entropy

MDPI

Article

Partial and Entropic Information Decompositions of a Neuronal Modulatory Interaction

Jim W. Kay [1,*], Robin A. A. Ince [2], Benjamin Dering [3] and William A. Phillips [3]

[1] Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK
[2] Institute of Neuroscience and Psychology, University of Glasgow, Glasgow G12 8QQ, UK; robin.ince@glasgow.ac.uk
[3] Faculty of Natural Sciences, University of Stirling, Stirling FK9 4LA, UK; b.r.dering@stir.ac.uk (B.D.); w.a.phillips@stir.ac.uk (W.A.P.)
* Correspondence: jimkay049@gmail.com; Tel.: +44-141-391-3288

Received: 30 June 2017; Accepted: 23 October 2017; Published: 26 October 2017

Abstract: Information processing within neural systems often depends upon selective amplification of relevant signals and suppression of irrelevant signals. This has been shown many times by studies of contextual effects but there is as yet no consensus on how to interpret such studies. Some researchers interpret the effects of context as contributing to the selective receptive field (RF) input about which neurons transmit information. Others interpret context effects as affecting transmission of information about RF input without becoming part of the RF information transmitted. Here we use partial information decomposition (PID) and entropic information decomposition (EID) to study the properties of a form of modulation previously used in neurobiologically plausible neural nets. PID shows that this form of modulation can affect transmission of information in the RF input without the binary output transmitting any information unique to the modulator. EID produces similar decompositions, except that information unique to the modulator and the mechanistic shared component can be negative when modulating and modulated signals are correlated. Synergistic and source shared components were never negative in the conditions studied. Thus, both PID and EID show that modulatory inputs to a local processor can affect the transmission of information from other inputs. Contrary to what was previously assumed, this transmission can occur without the modulatory inputs becoming part of the information transmitted, as shown by the use of PID with the model we consider. Decompositions of psychophysical data from a visual contrast detection task with surrounding context suggest that a similar form of modulation may also occur in real neural systems.

Keywords: information theory; partial information decomposition; entropic information decomposition; synergy; redundancy; contextual modulation; neural information processing

1. Introduction

Amplifiers, such as hearing aids, for example, are designed to increase signal strength without distorting the informative content that it transmits, i.e., its "semantics". Though independence of semantics has been a truism of information theory since its inception, information decomposition may help distinguish the effects of amplifying inputs from driving inputs which determine what the output transmits information about, which is what we will refer to here as its "semantics". It may seem intuitively obvious that any output must necessarily transmit information about all inputs that affect it, but that intuition is misleading. Here, we use information decomposition to show that a modulatory input can influence the transmission of information about other inputs while remaining distinct from that information.

This may help resolve a long-standing controversy within the cognitive neurosciences concerning the nature of "contextual modulation". Many see the wide variety of psychophysical and physiological

phenomena that are grouped under this heading as demonstrating that the concept of a neuron's receptive field, i.e., what the cell transmits information about, needs to be extended to include an extra-classical receptive field; see e.g., [1]. In contrast to that many others see these phenomena as evidence that contextual modulation does not change the cell's receptive field semantics; see e.g., [2–4].

Resolution of this issue requires an adequate definition of "modulation", which is used in several different, and often undefined, ways. It is frequently used to mean simply that one thing affects another. That unnecessary use of the term introduces substantial confusion, however, because the term is also often used to refer to a three-term interaction. It could be used to refer to any three-way interaction in which A effects the transmission of information about B by C. Our use is more specific than that, however. The essence of the modulatory interaction that we study here is that the modulator affects transmission of information about something else without becoming part of the information transmitted. The effect of the volume control on a radio provides a simple example. It changes signal strength without becoming part of the message conveyed. The use of the term "modulation" in telecommunications potentially adds further confusion, however, because in either amplitude modulation (AM) or frequency modulation (FM) it is the "modulatory" signal that is used to convey the message to be transmitted. That is the opposite of what we and many others in the cognitive and neurosciences refer to as "modulation". While awaiting a consensus that resolves this terminological confusion we define our usage of the term "modulation" as explicitly and as clearly as we can. Modulation that increases output signal strength is referred to as "amplification" or "facilitation". Modulation that decreases output signal strength is referred to as "disamplification", "suppression", or "attenuation".

Information decomposition could help clarify the notion of "modulation" as used within the cognitive and neurosciences in at least three ways. First, by requiring formal specifications to which decompositions can be applied it enforces adequate definition. Second, by being applied to a transfer function explicitly designed to be modulatory, it deepens our understanding of the information processing operations performed by such interactions. Third, decomposition of a modulatory interaction that is formally specified shows the conditions under which it can be distinguished from additive interactions and provides patterns of decomposition to which empirically observed patterns can be compared.

In this paper we apply information decomposition to a transfer function specifically designed to operate as a modulator within a formal neural network that uses contextually guided learning to discover latent statistical structure within its inputs [5]. We show that this transfer function has the properties required of a modulator, and that they can be clearly distinguished from additive interactions that do contribute to output semantics. A thorough understanding of this modulatory transfer function is of growing importance to neuroscience because recent advances suggest that something similar occurs at an intracellular level in neocortical pyramidal cells, and may be closely related to consciousness [6,7]. It is also important to machine learning because the information processing capabilities of networks such as those used for deep learning might be greatly enhanced if given the context-sensitivity that such modulatory interactions can provide.

Modulatory interactions distinguish the contributions of two distinct inputs to an output, so they imply some form of multivariate mutual information decomposition. Various forms of decomposition have been proposed, however, and they may offer different resolutions to this issue. We therefore compare resolutions that arise from two proposals discussed elsewhere in this Special Issue. One is Partial Information Decomposition [8–11]. The other is Entropic Information Decomposition [12,13]. We find that though there are important differences between these two proposed forms of decomposition, they are in agreement with respect to their implications for the issue of distinguishing between additive and modulatory interactions.

The notion of modulation is essentially a three-term interaction in which one input variable modulates transmission of information about a second input variable by an output. The two inputs therefore make fundamentally different kinds of contribution to the output. In contrast to that,

additive interactions do not require the two inputs to remain distinct because their contributions can be summarized via a single integrated value. Many information decomposition spectra and surfaces are displayed in the following, demonstrating their expressive power and the variety of information processing operations that a single transfer function can perform.

2. Notation and Definitions

In this section we describe our notation and define the information-theoretic concepts which are used in the sequel. A generic "p" is used to denote a probability mass function, with the argument of the function signifying which distribution is being described. Capital letters are used to denote random variables, with their realised values appearing in lower-case. We denote the conditional probability that $Y = y$, given that $X_1 = x_1$ and $X_2 = x_2$ by the conditional mass function $p(y|x_1, x_2)$ for $y \in B$, and $(x_1, x_2) \in B^2$, where $B = \{-1, +1\}$.

In [14], the RF and contextual field (CF) inputs were multivariate, but here we consider the special case of the local processor in [14] having two binary inputs, X_1 and X_2, and one binary output, Y, with all three random variables having range space B. The joint distribution of (Y, X_1, X_2) is given by the probability mass function (p.m.f.) $p(y, x_1, x_2)$, where

$$p(y, x_1, x_2) = \Pr(Y = y, X_1 = x_1, X_2 = x_2), \quad (y, x_1, x_2) \in B^3.$$

This distribution will be considered in the form

$$p(y, x_1, x_2) = p(y|x_1, x_2)p(x_1, x_2), \tag{1}$$

and we will separately specify a joint p.m.f $p(x_1, x_2)$ and a conditional p.m.f. $p(y|x_1, x_2)$.

In the local processor in Figure 1, the value of X_1 provides the receptive field (RF) input to the local processor, while the value of X_2 is the input from the contextual field (CF). The value of the RF input, X_1, is multiplied by the signal strength s_1 to form the integrated RF input and similarly for the CF input, X_2. Therefore, the values taken by the integrated RF and CF inputs are $r = s_1 x_1$ and $c = s_2 x_2$. These integrated values have both strength and a sign. The strength is a constant property of the defined system, while the sign can change from sample to sample. The signal strengths, s_i, are positive real numbers. The manner in which these signals are combined in the output unit will be described in Section 3.

In this study, it is assumed that $\Pr(X_1 = 1) = \Pr(X_2 = 1) = \frac{1}{2}$ and that the correlation between X_1 and X_2 is d, where $-1 < d < 1$. This means that

$$\lambda \equiv \Pr(X_1 = 1, X_2 = 1) \quad = \Pr(X_1 = -1, X_2 = -1) = \frac{1 + d}{4}, \tag{2}$$

$$\mu \equiv \Pr(X_1 = 1, X_2 = -1) = \Pr(X_1 = -1, X_2 = 1) \quad = \frac{1 - d}{4}. \tag{3}$$

It is also assumed that the conditional output probability has a logistic form, with

$$\Pr(Y = 1|X_1 = x_1, X_2 = x_2) = 1/(1 + \exp(-T(x_1, x_2))), \tag{4}$$

where T is a transfer function which depends also on the signal strengths, s_1, s_2. In Section 3, the two transfer functions that are used in this study are specified. It should be noted that we are actually considering a class of trivariate probability distributions that are indexed by (s_1, s_2, d), where $s_1 > 0, s_2 > 0, -1 < d < 1$, although this indexation is suppressed in the sequel for ease of notation. The various classical measures of information and measures of partial information used are calculated using a member of the class of trivariate probability distributions, defined in (1)–(4), that is given by a particular choice of (s_1, s_2, d).

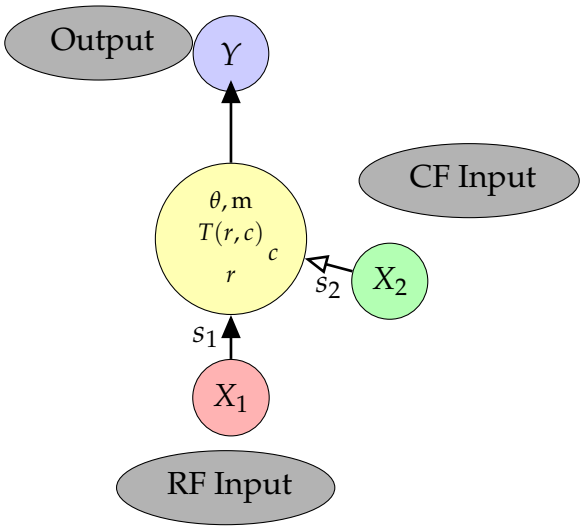


Figure 1. A local processor with binary receptive field (RF) input X_1, contextual field (CF) input X_2 and output Y. The weights on the connections from the RF and CF inputs into the output unit are s_1, s_2, which represent the strengths given to the input signals. The integrated RF input, r, and the integrated CF input, c, are passed through a transfer function T and a logistic nonlinearity within the output unit to produce the conditional output probability, θ, as well as the output conditional mean, m.

We now define the standard information theoretic terms that are required in this work and based on results in [15]. We denote by the function H the usual Shannon entropy, and note that any term with zero probabilities makes no contribution to the sums involved. The total mutual information that is shared by Y and the pair (X_1, X_2) is given by,

$$I[Y;(X_1,X_2)] = H(Y) + H(X_1,X_2) - H(Y,X_1,X_2). \tag{5}$$

The information that is shared between Y and X_1 but not with X_2 is

$$I[Y;X_1|X_2] = H(Y,X_2) + H(X_1,X_2) - H(X_2) - H(Y,X_1,X_2), \tag{6}$$

and the information that is shared between Y and X_2 but not with X_1 is

$$I[Y;X_2|X_1] = H(Y,X_1) + H(X_1,X_2) - H(X_1) - H(Y,X_1,X_2). \tag{7}$$

Finally, the co-information of (Y, X_1, X_2) has several equivalent forms

$$I[Y;X_1;X_2] = I[Y;X_1] - I[Y;X_1|X_2] = I[Y;X_2] - I[Y;X_2|X_1] = I[X_1;X_2] - I[X_1;X_2|Y], \tag{8}$$

where, for $i = 1, 2$,

$$I[Y;X_i] = H(Y) + H(X_i) - H(Y,X_i), \text{ and } I[X_1;X_2] = H(X_1) + H(X_2) - H(X_1,X_2). \tag{9}$$

We note that classical Shannon information measures have been used in neural coding studies to investigate measures of synergy and redundancy; see for example [16].

When we come to define measures of partial information it will be necessary to calculate these information quantities with respect to another p.m.f., say $q(y, x_1, x_2)$, and to denote this we add the subscript "q" to such terms, e.g., $I_q(Y;X_1;X_2)$. This means that the p.m.f. $q(y, x_1, x_2)$ is used in the computation rather than the original p.m.f. $p(y, x_1, x_2)$.

3. An Interaction Designed to Be Modulatory

Our concern here is with variables that can take either positive or negative values, which can be seen as being analogous to excitation and inhibition in neural systems. We model that decision as a probabilistic binary variable that chooses between the values 1 and -1. The criteria to be met by a modulatory transfer function in this case have been stated and discussed in many previous papers; see e.g., [17–19]. The criteria for a modulatory interaction were stated for a local processor receiving two inputs: the integrated RF input, r, and the integrated CF input, c. The requirements were stated in terms of the level of activation within the local processor, although in this paper we use this term to denote the value of the transfer function, and they are amended slightly here. Please note that the term 'integrated' was used in previous work to refer to the weighting and summing of the components of a multivariate input; we continue to use this term here even though the input to each field is univariate. The value of the transfer function is fed into a logistic function to compute the conditional probability that a 1 will be transmitted. Stated in those terms the CF input modulates transmission of information about the RF input if four criteria are met:

1. If the integrated RF input is extremely weak, then the value of the transfer function is close to zero.
2. If the integrated CF input is extremely weak, then the value of the transfer function should be close to the integrated RF input.
3. If the integrated RF and CF inputs have the same sign, then the absolute value of the transfer function should be greater than when based on the RF input alone. On the other hand, if the RF and CF inputs are of opposite sign then the absolute value of the transfer function should be less than when based on the RF input alone.
4. The sign of the value of the transfer function is that of the integrated RF, so that the context cannot change the sign of the conditional mean of the output.

