text{(A9)}
\]

In sum, the information theoretic framework introduced by (Panzeri et al., 2008; Schneidman et al., 2003) and reviewed here rests on the experimental determination of the joint stimulus–response distribution \( p(s, r_1, r_2) \) and the subsequent computation of a set of mutual information quantities which relate response signal features with the stimulus and one another. A summary of the information theoretic quantities is given in Table 1.

Appendix B. Bias correction

For the practical application of the framework introduced above it is important to note that the joint stimulus–response signal distribution \( p(s, r_1, r_2) \) is not known, but can only be estimated from a finite data set as the observed frequency distribution \( p_\theta(s, r_1, r_2) \). Here \( N \) indicates the total number of trials, given by \( N = |S| \cdot N_r \), where \( N_r \) indicates the number of repeats per stimulus (for the current purposes, we assume that all stimuli of the stimulus set are presented equally often). Accordingly, we use the subscript \( N \) for the information theoretic quantities estimated from limited simulated or experimental data, e.g. \( I_N(S; R_1) \) by which we mean the estimated mutual information between stimulus and response signal feature \( R_1 \). Dealing with finite (and in the case of EEG–fMRI experiments relatively small sample sizes) leads to systematic errors (biases) in the values of the computed quantities. The theoretical foundations, as well as methods for the correction of these systematic errors, have been studied extensively for the case of invasive neurophysiological data (Panzeri et al., 2007).

In the current work, for the purposes of assessing the relationship between two variables (i.e. all information theoretic quantities except \( I(S; R_1, R_2) \) and Synergy) we utilise the PT-correction scheme, proposed and implemented by (Panzeri and Treves, 1996). For the computation of bias-corrected \( I(S; R_1, R_2) \), we employ 'shuffling correction' (Montemurro et al., 2007b; Panzeri et al., 2007). In the following we provide a brief overview of these two correction schemes.
Denoting the real mutual information between two random variables by \( I \), it has been shown \( (\text{Panzeri and Treves, 1996}) \) that the average additional error, or bias, can be expressed as a series expansion in inverse powers of sample size \( N \):

\[
(I_N) - I = \sum_{m=1}^{\infty} C_m.
\]

(\text{B1})

Here \( (I_N) \) indicates the average Plug-In mutual information over repeated sampling from the real underlying probability distribution, and \( C_m \) represents successive contributions to the asymptotic expansion of the bias. Only the first term of the expansion is, on first approximation, independent of the underlying probability distributions \( (\text{Panzeri and Treves, 1996}) \) and can be estimated from the sampled data as

\[
C_1 = \frac{1}{2N\ln 2} \left( \sum_{j=1}^{N} \mathcal{R}_j - 1 \right) - (\mathcal{R} - 1).
\]

(\text{B2})

Here \( \mathcal{R} \) denotes the number of relevant response bins for the stimulus conditional distributions \( p_B(r | s) \), and \( \mathcal{R} \) the number of relevant responses for the marginal response distribution \( p_B(r) \). The relevant response bins are defined as the number of different response bins \( r_i \) with non-zero probability. \( (\text{Panzeri and Treves, 1996}) \) introduced a Bayesian approach to estimate these two numbers from the data for a single response variable. Following estimation of \( \mathcal{R} \) and \( \mathcal{R} \), Eq. (\text{B2}) can be evaluated and the resulting term subtracted from the Plug-In mutual information estimate, yielding the bias-corrected version.

For the case of multidimensional response variables, i.e. the case of response signal features of two modalities, the most powerful bias correction scheme is a shuffling approach \( (\text{Panzeri et al., 2007; Montemurro et al., 2007a}) \). The information contained in the joint response of multiple response variables \( R_1,\ldots,R_j \) can be computed indirectly as

\[
I_{\text{sh}}(S; R) = H(R) - H_{\text{ind}}(R | S) + H_{\text{sh}}(R | S) - H(R | S).
\]

(\text{B3})

In Eq. (\text{B3}), \( R \) indicates a multidimensional response variable, which is a two-dimensional vector in the context of the current study. Two new entropy measures, \( H_{\text{sh}}(R | S) \) and \( H_{\text{sh}}(R | S) \), are introduced in Eq. (\text{B3}). In the context of the current study these refer to the entropies of the observed frequency distributions \( P_{\text{sh}}(r_1, r_2 | s) \) and \( P_{\text{sh}}(r_1, r_2 | s) \), respectively. \( P_{\text{sh}}(r_1, r_2 | s) \) is the conditional joint response observed frequency distribution that would have been observed, were the two response signal feature distributions independent random variables. In this case, the respective joint distribution would factor according to

\[
P_{\text{sh}}(r_1, r_2 | s) = P_B(r_1 | s)P_B(r_2 | s).
\]

(\text{B4})

Hence, this non-empirical distribution can be calculated by pair-wise multiplication of the marginal distributions \( p_B(r_1 | s) \) and \( p_B(r_2 | s) \). \( P_{\text{sh}}(r_1, r_2 | s) \), on the other hand, is the conditional joint response observed frequency distribution created by removing response variable correlations in an alternative way, namely shuffling the sampled response variables for a given individual for each response element, thereby removing trial-by-trial dependencies. Further details of this approach can be found in \( (\text{Panzeri et al., 2007}) \) and \( (\text{Montemurro et al., 2007a}) \).

Finally, it should be noted that the problem of systematic errors in summary measures (statistics) of population parameters computed from finite samples is not specific to the framework applied here, but is a rather fundamental problem of quantitative empirical science, and encountered in many contexts, e.g. the computation of standard deviations from mean values \( (\text{Hays, 1994}) \).