In general terms, the CF input would have no modulatory effect on the output when the output and the CF input are conditionally independent given the value of the RF input, which is equivalent to the conditional mutual information $I[Y; X_2|X_1]$ being equal to zero. One case where this happens for any member of the class of trivariate binary distributions defined in (1)–(4) is when the correlation between the inputs X_1, X_2 is ± 1, for then $I[Y; X_2|X_1] = 0$; see Theorem 5. On the other hand, in situations where this conditional mutual information is non-zero then X_2 influences the prediction of the output Y by the input X_1 in the sense that

$$\Pr(Y = y|X_1 = x_1, X_2 = x_2) \neq \Pr(Y = 1|X_1 = x_1),$$

for at least one $(y, x_1, x_2) \in B^3$. This is a very general form of modulation, but the type of modulation defined in requirements 1–4 is very specific and we call it "contextual modulation". This contextual modulation is relevant within the local processor at the level of individual system inputs and outputs. On the other hand, the following conditions express the notion of contextual modulation for the whole ensemble of inputs and outputs:

M1: If the *RF* signal is strong enough and the CF input is extremely weak then $I[Y; X_1|X_2]$ can have its maximum value, $I[Y; X_1]$ can be maximised and $I[Y; X_2|X_1]$ is close to zero. This shows that the RF input is sufficient, thus allowing the information in the *RF* to be transmitted, and that the CF input is not necessary.

M2: $I[Y; X_2|X_1]$ and $I[Y; X_1]$ are close to zero when the RF input is extremely weak no matter how strong the CF input. This shows that the RF input is necessary for information to be transmitted, and that the CF input is not sufficient to transmit the information in the RF input.

M3: When $s_1 < s_2$ and when the RF input is weak, $I[Y; X_1]$ and $I[Y; X_1|X_2]$ are both larger when the CF input is moderate than when the CF input is weak. Thus the CF input modulates the transmission of information about the RF input.

One might expect that these two definitions of contextual modulation are linked. In the limiting situation of $s_1 \to 0$ it is possible to show that requirement 1 implies M1, and as $s_2 \to 0$ one finds that requirement 2 implies M2. It seems difficult to prove more general connections and so this matter is considered computationally in Section 3.1.

Multivariate binary processors were also considered in [5], thus allowing for choice between many more than two alternatives. It was also shown that the coherent infomax learning rule also applies to this multivariate case such that the contextually guided learning discovers variables defined on the RF input space that are statistically related to variables specified in, or discovered by, other streams of processing within the network. Thus it implements a multi-stream, non-linear, form of latent structure analysis. There are two distinct aspects of semantics in this system, i.e., the receptive field selectivity of each unit within a local processor and the positivity or negativity of its output. Here we are primarily concerned with that latter aspect. We show below that:

(i) the modulatory input affects output only when the primary driving integrated RF input is non-zero but weak;
(ii) that even when it does have an effect it has no effect on the sign of the conditional mean output, and
(iii) that it can have those modulatory effects without the binary output transmitting any unique information about the modulator.

In the case where the processor has a binary output, the transfer function has the form

$$T(x_1, x_2) = r\left[k_1 + (1 - k_1)\exp\left(k_2 rc\right)\right], \quad (k_2 > 0,\, 0 < k_1 < 1),$$

where $r = s_1 x_1, c = s_2 x_2$, k_1 and k_2 are constants, and here we take $k_1 = \frac{1}{2}$ and $k_2 = 1$.

This transfer function was designed to effect a modulatory interaction between two input sources, with one source being the primary driver while the role of the the second "contextual" source is to modulate transmission of information about the primary source. The effect of the contextual source is to amplify or disamplify the strength of the signal from the primary source in such a way that the semantic content (the sign) of the primary source is not changed. Neither of the PID and EID considered in this paper has previously been applied to this kind of signal and we now show this to be possible.

In this paper, the version of the modulatory transfer function we use takes the form

$$T_M(x_1, x_2) = \tfrac{1}{2} r\left[1 + \exp\left(rc\right)\right] = \tfrac{1}{2} s_1 x_1\left[1 + \exp\left(s_1 x_1 \times s_2 x_2\right)\right], \tag{10}$$

for given values x_1, x_2 of the random variables X_1, X_2, and given signal strengths s_1, s_2. Here the integrated RF input is $r = s_1 x_1$ and the integrated CF input is $c = s_2 x_2$, and they both have a sign and a strength. The output conditional probability is given by

$$\theta = \Pr(Y = 1 | X_1 = x_1, X_2 = x_2) = 1/[1 + \exp\left(-T_M(x_1, x_2)\right)]. \tag{11}$$

Whether this probability is greater than or less than $\frac{1}{2}$ is determined solely by the value of $x_1(\pm 1)$, and the form of T_M ensures that the contextual signal cannot change the sign of the output conditional mean. Thus the output produced has semantic content, and also the value of the output conditional probability, θ, gives the semantic content a measure of strength in the sense that values of θ closer to 0 or 1 indicate a more definite decision. The conditional variance of Y is $4\theta(1 - \theta)$, and so uncertainty in the output decision is largest when $\theta = 1/2$ and zero when $\theta = 0$ or 1. An alternative description is to say that the precision (reciprocal variance) is least when $\theta = 1/2$ and it tends to infinity as θ approaches 0 or 1. Within the local processor the conditional mean of the output, $\mathrm{m} = 2\theta - 1$, is also computed. It has both a sign and a strength.

Given the form of T_M, the integrated RF will be amplified in magnitude whenever the signs of x_1 and x_2 agree, and it will be disamplified when these signs do not agree. The role of the integrated CF is to modify the strength of the conditional mean output without conveying its own semantic content (i.e., its sign). This form of transfer function ensures that the maximum extent of any disamplification of the primary signal is by a factor of 2.

By way of contrast, we also consider an additive transfer function by simply adding together the integrated RF and CF inputs, r, c, to give

$$T_A(x_1, x_2) = r + c = s_1 x_1 + s_2 x_2, \tag{12}$$

with the output conditional probability given by

$$\Pr(Y = 1 | X_1 = x_1, X_2 = x_2) = 1/[1 + \exp(-T_A(x_1, x_2))]. \tag{13}$$

The use of this transfer function also affects the values of θ and m but, unlike the modulatory transfer function, this additive transfer function can change the sign of the output conditional mean m, which is not consistent with the fourth condition for a modulatory transfer function described above. The additive transfer function does satisfy condition M1 but does not satisfy condition M2 or M3. This additive transfer function can be seen as a simple version of the common assumption within neurobiology that neurons function as integrate-and-fire point processors. While this assumption does not imply that all integration is linear it does mean that such integration computes a single value per local processor. The results produced using the these two different transfer functions will be discussed in Sections 5–8.

Please note that in the sequel we normally abbreviate the terms "integrated RF input" and "integrated CF input" by using just "RF input" and "CF input", respectively. In particular, whenever a strength is implied for the RF or CF input, then we mean that the 'integrated' values of these inputs are being considered.

3.1. Analysis Using Classical Shannon Measures

We start in this section by presenting results involving the classical Shannon measures for the system defined in Sections 2 and 3. First we recall that λ and μ are defined in (2) and (3) and set up some further simplifying notation which is used in the results. We set

$$u = \Pr(Y = 1 | X_1 = 1, X_2 = 1), \quad \text{and} \quad v = \Pr(Y = 1 | X_1 = 1, X_2 = -1). \tag{14}$$

The parameters u and v are function of s_1 and s_2, and u takes the value u_M or u_A depending on which transfer function is being used; similarly for v. From (10) for transfer function T_M

$$u_M = 1/(1 + \exp(-\tfrac{1}{2} s_1 (1 + \exp(s_1 s_2))), \quad \text{and} \quad v_M = 1/(1 + \exp(-\tfrac{1}{2} s_1 (1 + \exp(-s_1 s_2))), \tag{15}$$

whereas, from (12), for transfer function T_A

$$u_A = 1/(1 + \exp(-(s_1 + s_2)), \quad \text{and} \quad v_A = 1/(1 + \exp(-(s_1 - s_2)). \tag{16}$$

Finally, we define

$$z = 2\lambda u + 2\mu v, \quad w = 2\lambda u + 2\mu(1 - v) \ \text{ and } \ h(v) = -v \log(v) - (1 - v) \log(1 - v), \tag{17}$$

where $0 < v < 1$. We note also that the value of z has two forms: z_M when transfer function T_M is used and z_A when transfer function T_A is employed; similarly for w. We now collect together our results in the following theorem, proof of which is relegated to the appendix.

Theorem 1. *It is assumed that $s_1 > 0, s_2 > 0$. For the probability distribution defined in (1)–(4), the following results hold.*

(a) $I[Y; X_1|X_2] = h(w) - 2\lambda h(u) - 2\mu h(v)$;
(b) $I[Y; X_2|X_1] = h(z) - 2\lambda h(u) - 2\mu h(v)$;
(c) $I[Y; X_1] = 1 - h(z)$;
(d) $I[Y; X_2] = 1 - h(w)$;
(e) $I[Y; X_1; X_2] = 1 - h(z) - h(w) + 2\lambda h(u) + 2\mu h(v)$;
(f) $I[Y; (X_1, X_2)] = 1 - 2\lambda h(u) - 2\mu h(v)$,

where from (15) and (16), $u = u_M, v = v_M$ when the transfer function T_M is employed and $u = u_A, v = v_A$ when the transfer function T_A is used.

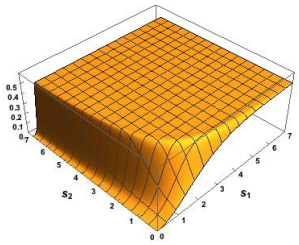
Since we are particularly interested in interactions among the three variables, Y, X_1, X_2, we now show the classic Shannon information measures defined in (6)–(9), with surface plots given in Figures 2 and 3. A correlation between the inputs of 0.78 was considered to ensure that these measures have the same maximum possible value of 0.5 bits, and a zero correlation was considered to represent the case of independent inputs. One purpose is to discuss the general links between requirements 1–4 and conditions M1–M3 from Section 3 and also the use of the transfer functions defined in (10) and (12).

First, we notice in Figure 2 that the modulatory and additive transfer functions produce very different surfaces. In Figure 2a,b, the surface for T_M rises more quickly to its maximum than the surface for T_A, and in Figure 2a sections parallel to the s_1 axis are similar for $s_2 \geq 2$, whereas the surface for T_A is symmetric about the line $s_1 = s_2$. Figures 2d,f,h,j and 3d,f,h,j for T_A show clear asymmetry about the line $s_1 = s_2$.

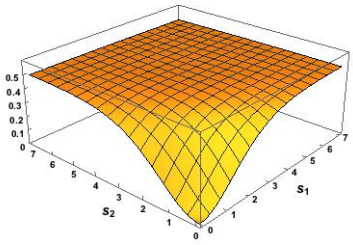
When the strength of the CF input, s_2, is very small we notice in Figures 2e and 3e that $I[Y; X_2|X_1]$ is close to zero. Figure 2c shows that $I[Y; X_1|X_2]$ rises quickly, then gradually, towards its maximum at 0.5 as the strength of the RF input, s_1, increases, as does the surface in Figure 3c although there the maximum value is higher at 1. Figures 2g and 3g show that $I[Y; X_1]$ rises towards a maximum value of 1; this rise is much steeper when the correlation is 0.78 than when it is zero. These observations provide support for condition M1 when the modulatory transfer function is used. Similar observations on the corresponding figures based on the use of the additive transfer function show that condition M1 is satisfied in this case also.

Figures 2e,g and 3e,g show, when s_1 is close to zero, that $I[Y; X_2|X_1]$ and $I[Y; X_1]$ are both close to zero, thus supporting condition M2 when the modulatory transfer function is used. This is not the case when the additive transfer function is employed, as can be seen from Figures 2f,h and 3f,h. It is important to note that these figures do not all use the same scales for the heights of the surface. For example, the scales of Figures 2e and 3e are expanded because $I[Y; X2|X1]$ is always small when the transfer function is modulatory.

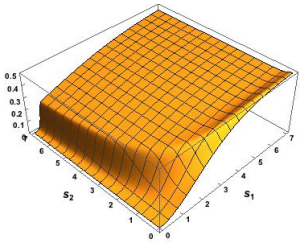
Also, when the strength of the RF input is weak (say $s_1 = 1$), we notice in Figures 2c,g and 3c,g that both $I[Y; X_1]$ and $I[Y; X_1|X_2]$ are larger for moderate CF strengths (say $s_2 = 5$) than when the the strength of the CF input is extremely weak ($s_2 = 0.05$, say), with this effect being stronger when the correlation between inputs is 0.78. This provides support for condition M3 when the modulatory transfer function is used. Inspection of the corresponding plots based on the additive transfer function show this effect only for $I[Y; X_1]$ in Figure 2h, and so condition M3 does not hold for the additive function.



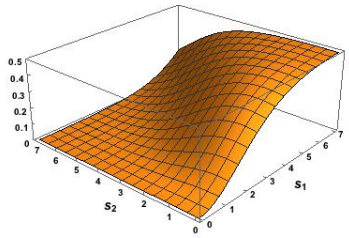
(**a**) Modulatory, $I[Y; X_1; X_2]$



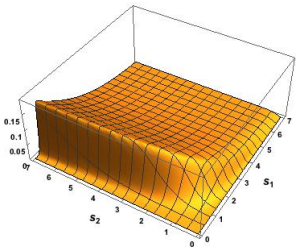
(**b**) Additive, $I[Y; X_1; X_2]$



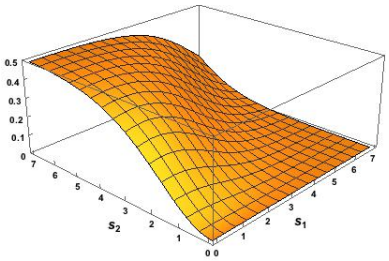
(**c**) Modulatory, $I[Y; X_1|X_2]$



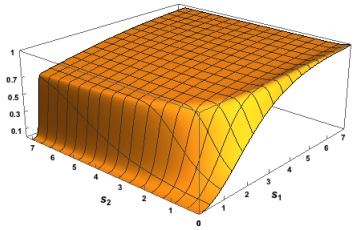
(**d**) Additive, $I[Y; X_1|X_2]$



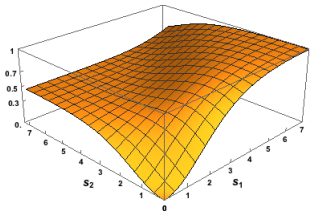
(**e**) Modulatory, $I[Y; X_2|X_1]$



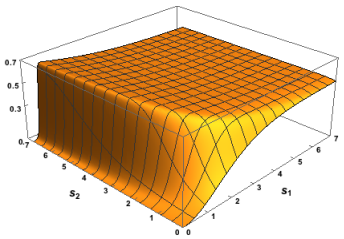
(**f**) Additive, $I[Y; X_2|X_1]$



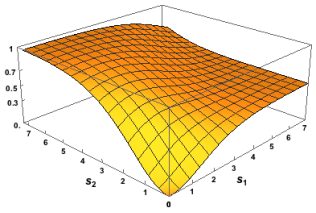
(**g**) Modulatory, $I[Y; X_1]$



(**h**) Additive, $I[Y; X_1]$



(**i**) Modulatory, $I[Y; X_2]$



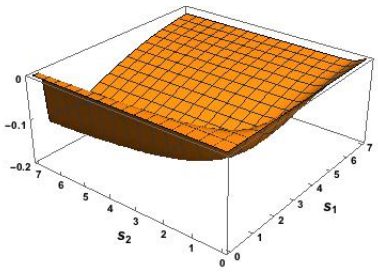
(**j**) Additive, $I[Y; X_2]$

Figure 2. Classical Shannon measures (in bits), based on additive and modulatory transfer functions, and a correlation between inputs of 0.78.

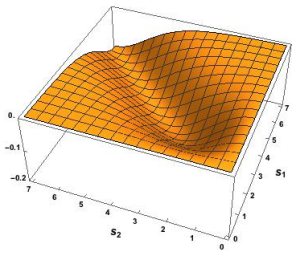
In Figure 3a,b, the surfaces of the co-information $I[Y; X_1; X_2]$ are negative, as expected from (8), since the correlation between X_1 and X_2 is zero and so their mutual information is zero.

Finally we focus discussion on the phenomenon of particular relevance to the subject of this paper by considering the surface plots of $I(Y; X_2|X_1)$. In Figure 2e, an interesting pattern emerges. There is a steep rise for small values of s_1 and for all values of $s_2 \geq 2$, and then the surface quickly dies away. This pattern is repeated in Figure 3e. This suggests that X_2 is affecting the information shared between Y and X_1, indicating that modulation of some form might be taking place.

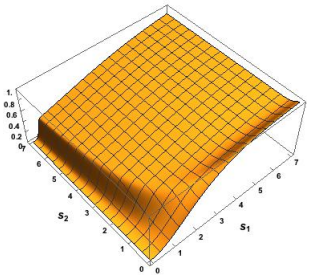
It could be argued, however, that X_2 is part of the output semantics in the sense that the output contains information specifically about X_2 itself. Since $I[Y; X_2|X_1]$ is clearly positive for these values of s_1, s_2, it is impossible to know whether or not this is the case based on this classical Shannon measure. It was shown in [8], that $I[Y; X_2|X_1]$ could be decomposed into two terms: the unique information that X_2 conveys about Y as well as synergistic information that is not available from X_2 alone, but rather gives the information that X_1 and X_2, acting jointly, have about the output Y. We now apply information decompositions in order to resolve these different interpretations. For discussion of some limitations of classical Shannon measures and the need for new measures of information, see [20].



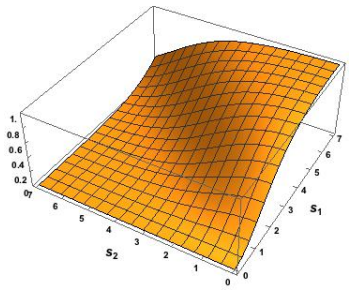
(**a**) Modulatory, $I[Y; X_1; X_2]$



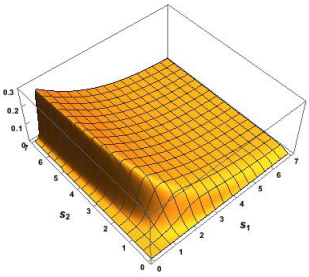
(**b**) Additive, $I[Y; X_1; X_2]$



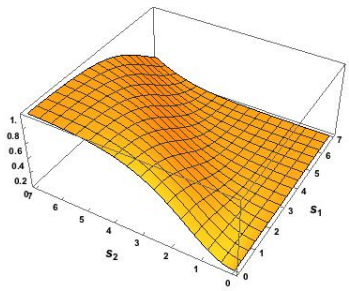
(**c**) Modulatory, $I[Y; X_1|X_2]$



(**d**) Additive, $I[Y; X_1|X_2]$

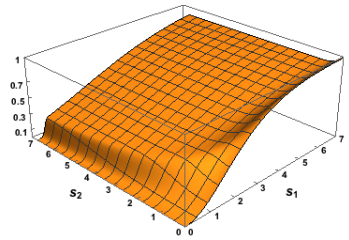


(**e**) Modulatory, $I[Y; X_2|X_1]$

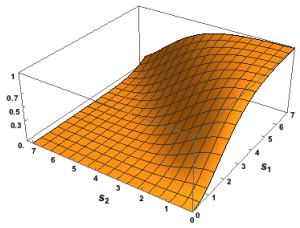


(**f**) Additive, $I[Y; X_2|X_1]$

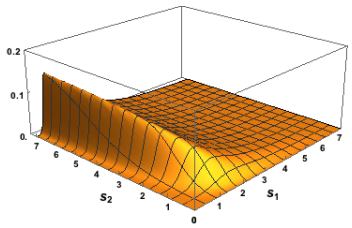
Figure 3. *Cont.*



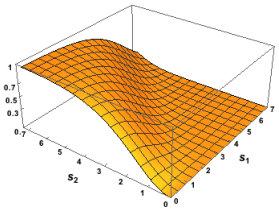
(**g**) Modulatory, $I[Y;X_1]$



(**h**) Additive, $I[Y;X_1]$



(**i**) Modulatory, $I[Y;X_2]$



(**j**) Additive, $I[Y;X_2]$

Figure 3. Classical Shannon measures (in bits), based on additive and modulatory transfer functions, and a zero correlation between inputs.

4. Information Decompositions

Williams and Beer [8] introduce a framework called the Partial Information Decomposition (PID) which decomposes mutual information between a target and a set of multiple predictor variables into a series of terms reflecting information which is shared, unique or synergistically available within and between subsets of predictors. Here we focus on the case of two input predictor variables, denoted X_1, X_2, and an output target Y. The information decomposition can be expressed as

$$I[Y;(X_1,X_2)] = I_{unq}[Y;X_1|X_2] + I_{unq}[Y;X_2|X_1] + I_{shdS+M}[Y;(X_1,X_2)] + I_{syn}[Y;(X_1,X_2)]$$

and it is the basis of both the information decompositions described in Sections 4.1 and 4.2. Adapting the notation of [21] we express our joint input mutual information in four terms as follows:

$\text{Unq}X_1 \equiv I_{unq}[Y;X_1\|X_2]$	denotes the unique information that X_1 conveys about Y;
$\text{Unq}X_2 \equiv I_{unq}[Y;X_2\|X_1]$	is the unique information that X_2 conveys about Y;
$\text{Shar}_{\text{S+M}} \equiv I_{shdS+M}[Y;(X_1,X_2)]$	gives the common (or redundant or shared) information that both X_1 and X_2 have about Y;
$\text{Syn} \equiv I_{syn}[Y;(X_1,X_2)]$	is the synergy or information that the joint variable (X_1,X_2) has about Y that cannot be obtained by observing X_1 and X_2 separately.

It is possible to make deductions about a PID by using the following four equations which give a link between the components of a PID and certain classical Shannon measures of mutual information. The following are from Equations (4) and (5) in [21], with amended notation; see also [8].

$$I[Y;X_1] = \text{Unq}X_1 + \text{Shar}_{\text{S+M}}, \tag{18}$$

$$I[Y;X_2] = \text{Unq}X_2 + \text{Shar}_{\text{S+M}}, \tag{19}$$

$$I[Y;X_1|X_2] = \text{Unq}X_1 + \text{Syn}, \tag{20}$$

$$I[Y;X_2|X_1] = \text{Unq}X_2 + \text{Syn}. \tag{21}$$

We will refer to these results in Section 5 and use them in Section 6.

We consider here two different information decompositions. Although there are clear conceptual differences between the two, where they agree we can have some confidence we are accurately decomposing information as we would like. Where they disagree, we hope this may shed light on particular properties of the modulatory systems we study here, and also provide interesting comparisons of the two approaches.

It has been noted [22] that there are two different ways shared information can emerge. *Source* shared information refers to shared information that arises simply because the two inputs are correlated. For example, if $Y = X_1$ but X_1 and X_2 are correlated then there will be some $I(Y; X_2)$ and some redundancy $I_{\text{shdS+M}}[Y; (X_1 X_2)]$, even though X_2 plays no role in the computation implemented by the local processor. However, redundancy can also occur in systems where the inputs are statistically independent—in this case, it is referred to as *mechanistic* shared information, since it arises as a property of the function of the local processor. We denote I_{shdS+M} as the standard PID measure of shared information which quantifies both of these types together. However, both decompositions we consider provide a way to separately quantify these two types of shared information, which we denote by I_{shdS} and I_{shdM} for source and mechanistic respectively.

4.1. The Ibroja PID

In the Ibroja PID [9,10], the shared information component is based on an assumption that the information shared between two predictors about a target should not be affected by the marginal distribution of the two inputs (X_1, X_2) when the output is ignored. Instead, the shared information is a function only of the individual input-output marginal distributions of (Y, X_1) and (Y, X_2). In other words, the information about the output which is shared between the two inputs is independent of the correlation between the two inputs. In [9], this is motivated with an operational definition of unique information based on decision theory. It is claimed that unique information in input X_1 should correspond to the existence of a decision problem where two agents must try to guess the value of the output Y in which an agent acting optimally on evidence from X_1 can do systematically better (higher expected utility) than an agent acting optimally based on evidence from X_2; see also Appendix B2 in [21].

Following notation in [9], we consider a given joint distribution p for (Y, X_1, X_2), we let Δ be the set of all joint distributions of Y, X_1 and X_2, and define

$$\Delta_p = \{q \in \Delta : q(y, x_1) = p(y, x_1) \text{ and } q(y, x_2) = p(y, x_2), \text{ for all } (y, x_1, x_2) \in B^3\} \tag{22}$$

as the set of all joint distributions which have the same (Y, X_1) and (Y, X_2) marginal distributions as p.

In Lemma 4 in [9] five equivalent optimisation problems are defined involving various information components. In this work we chose to minimise the total mutual information $I[Y; (X_1, X_2)]$ in order to find the optimal distribution q, denoted by $\hat{q}$. For the description of EID in Section 4.2, we note that this is equivalent to finding the distribution in Δ_p which maximizes the co-information $I[Y; X_1; X_2]$. This optimal distribution $\hat{q}$ is then used to calculate the four partial information measures:

$$\begin{aligned}
\text{Unq}X_1 &= I_{\hat{q}}[Y; X_1 | X_2], && (23)\\
\text{Unq}X_2 &= I_{\hat{q}}[Y; X_2 | X_1], && (24)\\
\text{Shar}_{\text{S+M}} &= \text{I}_{\hat{q}}[Y; X_1; X_2], && (25)\\
\text{Syn} &= I_p[Y; (X_1, X_2)] - I_{\hat{q}}[Y; (X_1, X_2)], && (26)
\end{aligned}$$

and the information quantities, except $I_p[Y; (X_1, X_2)]$, are calculated with respect to the optimal distribution $\hat{q}$.

Using equations (7) & (8) from [23], the shared information can be split into non-negative *source* and *mechanistic* components that are defined as follows (in amended notation).

$$\begin{aligned} I_{shdS}[Y;(X_1,X_2)] = \max\{&\min(I_{shdS+M}[Y;(X_1,X_2)], I_{shdS+M}[X_1;(X_2,Y)]), \\ &\min(I_{shdS+M}[Y;(X_1,X_2)], I_{shdS+M}[X_2;(X_1,Y)])\} \\ I_{shdM}[Y;(X_1,X_2)] = I_{shdS+M}[&Y;(X_1,X_2)] - I_{shdS}[Y;(X_1,X_2)] \end{aligned}$$

A particular advantage of the Ibroja approach is that it results in a decomposition consisting of non-negative terms. A possibly counter-intuitive feature is that in our two input, one output local processor context, one might expect that $I_{shdS+M}[Y;(X_1,X_2)]$ should change depending on the marginal distribution of the inputs, (X_1,X_2), in that source shared information should increase as the correlation between the inputs increases (assuming the individual input-output marginals are fixed). In the systems defined in Section 2, however, the marginal distributions of (Y,X_1) and (Y,X_2) do depend on the correlation between the inputs, and so the Ibroja PID does change as this correlation changes.