Appendix C. Linear Gaussian model specifications

A linear Gaussian model with one ‘stimulus’ variable \( S \) and two ‘response’ variables, \( R_1 \) and \( R_2 \), can be regarded as a linear mapping \( f \) from a ‘stimulus set’ \( S = \{s_1,\ldots,s_N\} \subset \mathbb{R} \) to a response set \( R = \left\{ \left( \begin{array}{c} r_1 \\ r_2 \\ \end{array} \right) \right\} \subset \mathbb{R}^2 \). \( f:S \rightarrow R \), given by

\[
f(s) = as + b + e, \ \forall s \in S, \ a, b \in \mathbb{R}^2 \text{ and } e \sim N(0, \Sigma_{ab\Sigma}).
\]

(\text{C1})

Here, \( a \) is a multiplicative coefficient, \( b \) an additive bias, and \( e \) an additional error or noise term which is normally distributed with mean \( 0 \) and covariance matrix \( \Sigma \). Equivalently, and for practical purposes more conveniently, one can obtain samples from the specified models by sampling from a bivariate Gaussian distribution with model specific mean \( \langle \mu_M \rangle \) and covariance matrix \( \langle \Sigma_M \rangle\). Hence, we can write the distribution of the response variables given the stimulus as

\[
r = \left( \begin{array}{c} r_1 \\ r_2 \\ \end{array} \right) \sim N(\mu_M | s), \langle \Sigma_M \rangle)
\]

(\text{C2})

where

\[
N(\mu_M | s), \langle \Sigma_M \rangle = \frac{1}{2\pi \langle \Sigma_M \rangle^{1/2}} \exp \left\{ -\frac{1}{2} (r - \langle \mu_M | s \rangle)^T \langle \Sigma_M \rangle^{-1} (r - \langle \mu_M | s \rangle) \right\}.
\]

(\text{C3})

Note that the marginal stimulus distribution is usually not empirical but rather under the control of the experimenter. In the simulations (as well as in the real experiment) the marginal stimulus distribution was set to the discrete uniform distribution, i.e.

\[
p(s_i) = \frac{1}{|S|}.
\]

(\text{C4})

Therefore, the probability for stimulus \( s_i \) to occur on any given trial is the inverse of the cardinality \( |S| \) of the stimulus set.

In model M0, no dependence of the response variables on the stimulus was postulated. Hence, the dependence of the mean on the stimulus and interaction between the response variables additive error components was removed by setting

\[
\mu_{M0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } \Sigma_{M0} = \begin{bmatrix} \sigma_{1}^2 & 0 \\ 0 & \sigma_{2}^2 \end{bmatrix}.
\]

(\text{C5})

Hence, \( \sigma_{1}^2 \) and \( \sigma_{2}^2 \) are the only free parameters of model M0, and both individual response variables \( R_1 \) and \( R_2 \) are normally distributed with mean \( 0 \) and respective variances \( \sigma_{1}^2 \) and \( \sigma_{2}^2 \).

In model M1, both response variables depend on the stimulus value, but not on each other. In order to implement the linearity constraints introduced above, we set the mean vector and covariance matrix for M1 to

\[
\mu_{M1} = \begin{bmatrix} a_{R1} s + b_{R1} \\ a_{R2} s + b_{R2} \end{bmatrix} \text{ and } \Sigma_{M1} = \begin{bmatrix} \sigma_{1}^2 & 0 \\ 0 & \sigma_{2}^2 \end{bmatrix}.
\]

(\text{C6})

where \( a_{R1}, a_{R2}, b_{R1}, b_{R2}, \sigma_{1}^2 \) and \( \sigma_{2}^2 \) are free and independent parameters, and \( s \in S \) represents the stimulus value. The joint stimulus response probability distribution of model M1 can then be sampled as

\[
p_{M1} = (s, r_1, r_2) = N(r | \mu_{M1}, \Sigma_{M1}, s).
\]

(\text{C7})
For model $M_2$, $R_2$ depends on the value of $S$ indirectly by means of $R_1$, which depends directly on $S$. We therefore set for (a derivation of these results, see (Bishop, 2006), pp 370)

$$\mu_{M_2} = \left( a_{R1} + b_{R1} \right) \text{ and } \Sigma_{M_2} = \left( \begin{array}{cc} a_{R2} & a_{R2} \sigma_1 \\ a_{R2} \sigma_1 & \sigma_1^2 \end{array} \right) \left( \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right)$$

Again with free parameters $a_{R1}, a_{R2}, b_{R1}, b_{R2}, \sigma_1^2$ and $\sigma_2^2$. Hence, the joint response probability distribution of model $M_2$ can be sampled from

$$P_{M_2} = \left( s_1, r \right) = N\left( \mu_{M_2}, \Sigma_{M_2}, s \right).$$

Finally, model $M_3$ was created as an intermediate scenario between models $M_1$ and $M_2$. In model $M_3$, response variable $R_2$ is sensitive to both changes in variable $S$ and variable $R_1$ depending on a linear weighting constant $t \in [0,1]$, such that for $t = 0$ model $M_3$ is identical with model $M_1$ and for $t = 1$ model $M_3$ is identical to model $M_2$. Hence, the mean vector and covariance matrix of model $M_3$ are for $t \in [0,1]$

$$\mu_{M_3} = \left( a_{R1} + b_{R1} \right) \left( 1 - t \right) a_{R2} \sigma_1 + b_{R1} \left( 1 - t \right) a_{R2} \sigma_1 + b_{R2} \right) \left( 1 - t \right) a_{R2} \sigma_1 + b_{R2} \right)$$

and the joint response probability distribution of model $M_3$ can be sampled from

$$P_{M_3} = \left( s_1, r \right) = N\left( \mu_{M_3}, \Sigma_{M_3}, s \right).$$

Apppendix D. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2009.07.038.

References