4.2. The EID Using I_{ccs}

An alternative measure of shared information was recently proposed in [12]. Since at a local or pointwise level [24–28] (i.e., the terms inside the expectation), information is equal to change in surprisal, I_{ccs} seeks to measure shared information as the change in surprisal that is common to the input variables (hence CCS, Common Change in Surprisal). For two inputs, I_{ccs} is defined as:

$$\begin{aligned} I_{\text{ccs}}[Y;(X_1,X_2)] &= \sum_{y,x_1,x_2} p(y,x_1,x_2) h_y^{\text{com}}(x_1,x_2) \\ h_y^{\text{com}}(x_1,x_2) &= \begin{cases} i_{\tilde{q}}(y;x_1;x_2) & \text{if } \operatorname{sgn} i_{\tilde{q}}(y;x_1;x_2) = \operatorname{sgn} i_{\tilde{q}}(y;x_1) = \operatorname{sgn} i_{\tilde{q}}(y;x_2) = \operatorname{sgn} i_{\tilde{q}}(y;x_1,x_2) \\ 0 & \text{otherwise} \end{cases} \\ i_{\tilde{q}}(y;x_1;x_2) &= i_{\tilde{q}}(y;x_1) + i_{\tilde{q}}(y;x_2) - i_{\tilde{q}}(y;x_1,x_2) \\ \tilde{q} &= \arg\max_{q \in \Delta_p^2} \sum_{y,x_1,x_2} -q(y,x_1,x_2) \log q(y,x_1,x_2) \\ \Delta_p^2 &= \left\{ q \in \Delta : \begin{array}{l} q(y,x_1) = p(y,x_1), q(y,x_2) = p(y,x_2) \\ q(x_1,x_2) = p(x_1,x_2), \text{ for all } (y,x_1,x_2) \in B^3 \end{array} \right\} \end{aligned}$$

where lower case symbols indicate the local or pointwise values of the corresponding information measures, i.e., $I_{\tilde{q}}(Y;X_1) = \sum_{y,x_1} p(y,x_1) i_{\tilde{q}}(y,x_1)$. The sign conditions ensure that only terms corresponding to genuine shared information are included; terms not meeting the sign equivalence represent either synergistic or ambiguous effects [12].

This approach has two fundamental conceptual differences from the Ibroja PID. The first is that in [12] a game theoretic operational definition of unique information is introduced. This is very similar to the decision theoretic argument in [9] but extends the considered situations to include games where the utility function is asymmetric or the game is zero-sum. Both of these extensions induce a dependency on the marginal distribution of (X_1,X_2). A specific example system is provided in [12] as well as a specific game which demonstrates unique information even when there is none available from the decision theoretic perspective.

The second conceptual difference is the way in which shared information is actually measured, within the constraints imposed by the respective operational definitions. In the Ibroja PID, shared information is measured as the maximum co-information over the optimization space Δ_p. I_{ccs} also relies on co-information, but breaks down the pointwise contributions and includes only those terms that unambiguously correspond to redundant information between the inputs about the output. This is important because co-information conflates redundant and synergistic effects [8,12] so cannot itself be expected to fully separate them. I_{ccs} is calculated using the distribution with maximum entropy

subject to the game theoretic operational constraints (equality of all pairwise marginals). However, note that maximizing co-information subject to the extended game theoretic constraints is equivalent to maximizing entropy.

A decomposition of mutual information can be obtained using I_{ccs} following the partial information decomposition framework [8].

$$\begin{aligned}
\text{Unq}X_1 &= I[Y;X_1] - I_{\text{ccs}}[Y;(X_1,X_2)],\\
\text{Unq}X_2 &= I[Y;X_2] - I_{\text{ccs}}[Y;(X_1,X_2)],\\
\text{Shar}_{\text{S+M}} &= I_{\text{ccs}}[Y;(X_1,X_2)],\\
\text{Syn} &= I[Y;(X_1,X_2)] - I[Y;X_1] - I[Y;X_2] + I_{\text{ccs}}[Y;(X_1,X_2)],
\end{aligned}$$

The inclusion of $p(x_1,x_2)$ in the constraints for $\tilde{q}$ means that the measured shared and unique information is not invariant to the predictor-predictor marginal dependence. With I_{ccs} this affects the decomposition in an intuitive way: negative or no correlation between predictors results in more unique information, while when correlation between the predictors increases, shared information increases (driven by increased source shared information) and unique information decreases; see Figure 7 in [12]. However, the PID computed with I_{ccs} is not non-negative. In particular, the unique information terms can take negative values, which can be challenging to interpret.

In [13], it was recently suggested that the PID formalism could be applied to decompose multivariate entropy directly. The concepts of redundancy and synergy can apply just as naturally to entropy, resulting in a Partial Entropy Decomposition (PED) which can separate a bivariate entropy into four terms representing shared uncertainty, unique uncertainty in each variable, and synergistic uncertainty which arises only from the system as a whole. This approach shows that mutual information is actually the difference between redundant and synergistic entropy:

$$I[Y;X] = H_{\text{shd}}[(Y,X)] - H_{\text{syn}}[(Y,X)]$$

and this relationship holds for any measure of shared entropy which satisfies the PED axioms. This shows that mutual information does not only quantify common, shared or overlapping entropy, but is also affected by synergistic effects between the variables. At the global level since joint entropy is maximised when the two variables are independent (alternatively mutual information is non-negative), this implies that $H_{\text{shd}}[(Y,X)] \geq H_{\text{syn}}[(Y,X)]$. Mutual information is the expectation over local information terms that can themselves be positive, representing an decrease in the surprisal of event y when event x is observed, or negative, representing an increase in the surprisal of y when x is observed. Negative local information terms, which have been called "misinformation" [26], arise for symbols where $h(x,y) > h(x) + h(y)$; that is, those symbols provide a synergistic contribution to the joint entropy expectation sum. The existence of such locally synergistic entropy terms suggest that synergistic entropy is a reasonable thing to quantify within the PED framework. A shared entropy measure (H_{cs}) can be defined in a manner consistent with I_{ccs} as [13]:

$$\begin{aligned}
H_{\text{cs}}(Y,X_1,X_2) &= \sum_{y,x_1,x_2} \tilde{q}(y,x_1,x_2) h_{\text{cs}}(y,x_1,x_2)\\
h_{\text{cs}}(y,x_1,x_2) &= \max\left[-i_{\tilde{q}}(y,x_1,x_2), 0\right]
\end{aligned}$$

This entropy perspective can give some insight into the meaning of negative terms within the I_{ccs} PID. With I_{ccs}, shared information is calculated as shared entropy with the target that is common to both inputs (positive local co-information terms in I_{ccs}) minus synergistic entropy with the target that is common to both inputs (negative local co-information terms in I_{ccs}). Negative unique information terms can therefore arise when there is more unique synergistic entropy between a target and the predictor than there is unique shared entropy between the target and the predictor. Unique synergistic entropy means there is synergistic entropy between say X_2 and Y which is not shared with X_1. This can

arise for example, whenever the calculation of $I[Y; X_2]$ includes negative local terms in the expectation (for some values of y, x_2), but $I[Y; X_1]$ does not. In such cases, these negative local contributions to the mutual information must be unique; they do not appear in $I[Y; X_1]$ since that calculation has no negative terms.

The PED of our three variables also provides a way to separate the I_{ccs} shared information into mechanistic and source shared terms. The source shared information can be obtained from the three way partial entropy term, $H_{\text{shd}}[(Y, X_1, X_2)]$. This term represents the entropy that is common to all three variables, therefore it is included in the calculation of both $I[Y; X_1]$ and $I[Y; X_2]$ and so is shared information. However, it is possible that this quantity also includes some mechanistic shared information. This can only happen if $H_{\text{cs}}[(Y, X_1, X_2)] > H_{\text{cs}}[(X_1, X_2)]$—i.e., the two inputs share more entropy in the context of the full system then they do when ignoring (by marginalising away) the output. This corresponds to a negative partial entropy term $H_{\text{shd}}[(X_1, X_2)]$. Therefore we calculate source and mechanistic shared information, from Equation (32) in [13], as:

$$I_{\text{shdS}}[Y; (X_1, X_2)] = \min\left(H_{\text{cs}}[(Y, X_1, X_2)], H_{\text{cs}}[(X_1, X_2)]\right),$$
$$I_{\text{shdM}}[Y; (X_1, X_2)] = I_{\text{ccs}}[Y; (X_1, X_2)] - I_{\text{shdS}}[Y; (X_1, X_2)]$$

The first expression quantifies the source shared entropy: it is the three-way shared entropy with any mechanistic shared entropy removed. Since I_{ccs} quantifies source and mechanistic shared information together, we obtain the mechanistic shared information by subtracting off the calculated source shared information. Source shared information defined in this way is always positive, but mechanistic shared information can be negative. Negative mechanistic shared information can arise when, for example, both $I[Y; X_1]$ and $I[Y; X_2]$ contain negative local information terms, and those local information terms are common, reflected in a negative local co-information term. Alternatively, there is synergistic entropy between Y and X_1 that overlaps with synergistic entropy between Y and X_2. Synergistic entropy between the target and a predictor is by definition a mechanistic effect, since it is uncertainty that does not arise in the predictor alone, but is only obtained when the output (i.e., the mechanism) is considered. Please see [13] for further details. Since this approach relies on terms from the partial entropy decomposition as well as the partial information decomposition using I_{ccs}, we refer to it here as an Entropic Information Decomposition (EID).

5. Information Decomposition (ID) Spectra

We now describe a simple visual display [29] in which all the transmitted mutual information components appear, together with the residual output entropy. These displays are referred to as "spectra" because different colours are used for different components. Here the spectra are shown as stacked bar charts, which facilitates presentation of many spectra in a single figure. These spectra convey a simple but important message when applied to the goal of distinguishing between modulatory and additive interactions, whether in real or artificial neural systems. The important message is that modulatory and additive forms of interaction can have similar or even identical effects under some conditions, but very different effects under others. Such plots can also be used to compare the information processing performed in a system under different parameter regimes. They can also be used to compare the kinds of information processing performed by individual subjects or groups of subjects when completing psychophysical tasks; see Section 8.

5.1. Definition and Illustrations

The first five components are the partial information measures considered in Section 4: unique informations, shared source and mechanistic information and synergy. To this is added the residual output entropy.

The residual output entropy is $H(Y)_{res} = H(Y|X_1, X_2)$, which appears in the following decomposition, from Equation (6) in [21],

$$H(Y) = I_{unq}[Y; X_1|X_2] + I_{unq}[Y; X_2|X_1] + I_{shdS+M}[Y; (X_1, X_2)] + I_{syn}[Y; (X_1, X_2)] + H(Y|X_1, X_2) \quad (27)$$

and here we also use the decomposition

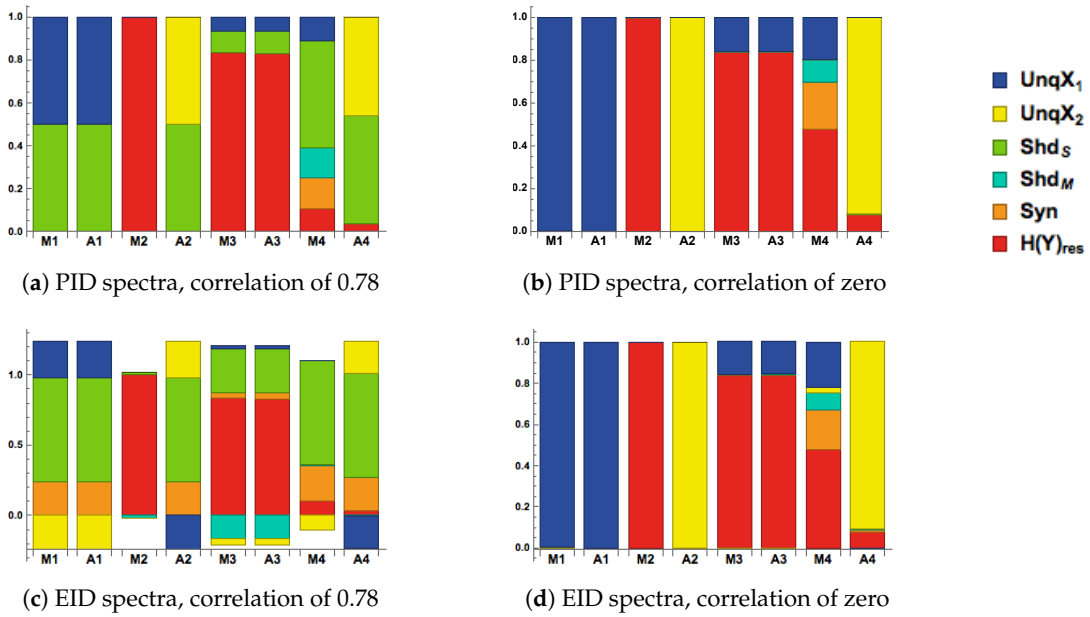
$$I_{shdS+M}[Y; (X_1, X_2)] = I_{shdS}[Y; (X_1, X_2)] + I_{shdM}[Y; (X_1, X_2)].$$

In our discussion, we consider four different spectra as an illustrative test set. First, we take $s_1 = 10.0$ and $s_2 = 0.05$ to represent the situation where the RF input is strong and the CF input is extremely weak. Secondly, in the case where $s_1 = 0.05, s_2 = 10.0$, the RF input is extremely weak while the CF input is strong. Thirdly, when $s_1 = 1.0, s_2 = 0.05$ the RF input is weak and the CF input is extremely weak. Finally, when $s_1 = 1.0, s_2 = 5.0$, the RF input is weak and the CF input is of moderate strength.

5.2. Ibroja Spectra

It is useful to bear in mind when interpreting these spectra that the information components are not independent quantities since they satisfy the constraints (18)–(21) and (27); so these non-negative components are negatively correlated. Figure 4a,b show PID decompositions when the two inputs have a correlation of either 0.78 or 0. In both cases modulatory and additive transfer functions lead to very similar decompositions when the RF input is strong (charts M1 and A1), or of moderate strength (charts M3 and A3), and the CF input is very weak, since there is little or no difference between charts M1 and A1 and between M3 and A3. Thus, when context is absent or very weak the modulatory transfer function becomes effectively equivalent to an additive function.

When the RF input is either very weak (charts M2 and A2) or less weak but with strong CF input (charts M4 and A4), modulatory and additive transfer functions have very different effects. Consider the case where the RF input is very weak and the CF is strong. The modulatory function transmits little or no input information (chart M2), implying that RF input is necessary to information transmission. In contrast, the additive transfer function in that case transmits information unique to the CF input with shared information if the two inputs are correlated (chart A2). Cases where RF input is present but weak show the modulatory effect of the CF input. Consider transmission in the case of weak RF input with extremely weak CF input (charts M3 and A3). The output residuals are then high, showing that little information is transmitted. What is transmitted is a combination of shared information and information unique to the RF input. If the RF input is weak but the CF input is strong, however, then the modulatory function transmits more unique information about the RF than when the CF input is weak, together with some synergy, some mechanistic shared, and some source shared if the inputs are correlated (chart M4). In contrast, the additive transfer function transmits no information unique to the RF but only information unique to the CF and shared information if the inputs are correlated (chart A4).



(**a**) PID spectra, correlation of 0.78

(**b**) PID spectra, correlation of zero

(**c**) EID spectra, correlation of 0.78

(**d**) EID spectra, correlation of zero

Figure 4. Partial information decomposition (PID) and entropic information decomposition (EID) spectra (in bits), based on additive (A) and modulatory (M) transfer functions for four combinations of signal strengths: 1. ($s_1 = 10.0, s_2 = 0.05$), 2. ($s_1 = 0.05, s_2 = 10.0$), 3. ($s_1 = 1.0, s_2 = 0.05$), 4. ($s_1 = 1.0, s_2 = 5.0$), and two values of the correlation between inputs: 0.78 and zero.

5.3. EID Spectra

The EID spectra can have negative partial information measures, and so when interpreting them it is useful to bear in mind the constraints (18)–(21). Therefore, for example, if the UnqX_1 component is negative then, since the classical Shannon measures are fixed, it would follow from (18) and (20) that the components Shar$_{S+M}$ and Syn would be larger than if the UnqX_1 component were equal to zero; of course the component Shar$_{S+M}$ is split further into *Source* and *Mechanistic* terms, as discussed in Section 4. In particular, if it were the case that $I[Y; X_1|X_2]$ were equal to zero then the synergy component would be positive and equal in magnitude to the UnqX_1 component. Therefore, when a negative component is present this is likely to make the relative magnitudes of the partial information components appear different than in the corresponding Ibroja spectra, even though the same essential message might be being expressed.

Consider Figure 4c. We note that the use of the modulatory and the additive transfer functions leads to very similar spectra in charts M1 and A1, and M3 andA3. In charts M1 and A1, we see that when the RF input is strong the residual output is zero and the information is transmitted mainly via the source-shared component, but with some synergy and some unique information about the RF, as well as some unique misinformation from the CF. Charts M2 and A2 reveal a marked difference in the spectra due to the transfer functions. When the modulatory transfer function is employed and the RF input is extremely weak then almost no information is transmitted. In contrast, the use of the additive transfer function leads to all the information being transmitted, mostly in the form of source shared information, with some synergy, some unique information about the CF and some misinformation from the RF. In charts M3 and A3, the output residual is very high and so very little information is transmitted when the RF input is weak and the CF input is extremely weak, and what is transmitted is a combination of positive source shared information and negative mechanistic shared information. Chart M4, where the CF input is moderate but the RF input is weak, indicates that more information about the RF is transmitted than was the case in chart M3, since the output residual is smaller. This information is transmitted mainly via source shared information and synergy, with some unique misinformation from the CF.

We now briefly consider Figure 4d. Charts M1 and A1 show that all the information is transmitted in a form unique to the RF. We see a striking difference between charts M2 and A2, with no information being transmitted in M1 and all the information unique to the CF being transmitted in A2. Charts M3 and A3 appear to be identical, with some information unique to the RF being transmitted and a high output residual. Chart M4 shows that about one-half of the information is transmitted, mainly due to that unique to the RF and synergy but also with some mechanistic shared and a little unique to the CF. Much more information is transmitted in A4, predominantly in a form unique to the CF. A pleasing feature of Figure 4d is that the source shared information component is zero in all the charts, while the mechanistic shared component in chart M4 is positive; this is exactly what would be expected when the inputs are uncorrelated, and here there are no negative mechanistic shared components unlike in Figure 4c where the inputs are strongly correlated.

5.4. Contextual Modulation and Information Decompositions

In Section 3, the conditions M1–M3 express the notion of contextual modulation. Here, we translate these conditions using (18)–(21) into corresponding expressions of contextual modulation for ID measures, denoted by S1–S3 for non-negative decompositions, with amended conditions S1′–S2′ for the EID when it has negative components.

S1: If the *RF* signal is strong enough, and the CF input is extremely weak, then both UnqX2 and Syn are close to zero, UnqX1 can have its maximum value, and the sum of UnqX1 and $\text{Shar}_{\text{S+M}}$ can equal the total output entropy. This shows that the RF input is sufficient, thus allowing the information in the *RF* to be transmitted, and that the CF input is not necessary.

S2: All five partial information components are close to zero when the RF input is extremely weak no matter how strong the CF input. This shows that the RF input is necessary for information to be transmitted, and that the CF input is not sufficient to transmit the information in the RF input.

S3: When $s_1 < s_2$ and when the RF input is weak, then the sum of UnqX1 and Syn is larger when the CF input is moderate than it is when the CF input is weak. The same is true of the sum of UnqX1 and $\text{Shar}_{\text{S+M}}$. Thus the CF input modulates the transmission of information about the RF input.

The following conditions provide amendments to S1-S2 when the EID has negative components:

S1′: When UnqX2 < 0, UnqX2 and Syn are approximately of the same magnitude, the sum of UnqX1 and Syn can have its maximum value, and the sum of UnqX1 and $\text{Shar}_{\text{S+M}}$ can equal the total output entropy.

S2′: If at least one component is negative, then we can set the left-hand sides of (18)–(21) to zero and use the rule that the sum of the magnitudes of the negative components is approximately equal to the sum of the magnitudes of the positive components. If in any of (18)–(21) there is no negative term then all terms on the right-hand side are close to zero.

We now discuss the spectra in relation to the these conditions. First we discuss the PID charts in Figure 4a. In charts M1 and A1, we see that Syn and UnqX2 are apparently equal to zero and that the sum of UnqX1 and $\text{Shar}_{\text{S+M}}$ is equal to 1, the value of the total output entropy; UnqX1 is equal to 0.5 which is presumably the maximum value it can take. Therefore Condition S1 is satisfied for the modulatory and the additive transfer function. For charts M2 and A2, we see in M2 that all five of the components are apparently zero, and hence condition S2 holds for the modulatory transfer function, but this is not the case with the additive transfer function in A2 since the values of UnqX2 and Shar_{S} are appreciable. Inspection of charts M3 and M4 shows that the sum of Syn and UnqX1 and the sum of UnqX1 and $\text{Shar}_{\text{S+M}}$ are larger in M4 than in M3, thus supporting condition S3. In charts A3 and A4 we see the same for the sum of UnqX1 and $\text{Shar}_{\text{S+M}}$, but the opposite for the sum of Syn and UnqX1, and so S3 is not fully supported in the additive case.

We now consider the EID charts in Figure 4c. In charts M1 and A1, UnqX2 is negative and UnqX2 and Syn have approximately the same magnitude. Therefore, the sum of UnqX1 and $\text{Shar}_{\text{S+M}}$

is equal to 1, the value of the total output entropy. Also, UnqX1 is just larger than 0.2, presumably the largest value it can take. Therefore, the conditions of S1′ are satisfied in both the modulatory and additive cases. For charts M2 and A2, we see in M2 that the residual output entropy is almost equal to 1, that UnqX1, UnqX2 and Syn are apparently zero and that the little negative mechanistic shared information is counterbalanced by a similar amount of positive source shared information, thus supporting condition S2′, since all the right-hand sides in (18)–(21) are close to zero. This condition is, however, not supported in the additive case since the values of UnqX2, Shar_{S}, Syn and UnqX1 (negative) are all appreciable. Considering charts M3 and M4, we notice that the sum of Syn and UnqX1 and also the sum of UnqX1 and $\text{Shar}_{\text{S+M}}$ are larger in M4 than in M3, thus supporting condition S3. In charts A3 and A4 we see the same for the sum of UnqX1 and $\text{Shar}_{\text{S+M}}$, but the opposite for the sum of Syn and UnqX1, and so S3 is not fully supported in the additive case. Hence, when the correlation between inputs is strong, we find that the conclusions for both PID and EID are the same with regard to the use of modulatory and additive transfer functions.

In Figure 4b,d, the respective PID and EID spectra are virtually identical, and so the same conclusions will hold for both decompositions. In charts M1 and A1, UnqX2 and Syn are apparently zero, the sum of UnqX1 and $\text{Shar}_{\text{S+M}}$ is equal to the total output entropy and this time UnqX1 is fully maximized. Therefore condition S1 is supported in both charts. In chart M2 the residual output entropy is close to 1 and so all five information components are close to zero, thus supporting condition S2. We notice that the sum of Syn and UnqX1 and also the sum of UnqX1 and $\text{Shar}_{\text{S+M}}$ are larger in M4 than in M3, thus supporting condition S3. In charts A3 and A4 we see that both these sums are smaller in A4 than in A3, and so S3 is not supported in the additive case.

5.5. Comparison of PID and EID

Close comparison of the EID and PID spectra sheds light on both the information processing properties of the form of modulation considered here, and on relations between PID and EID. Most importantly for the purposes of this paper both PID and EID show the distinctive properties of the modulatory interaction, in which the modulatory transfer function is employed. First, no information dependent on the inputs is transmitted when the RF input is very weak whatever the value of the CF input. This shows that the RF input is necessary for this transfer function to transmit information about the input and that the CF input is not sufficient. Second, information is transmitted about the RF input for all states of the CF input including those in which it is absent or very weak. This shows that the RF input is sufficient for this transfer function to transmit information about the input and that the CF input is not necessary. Third, when the RF input is strong no information dependent on the CF input is transmitted by the output, but when the RF input is present but weak then the output transmits less information dependent on the the RF input when context is very weak.

This shows the modulatory effect of the CF input. Fourth, modulatory interactions produce the same components as additive interactions when the CF input is very weak, but very different components when the CF input is stronger and the RF input is present but weak. This shows conditions that distinguish these two forms of interaction. In general, the two inputs have equivalent opportunities to effect the output for additive interactions, whereas the effects of the CF input are conditional upon the RF input for the modulatory interaction. Fifth, when the two inputs are uncorrelated there is little difference between the EID and PID decompositions other than the splitting of shared into source and mechanistic by EID.

The spectra displayed may also shed some light on the negative components of EID, which still await a clear and widely accepted interpretation. First, negative components are zero or tiny when the two inputs are uncorrelated. Second, synergy and source shared were never negative in the conditions studied. Third, negative unique components seem to be compensated for by positive synergistic components. Fourth, source shared is never negative and positive only when the two inputs are correlated. Whether these observations will aid interpretation of the negative components remains to be seen.

The spectra shown here are all for specific values of the two input strengths, so to see whether the observations listed in the two preceding paragraphs hold for other values of those strengths the following section presents surfaces showing each of the output components that depend on input as a function of the two input strengths.

6. Analysis of the Transfer Functions Using the Ibroja PID over a Wide Range of Input Strengths

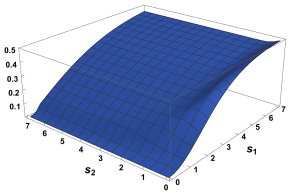
The five Ibroja surfaces were constructed as a function of the RF and CF signal strengths, s_1 and s_2. In Figure 5, we notice the striking differences in the surfaces for each measure between the use of the modulatory and the additive transfer function. In Figure 5b,d, there is a clear asymmetry that mimics that shown in Figure 2d,f.

We notice, in particular, that it appears that $\mathrm{Unq}X_1$ is zero when $s_2 > s_1$ while $\mathrm{Unq}X_2$ is zero for $s_1 > s_2$. In Figure 5a, $\mathrm{Unq}X_1$ rises towards its maximum as s_1 increases, and the rise is similar for $s_2 > 2$. For $s_1 > 2$ the shape of this plot matches that in Figure 2c. In Figure 5c, we note that $\mathrm{Unq}X_2$ appears to be zero for all values of s_1 and s_2. In Figure 5f,h,j plots of $\mathrm{Shar_S}$, $\mathrm{Shar_M}$ and Synergy are symmetric about the line $s_1 = s_2$ when based on the additive transfer function, and the maximum values of $\mathrm{Shar_M}$ and Synergy happen along the line $s_1 = s_2$, while $\mathrm{Shar_S}$ flattens quite quickly onto a plateau for most values of s_1 and s_2. On the other hand, there is no symmetry in Figure 5e,g,i, where the surfaces of $\mathrm{Shar_M}$ and Synergy rise and fall as s_1 increases and the pattern is similar for $s_2 > 2$, while the $\mathrm{Shar_S}$ surface rises quickly onto a plateau. The plot of synergy in Figure 5g appears to match exactly the plot of $I[Y; X_2|X_1]$ in Figure 2e, as expected, since it appears from Figure 5c that $\mathrm{Unq}X_2 = 0$.

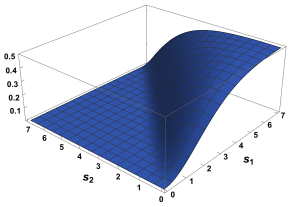
In Figure 6, the surfaces for $\mathrm{Unq}X_1$ and $\mathrm{Unq}X_2$ are similar to the corresponding plots in Figure 5. In particular, we note that again it appears from Figure 6c that $\mathrm{Unq}X_2 = 0$. Again, Figure 6g appears to match the corresponding plot of $I[Y; X_2|X_1]$ in Figure 2e. In Figure 6e,f, the $\mathrm{Shar_S}$ surface is zero for all values of s_1 and s_2; this is expected since the source shared information should be zero when the inputs are uncorrelated. By inspecting the surfaces in Figures 5e,f and 6e,f, we notice (as expected) that the source shared information is much larger when the inputs are strongly correlated than when they are uncorrelated. The plots of mechanistic shared information in Figures 5g,h and 6g,h indicate that the presence of strong correlation does not have much effect. In Figure 6h,j, symmetry is again apparent, with the maximum values occurring along the line $s_1 = s_2$.

Of special interest is the finding that $\mathrm{Unq}X_2$ appears to be zero. This suggests that X_2 can modify the transmission of information from the receptive field input X_1 to the output Y without transmitting any unique information about itself. This conclusion would be much stronger if it were possible to prove mathematically that $\mathrm{Unq}X_2 = 0$, given the system defined in Sections 2 and 3. We now state some formal results which indicate that this is indeed the case. We also define a class of transfer functions, that includes our modulatory transfer function T_M, for which $\mathrm{Unq}X_2 = 0$.

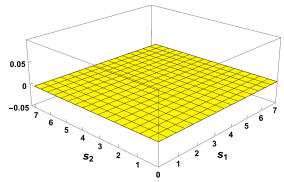
We saw also in the surfaces of $\mathrm{Unq}X_1$ and $\mathrm{Unq}X_2$, produced by the additive transfer function, that $\mathrm{Unq}X_2$ appears to be zero when $s_1 > s_2$, and also that $\mathrm{Unq}X_1$ appears to be zero when $s_1 < s_2$. We also state some mathematical results to confirm these impressions, as well as proving that when $s_1 = s_2$ both uniques are zero. Then, using (18)–(21), the exact Ibroja decomposition is derived. Proofs are given in the appendix. We now state the results.



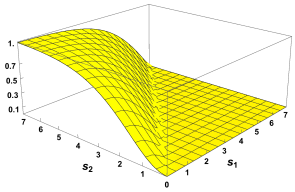
(**a**) Modulatory, UnqX_1



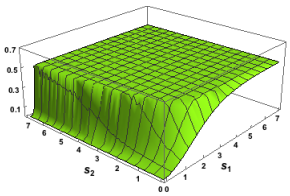
(**b**) Additive, UnqX_1



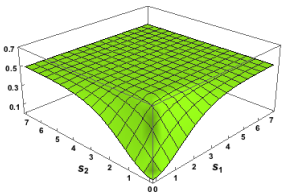
(**c**) Modulatory, UnqX_2



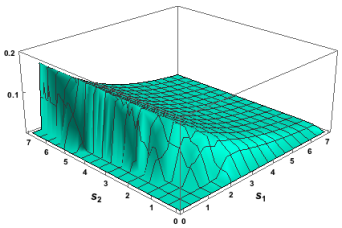
(**d**) Additive, UnqX_2



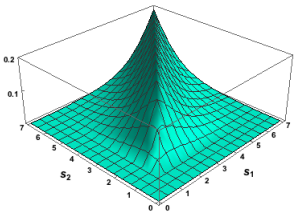
(**e**) Modulatory, Shar$_S$



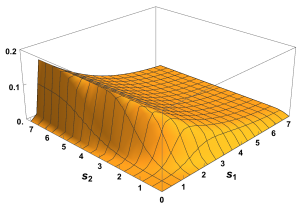
(**f**) Additive, Shar$_S$



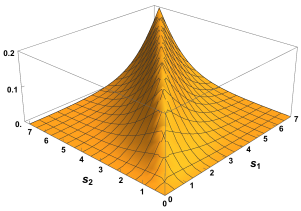
(**g**) Modulatory, Shar$_M$



(**h**) Additive, Shar$_M$

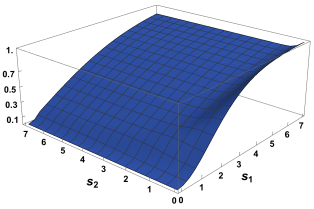


(**i**) Modulatory, Synergy

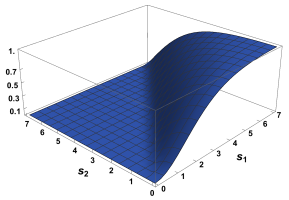


(**j**) Additive, Synergy

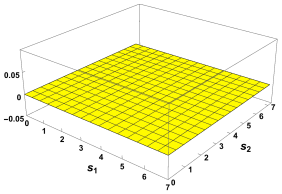
Figure 5. Ibroja surfaces, based on additive and modulatory transfer functions, and a correlation between inputs of 0.78.



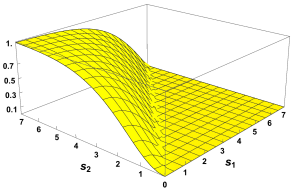
(**a**) Modulatory, UnqX_1



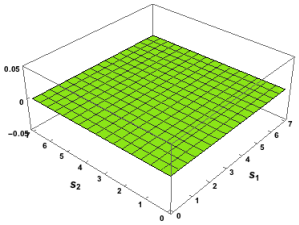
(**b**) Additive, UnqX_1



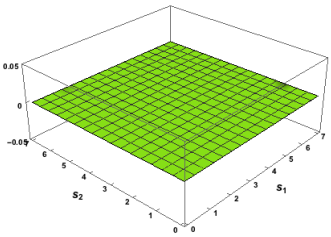
(**c**) Modulatory, UnqX_2



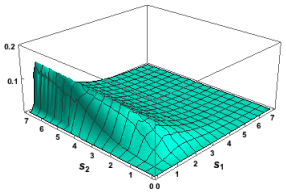
(**d**) Additive, UnqX_2



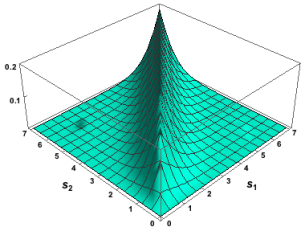
(**e**) Modulatory, Shar$_S$



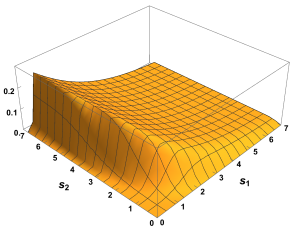
(**f**) Additive, Shar$_S$



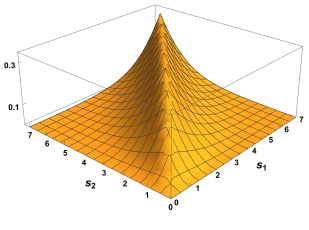
(**g**) Modulatory, Shar$_M$



(**h**) Additive, Shar$_M$



(**i**) Modulatory, Synergy



(**j**) Additive, Synergy

Figure 6. Ibroja surfaces, based on additive and modulatory transfer functions, and zero correlation between inputs.

Let F be a function of two real variables, x, y, which has the property that

$$F(-x,-y) = -F(x,y) \quad \text{and} \quad F(-x,y) = -F(x,-y), \text{ for } x > 0, y > 0. \tag{28}$$

We consider $F(r, c)$ as a transfer function, for integrated RF input r and integrated CF input c, and, as in Section 3, we pass the value of F through a logistic nonlinearity to obtain output conditional probabilities of the form, with $r = s_1 x_1$ and $c = s_2 x_2$,

$$\Pr(Y = 1 | X_1 = x_1, X_2 = x_2) = 1/(1 + \exp[-F(s_1 x_1, s_2, x_2)]). \tag{29}$$

We also assume that the joint p.m.f. for (X_1, X_2) has the form given in (1)–(3).

Theorem 2. *For the trivariate probability distribution defined in* (1)–(3), (29) *and a transfer function as defined in* (28), *suppose that* $g \geq \frac{1}{2}$ *and* $h \geq \frac{1}{2}$ *but* g *and* h *are not both equal to* $\frac{1}{2}$, *where* g *and* h *are defined by*

$$g = \Pr(Y = 1 | X_1 = 1, X_2 = 1) \quad \textit{and} \quad h = \Pr(Y = 1 | X_1 = 1, X_2 = -1). \tag{30}$$

Suppose also that $\lambda \neq 0$, $\mu \neq 0$, $s_1 > 0$, $s_2 > 0$. *Then, for such a system,* $UnqX_2 = 0$ *in the Ibroja PID.*

The conclusion of Theorem 2 also holds when the conditions on g, h are: $g \leq \frac{1}{2}, h \leq \frac{1}{2}$ but both g, h are not equal to $\frac{1}{2}$. The conclusion also holds when $g = \frac{1}{2}, h = \frac{1}{2}$, although in this case all of the information components are zero since the total mutual information $I[Y; (X_1, X_2)] = 0$, because Y is independent from (X_1, X_2).

We now state the results for the two transfer functions used in this study.

Corollary 1. *If the modulatory transfer function* T_M *is used in the system described in Theorem 2, and under the conditions stated there, then* $UnqX_2 = 0$ *in the Ibroja PID.*

Corollary 2. *If the additive transfer function* T_A *is used in the system described in Theorem 2, and under the conditions stated there, then* $UnqX_2 = 0$ *in the Ibroja PID when* $s_1 \geq s_2$.

It is shown by Theorem 2 that there is a general class of transfer functions which, when used in the system described in Sections 2 and 3, and which satisfy the conditions of the Theorem 2, have the property of not transmitting any unique information about the modulator. The modulatory transfer function used in this work is a member of this class. The additive transfer function T_A is also a member of this class but it does not satisfy the conditions required in Theorem 2 for all values of s_1 and s_2.

We now present a result regarding $\mathrm{Unq}X_1$ and $\mathrm{Unq}X_2$ when the additive transfer function is used in the system considered in Sections 2 and 3.

Theorem 3. *For the trivariate probability distribution defined in Sections 2 and 3, with the additive transfer function* T_A, *suppose that* $\lambda \neq 0$, $\mu \neq 0$, $s_1 > 0$, $s_2 > 0$. *Then, for such a system,* $UnqX_1 = 0$ *in the Ibroja PID when* $s_1 \leq s_2$. *When* $s_1 = s_2$ *then both* $UnqX_1$ *and* $UnqX_2$ *are zero in the Ibroja PID.*

Given the results of Theorems 2 and 3, and since the Ibroja PID is a non-negative decomposition, we can now state the following exact results.

Theorem 4. *For the trivariate probability distribution defined in* (1)–(4), *suppose that* $\lambda \neq 0$, $\mu \neq 0$, $s_1 > 0$, $s_2 > 0$. *Then, with* u_M, v_M, u_A, v_A *defined in* (15)–(16), *we have*

(a) When transfer function T_M *is employed then*

(i) $Syn = I(Y; X_2 | X_1) = h(z_M) - 2\lambda h(u_M) - 2\mu h(v_M)$;

(ii) $Shar_{S+M} = I(Y; X_2) = 1 - h(w_M)$;

(iii) $UnqX_1 = I(Y; X_1 | X_2) - I(Y; X_2 | X_1) = h(w_M) - h(z_M)$, *and* $UnqX_2 = 0$.

(b) *When the transfer function T_A is used and $s_1 = s_2$ then*

(i) $Syn = I(Y; X_2|X_1) = h(z_A) - 2\lambda h(u_A) - 2\mu;$

(ii) $Shar_{S+M} = I(Y; X_1) = 1 - h(z_A);$

(iii) $UnqX_1 = UnqX_2 = 0;$

(c) *When the transfer function T_A is used and $s_1 < s_2$ then*

(i) $Syn = I(Y; X_1|X_2) = h(w_A) - 2\lambda h(u_A) - 2\mu h(v_A);$

(ii) $Shar_{S+M} = I(Y; X_1) = 1 - h(z_A);$

(iii) $UnqX_2 = I(Y; X_2|X_1) - I(Y; X_1|X_2) = h(z_A) - h(w_A) \quad and \quad UnqX_1 = 0.$

(d) *When the transfer function T_A is used and $s_1 > s_2$ then*

(i) $Syn = I(Y; X_2|X_1) = h(z_A) - 2\lambda h(u_A) - 2\mu h(v_A);$

(ii) $Shar_{S+M} = I(Y; X_2) = 1 - h(w_A);$

(iii) $UnqX_1 = I(Y; X_1|X_2) - I(Y; X_2|X_1) = h(w_A) - h(z_A), \quad and \quad UnqX_2 = 0.$

For the trivariate binary system considered in Sections 2 and 3, these results show that the Ibroja PID is a minimum mutual information PID, as was found in [30,31] for the trivariate Gaussian system. Finally, we give the PID for any non-negative decomposition in the case where $\lambda = 0$ or $\mu = 0$, so that the correlation between inputs is -1 or $+1$, respectively.

Theorem 5. *Consider the probability distribution defined in* (1)–(4). *When the correlation between the inputs, X_1, X_2, is $+1$, we have that*

(a) $UnqX_1 = UnqX_2 = Syn = 0$, *and* $Shar_{S+M} = 1 - h(u)$.

when the correlation between the inputs, X_1, X_2, is -1, we have that

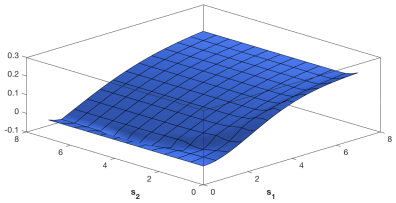
(b) $UnqX_1 = UnqX_2 = Syn = 0$, *and* $Shar_{S+M} = 1 - h(v)$,

where, from (15)–(16), *$u = u_M, v = v_M$ when the transfer function T_M is employed and $u = u_A, v = v_A$ when the transfer function T_A is used.*

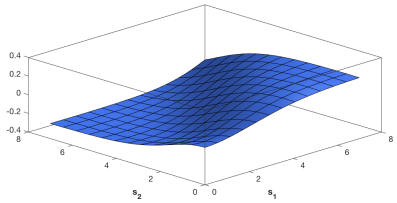
7. Analysis of the Transfer Functions Using EID over a Wide Range of Input Strengths

As in the previous section, five EID surfaces were constructed as a function of the RF and CF signal strengths, s_1 and s_2, in the definition of the trivariate binary system. Many of the properties of the resulting surfaces are common with the Ibroja PID surfaces: the opposite asymmeteries of the unique information terms for the additive system (Figures 7b,d and 8b,d), the symmetry in s_1 and s_2 of the other terms for the additive transfer function, and the asymmetries for the modulatory transfer function where the surfaces are relatively constant along the s_2 axis. However, there are also some differences, most noticeably the presence of negative terms.

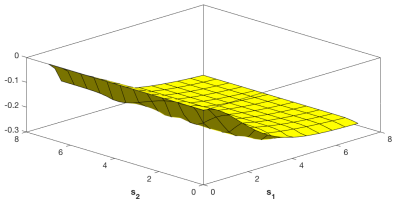
Figure 7c shows that for the modulatory transfer function, the EID shows negative unique information about X2.



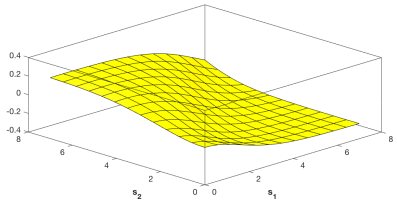
(**a**) Modulatory, UnqX_1



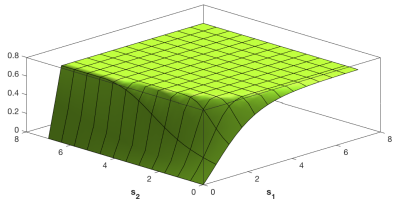
(**b**) Additive, UnqX_1



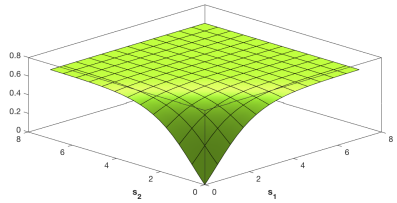
(**c**) Modulatory, UnqX_2



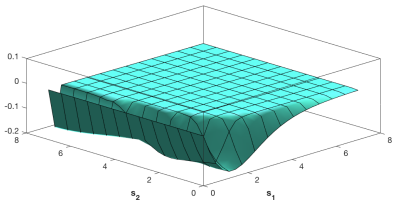
(**d**) Additive, UnqX_2



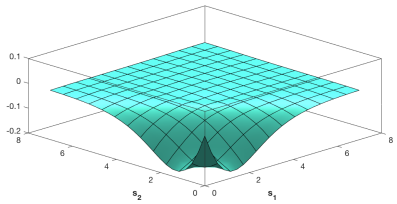
(**e**) Modulatory, Shar$_S$



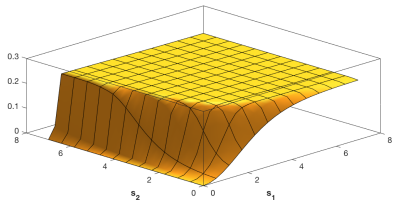
(**f**) Additive, Shar$_S$



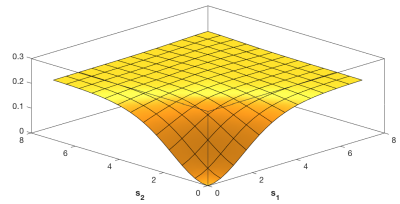
(**g**) Modulatory, Shar$_M$



(**h**) Additive, Shar$_M$

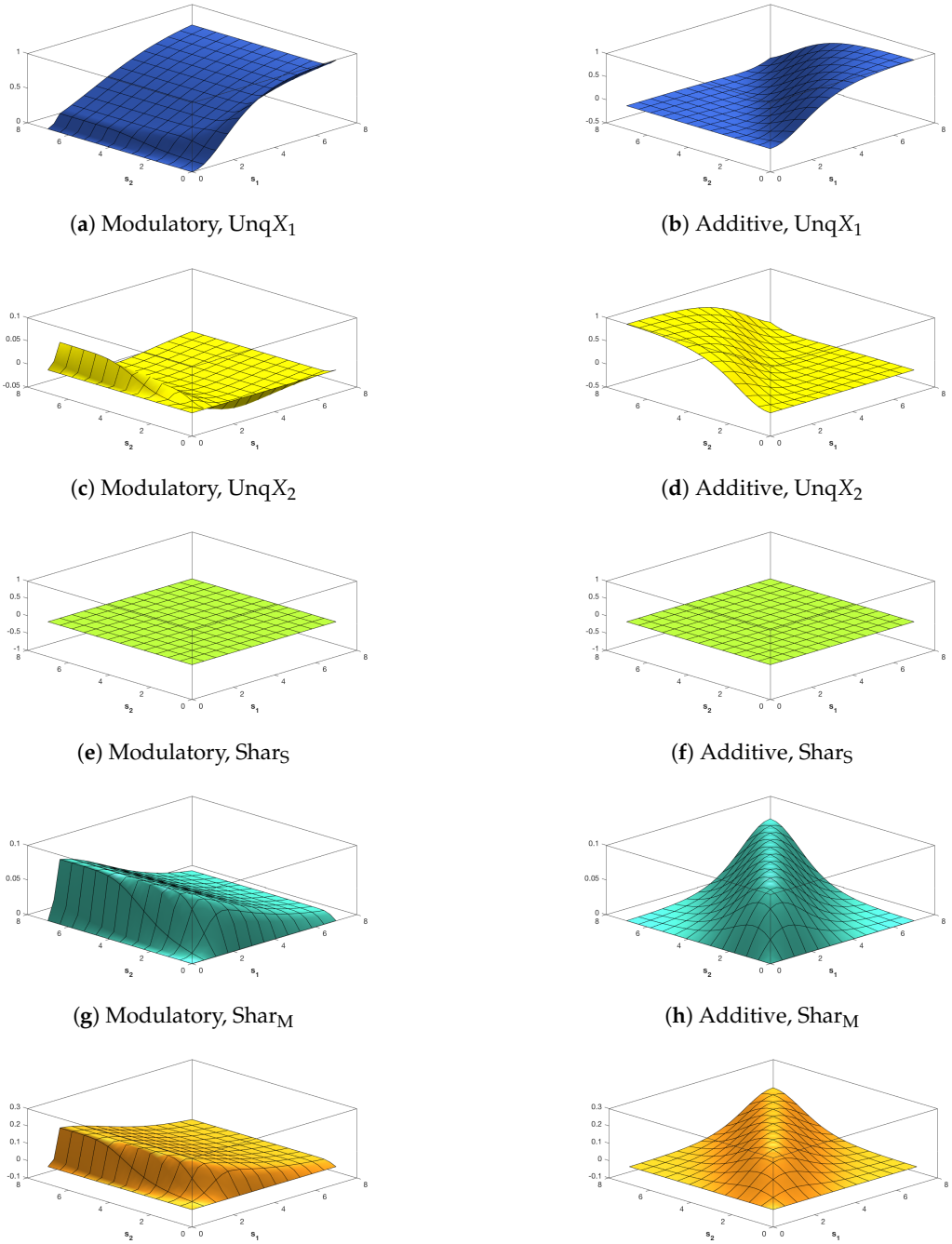


(**i**) Modulatory, Synergy



(**j**) Additive, Synergy

Figure 7. EID surfaces, based on additive and modulatory transfer functions, and a correlation between inputs of 0.78.



(**a**) Modulatory, UnqX_1 (**b**) Additive, UnqX_1

(**c**) Modulatory, UnqX_2 (**d**) Additive, UnqX_2

(**e**) Modulatory, Shar$_S$ (**f**) Additive, Shar$_S$

(**g**) Modulatory, Shar$_M$ (**h**) Additive, Shar$_M$

(**i**) Modulatory, Synergy (**j**) Additive, Synergy

Figure 8. EID surfaces, based on additive and modulatory transfer functions, and a zero correlation between inputs.

This is relatively constant irrespective of the strength of the CF signal, and increases in magnitude with stronger RF signals. Shar$_{S+M}$ is here split into separate source and mechanistic components. The source shared information for the modulatory transfer function plateaus for $s_1, s_2 > 2$ (Figure 7e). In this case, there is a very strong correlation between the two inputs, which is reflected in the shared source information. The source shared information is fixed due to the high correlation between the inputs, however, the univariate information in the CF decreases as a function of s_1. Therefore the unique X_2 information is negative. Similarly, as $I[Y; X_2|X_1]$ is UnqX_2 plus synergy in (21), the negative unique interacts with the plateau of positive synergy to result in the $I[Y; X_2|X_1]$ surface (Figure 2e).

In Figure 7g, we note that the mechanistic shared component is negative for small values of s_1, while in Figure 7h it is negative for some small values of s_1, s_2. In contrast, Figure 8g,h show that the mechanistic component is non-negative when the correlation between the inputs is zero.

In general, the univariate mutual information $I[Y; X_2]$ is a sum of positive and negative terms, representing shared and synergistic entropy respectively between the two variables in the calculation. Since mutual information is non-negative, the positive terms always outweigh the negative terms in the mutual information expectation summation. However, if some of the positive terms in the calculation of $I[Y; X_2]$ are shared, or overlapping, with corresponding positive local information terms of $I[Y; X_1]$, those terms will contribute to the shared information term of the decomposition, and not be counted in the unique information terms. If enough of the shared entropy between X_2 and Y is overlapping with that shared between X_1 and Y, and the negative synergistic entropy terms in $I[Y; X_2]$ are not shared with X_1, then the unique synergistic entropy between Y and X_2 can be larger than the unique redundant entropy between Y and X_2, resulting in a net negative UnqX_2 information term.

To illustrate this consider a specific example, when $s_1 = s_2 = 2$, with correlation between inputs of 0.78. We can consider the local contributions to the univariate mutual information $I[Y; X_1]$. As $I[Y; X_1]$ is an expectation computed with a summation we can consider each local term in the summation which we denote $e(y, x_1) = p(y, x_1)i(y, x_1)$:

$$\begin{aligned} e(-1,-1) = e(1,1) &= 0.46 \\ e(-1,1) = e(1,-1) &= -0.06 \end{aligned}$$

and similarly for $I(Y; X_2)$, the $e(y, x_2)$ are:

$$\begin{aligned} e(-1,-1) = e(1,1) &= 0.40 \\ e(-1,1) = e(1,-1) &= -0.11 \end{aligned}$$

Note that here the strong similarity in the profile of the local information terms results from the high correlation between the two inputs. Local co-information values when $x_1 = x_2 = y = -1$ and when $x_1 = x_2 = y = 1$ show that the terms are largely, but not completely, overlapping (0.37 bits). There are no other local contributions to the I_{ccs} shared information measure.

Further consideration of these pointwise terms reveals that there are some positive and some negative local unique contributions to the univariate information for both predictors. The shared local information for the state $(y, x_1, x_2) = (-1, -1, -1)$ is 0.37 bits. The corresponding $(y, x_1) = (-1, -1)$ term in the calculation of $I(Y; X_1)$ gives 0.46 bits of information. Since 0.37 bits of that is shared with X_2, $0.46 - 0.37 = 0.09$ bits are unique to X_1 for that local contribution. Similarly there is a contribution of 0.09 bits of unique X_1 information when $(y, x_1) = (1, 1)$. Considering the same local terms for X_2 there are again 0.37 bits shared with X_1 and now $0.40 - 0.37 = 0.03$ bits of unique X_2 information. So in total, when the output matches the RF X_1 input, those states contribute 0.18 bits to the unique X_1 information and 0.06 bits to the unique X_2 information.

Moving to the cross-terms, since there is no corresponding local shared information these contributions to the univariate mutual information are entirely unique. So for X_1 the unique information is $2 \times -0.06 = -0.12$ bits, and X_2 has $2 \times -0.11 = -0.22$ bits of unique information. So the total net unique information in X_1 is $0.18 - 0.12 = 0.06$ bits, and for X_2 there are $0.06 - 0.22 = -0.16$ bits of unique information. This shows that in this system both variables have both positive and negative contributions to unique information, and that a negative value results when the negative contributions are larger.

In this case, when the sign of either input matches the sign of the output, they have locally redundant entropy, some of which is shared with the other input, but a small fraction of which is unique to that variable (i.e., related to the residual variance over that determined by the correlations between the variables). Instead, when the sign of the input does not match the sign of the output,

there is local synergistic entropy between the variables. In other words, that particular local value of the input variable is misleading about the corresponding local output value, in the following sense.

Imagine a gambler was trying to predict the output of the system, starting with knowledge of the marginal distribution of the output $p(Y)$. They would determine a gambling strategy to optimise payout based on that distribution of Y. Observing the value of an input variable, combined with knowledge of the function of the system, would allow the gambler to form a new distribution of the output, $p(Y|X_2 = x_2)$. In this updated conditional distribution some specific values of the output would have higher probability than under $p(Y)$, and some would have lower probability. In the alternate sign cross terms in this example, the actual outcome is one of those that had lower probability under the conditional distribution obtained after observing the input. The particular (local) evidence provided by the value of the input on that trial moved the conditional distribution in the wrong direction for that output value—i.e., it was misleading about that particular output value, because it suggested it was less likely to happen, but then it did happen anyway. The fact that negative local values correspond to misleading evidence from the perspective of prediction explains why they have been termed misleading information or "misinformation" [26].

Therefore for both variables there are some unique information contributions that are both positive and negative (positive when the sign of the input is preserved in the output, and negative when the sign is changed in the output). Because a change in the sign of the output is rare, as a consequence of the design of the transfer function, that joint event is less likely to happen than would be predicted from the independent marginal local probability of the two events. The surprisal of the joint event is greater than the sum of the surprisal of the individual events. In conditional probability terms, $p(y|x_1) < p(y)$, the likelihood of seeing that value of y is decreased by conditioning on that value of x_1.

While in Figure 7c, the unique X_2 information is always negative, as shown in the example above there can be both positive and negative components. It would be possible to further split I_{ccs} to consider positive and negative terms separately, and so keep these shared vs. synergistic entropy effects separate throughout the decomposition. However here we focus on the net unique information effects to present a simpler decomposition and one that can be directly compared with the I_{broja} PID. Note that in Figure 8c the balance is different. Here the two inputs are independent. Without the strong correlation between the inputs the positive local information terms are smaller, and the balance between positive and negative contributions to unique information is closer. Therefore, there is a narrow parameter region, when $s_1 < 2$ in which there is net positive unique information about X_2. In Figure 8, which shows all the surfaces for independent inputs, the surfaces for the modulatory transfer function do not plateau so much. They remain mostly constant along s_2 axis, and along the s_1 axis UnqX_1 increases while Shar_M and Synergy decrease (Shar_S is always zero here due to the fact the inputs are independent.)

8. Applications of ID Measures to Psychophysical Data

We now turn our attention to demonstrating the practicality of using PID and EID to decompose spectra from real-world data. We use the example of a behavioural lateral masking paradigm whereby the driving RF input is a centrally presented gabor patch (a sinusoidal grating combined with a gaussian function) of varying contrast. CF input takes the form of high-contrast gabor patches that flank the central target in the upper and lower visual fields; see Figure 9 for example stimuli. Neurophysiological studies have demonstrated that, in this experimental setup, when flankers are presented concurrently with targets but placed outside the classical receptive field, the cell's response to the target is modulated [32,33]. Furthermore, due to the size of stimuli, orientation, contrast, and their wavelength, CF input can suppress detection of the centrally presented target gabor [32,33]. This paradigm is a suitable testbed for PID measures since it measures the influence of a modulatory input (CF), surrounding flanker stimuli, on performance, in this instance a contrast detection task on a centrally presented gabor (RF). Furthermore, the paradigm can be manipulated to conform to the predictions outlined in Section 3.

(**a**) Over threshold, no flankers

(**b**) Over threshold, with flankers

(**c**) No target, with flankers

Figure 9. Examples of gabor patch stimuli used in the psychophysical experiment. In all conditions, the task was to detect the presence of a centrally presented target gabor.

We tested 21 participants from the University of Stirling's undergraduate psychology programme (Mean age = 19.1 years, SD = 1.3), who all had normal or corrected to normal vision. Ethical approval for the study was obtained from the University of Stirling's research ethics committee. Participants first completed a two-alternative forced choice staircase experiment, in which individual contrast sensitivity thresholds were established. Participants were asked to report whether a Gabor patch appeared to the left or right of a central fixation cross; the Gabor patch steadily decreased in contrast over the course of the experiment until a threshold of 60% accuracy was determined. This procedure was run twice with participants, and the average contrast threshold was used. After thresholds were established, participants completed the main experiment in which they were tasked with detecting a central target gabor in three conditions: (1) Over threshold target; (2) At threshold target; (3) No target present. In all three conditions, flankers were either present or not with equal occurrence; see Figure 9 for example stimuli.

Participants completed 100 trials per condition (except in the "No target" conditions, where they viewed 25 trials per condition, giving 450 trials in total), and all stimuli were presented for 500 ms, with a 2000 ms inter-stimulus interval for participants to respond.

Gabor patch stimuli for both the staircase and the main experimental paradigms were viewed on a gamma corrected CRT monitor (Tatung C7BBR, 60 Hz refresh rate, Taipei, Taiwan) at a distance of 80 cm, had a spatial frequency of 0.5 cycles per degree, and subtended a visual angle of no more than 1.93° in horizontal and vertical dimensions. From upper to lower flanker, the whole image subtended no more than 8.22° of vertical visual angle. All stimuli were presented on a medium grey background (RGB, 128,128,128). Gabors were phase shifted by $\pm 90°$ to present equal weightings of black/white. Flanker gabors in the main experiment were presented at 0.85 Michelson contrast across all trials, whereas central target gabor contrast varied by individual (Mean = 0.012, SD = 0.003).

Table 1. Estimated accuracy, with estimated standard error, for each combination of the three conditions and the absence or presence of flankers.

	No Target	**At Threshold**	**Over Threshold**
Without Flankers	0.9096 (0.0273)	0.8797 (0.0289)	0.9824 (0.0037)
With Flankers	0.9629 (0.0150)	0.3766 (0.0532)	0.9849 (0.0039)

Summary statistics for the accuracy data are shown in Table 1. Of particular note is the suppression of contrast detection accuracy in the "At Threshold" condition when flankers are present. We found, using a 3 (Threshold: Over, At, No target) by 2 (Flankers: With vs. Without) repeated measures ANOVA model (Huynh-Feldt corrections reported where appropriate), that accuracy for detection of the central gabor patch was lower in "at threshold" conditions in comparison to "over threshold" [F(1.147, 22.937)

= 66.401, $p < 0.001, \eta^2 = 0.769$]; post hoc comparison, Mean difference = 0.356, $p < 0.001$]. Furthermore, the presence of flankers further reduced the contrast detection accuracy [F(1, 20) = 55.508, $p < 0.001$, $\eta^2 = 0.735$], however this was a consequence of flanker stimuli suppressing contrast detection when target was at threshold, but not when the target was over threshold [F(1.334, 26.678) = 85.042, $p < 0.001, \eta^2 = 0.81$]. These results indicate that the CF input in these conditions served to suppress contrast detection; however the nature of the suppressive effect found could be additive/subtractive or modulatory.

Group ID spectra for the analysis of this experiment show that in conditions where the central target gabor was presented over threshold, i.e., in a case of near certainty, the majority of information transmitted in Y is unique to X_1, the driving RF input. The influence of CF flanker stimuli in this condition makes very little contribution to the output (Figure 10). In contrast, in conditions of uncertainty, i.e., at threshold, the unique contributions of X_1 driving RF input is, by definition, much reduced, and the effect of the X_2 modulatory CF input is much increased via its contribution to the synergistic component. This latter effect occurs even though the unique contribution of the CF input at threshold is small. The pattern of decompositions observed when the target driving RF input is weak is similar to that of the modulatory transfer function examined in Section 5, except for the occurrence of a small amount of unique information from the X_2 modulatory CF input.

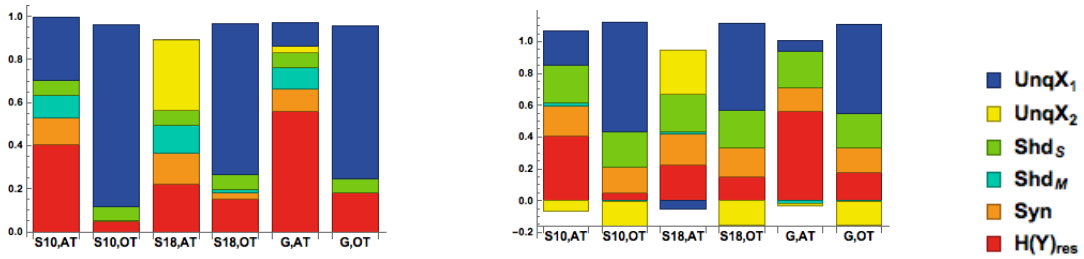
Figure 10 shows group decomposition spectra, however the decomposition may vary across subjects. Fortunately, enough data was collected for analysis of the individual data to be possible. We show Ibroja spectra for individual subjects of interest also in Figure 10. When the RF input is over threshold (i.e., strong), information transmitted is again unique to the RF in both subjects 10 and 18. However, at threshold (i.e., weak RF input) interactions that meet the criteria for modulation do occur for many subjects. Subject 10 is a clear example of a subject for whom the flanking context did indeed seem to function as a modulator. Information unique to the target stimulus was transmitted, but information unique to the flanking context, X_2, was at or near zero. X_2 must have contributed to output, however, because there is a substantial synergistic component. Such subjects therefore display a decomposition that is remarkably similar to that for the modulatory function studied in previous sections.

A few subjects performed very differently at threshold. Subject 18's responses at threshold conveyed no unique information about the target; unique information to CF input dominates, but again with substantial shared information and synergy between RF and CF. Therefore, the target, X_1, input contributed to the synergy, but the subject's response conveyed unique information only about X_2. Thus, under these conditions for these subjects, the central target, X_1, modulated transmission of information about the flankers, X_2, not the other way round. This demonstrates the value of using ID spectra to analyze such data.

Accuracy data for subject 18 suggests a very strong suppressive effect of CF input on contrast detection when the central target was presented at threshold (Accuracy in at threshold condition with flankers is 3%). The presence of some information unique to X_2 in the group data is therefore largely due to a few subjects whose performance at threshold was mainly transmitting information about the flankers. It may be that there were subjects for whom the threshold was underestimated. Overall, the decompositions of these psychophysical data confirm the rich expressive power of the decomposition spectra, and we expect to see far more use of them for such purposes in the near future.

To summarise, the nature of the modulation presented above is uncovered through use of decomposition measures. The suppression of contrast detection accuracy observed here when the RF input is weak coincides with less unique information transmitted about the RF in the output, and in addition, shared information and a synergistic relationship between RF and CF inputs. EID spectra suggest that the shared information is not mechanistic (see Section 4). Differing PID spectra between individual participants highlights the efficacy of PID for disambiguating modulatory interactions at the single subject level. The empirically observed spectra shown in this section may also cast some light on relations between PID and EID. Overall, these two forms of decomposition are mostly in

agreement. With respect to the negative EID components they again show that where negative unique components occur they seem to be compensated for by equivalent positive increases in the synergy. In addition, these results show that most EID components are positive, with negative components being the exception rather than the rule.



(**a**) PID spectra (**b**) EID spectra

Figure 10. Partial information decomposition (PID) and EID spectra (in bits) calculated for subject 10 (S10), subject 18 (S18) and the whole group of subjects (G) in the contrast detection experiment calculated at threshold (AT) and over threshold (OT).

9. Conclusions and Discussion

9.1. Implications of These Findings for Conceptions of 'Modulation' in the Cognitive and Neurosciences

Intuition suggests that any variable that affects output must transmit information specifically about itself in that output. That is clearly incorrect because output can be pure synergy. Furthermore, as shown in Figure 2e, conventional information theoretic analysis weakens that view by showing that the conditional mutual information transmitted about the modulator is at or near zero unless the primary drive is present but weak. The PID and EID analyses reported in Sections 5–7 now show that the conditional mutual information transmitted about the modulator was greater than zero when the primary input was present but weak because the synergistic component is then greater than zero, not because the modulator transmits unique information about itself. Thus, the intuitive view that to have any effect modulators must transmit specific information about themselves is shown to be seriously misleading.

Signals can have a kind of dual "semantics", one concerned with the message being transmitted, and one being concerned with the strength, salience, confidence, or precision with which that message is conveyed. The notion of contextual modulation requires a distinction between signal strength and signal semantics because it implies that the signal's strength can be modulated without changing its semantic content. A set of criteria to be met by what we call a modulatory transfer function were stated in Section 3. The surfaces given in Sections 6 and 7 for PID and EID analyses respectively show that our modulatory transfer functions meet these criteria. Section 5.2 showed a set of four ID spectra that together would imply that a transfer function is modulatory. ID spectra have substantial expressive power so it is possible that, when applied to empirical data from the cognitive and neurosciences, they may reveal that modulatory interactions take various and unexpected forms.

Another perspective from which to view our distinction between drive and modulation is that of the receiver of the output signal. Such a receiver can confidently infer the sign of the driving input from the output alone when the driving input is sufficiently strong. This is true whatever the strength of the modulatory input. Nothing can be confidently inferred from the output alone about the sign of the modulatory input, however, no matter what the strength of that modulatory input. This again supports our claim that modulatory inputs do not contribute to the message being conveyed by the semantics of output.

9.2. Comparisons between PID and EID

The most important outcome of the findings reported above is that they show that both EID and PID support all the main conclusions made above with respect to the defining properties and functions of modulatory interactions. Important strengths of EID shown here are that it distinguishes between source and mechanistic forms of shared information, and it relates them appropriately to the correlation between the two inputs. This is also the case with PID when the separation of shared information from [23] is included in the Ibroja decomposition.

9.3. Using EID and PID to Analyze and Interpret Psychophysical Data

The application of PID spectra to psychophysical data is useful in distinguishing ways in which two distinct inputs can contribute to a single measure of output. The methods outlined here can establish the underlying nature of statistical interactions in real world systems that cannot be studied with traditional multi-variate statistics alone. Future studies will apply these measures to continuous data streams to elucidate the strength of modulatory effects in complex neuroimaging data for example.

9.4. Using ID Spectra to Analyze and Interpret Empirical Data in General

The spectra and surfaces shown here were computed from a known transfer function, but the inverse problem may also arise. That is, to what extent can a transfer function, or properties of it, be inferred from an ID spectra, or set of spectra? For example, ID spectra could be computed from neurobiological observations, from psychophysical observations, from the activities of local processing elements in deep learning architectures, or from the input-output activity of a system as a whole. Work on information decomposition has so far focussed on the forward problem, i.e., on computing the spectra given a known transfer function. When ID spectra are computed from empirical data, however, then issues concerning the inverse problem will become more prominent and the application of formal statistical modelling will be required. Future studies will apply these measures to continuous data streams to elucidate the strength of modulatory effects in complex neuroimaging data for example. I_{ccs} and the EID can be easily computed for continuous Gaussian variables, which together with a semi-parametric Gaussian copula assumption results in a promising approach for robustly estimating these quantities from experimental data [34]. Further study of the statistical properties of these methods when applied to experimental data, for example in terms of limited sampling bias [35] and optimal permutation tests for valid statistical inference [36] are important areas for future work. For some recent work with fMRI data, see [37].

Empirical studies will rarely provide enough data to compute the equivalent of the surfaces shown above, so it is spectra that empirical studies will usually provide. The studies above show that the conditions under which the spectra are measured must be carefully chosen if modulatory and additive functions are to be distinguishable. We assume that transfer functions cannot be rigorously inferred from observed spectra, but they can be examined to see whether or not they meet the requirements for a modulatory interaction as described above. This will not fully constrain the unknown transfer function producing the observed output because those requirements can be met in many different ways. If an observed spectrum does meet our criteria for a modulatory interaction, then further experiments might be designed to distinguish between different ways in which those criteria can be met.

9.5. Modulatory Regulation of Activity as a Crucial and Non-Trivial Aspect of Information Processing

Though the topics dealt with in this Special Issue have implications for many disciplines they have special implications for the computational, cognitive, and neurosciences. This new perspective on multivariate information decomposition substantially enhances our notions of what "information processing" can be, and that is at the heart of all of those disciplines. Information processing is more than simply transferring information from one time or place to another. As others have argued it also includes creating new information via synergetic interactions between separate inputs; see [26,38]. Our argument

here is that in addition to "enhancing computational capabilities via synergy" information processing also includes distinguishing between currently relevant and currently irrelevant inputs. That is far from trivial, and though we have not considered the various criteria by which relevance can be assessed, we have done so elsewhere; see e.g., [19,39]. Here we have shown that it is possible to use any such assessment to amplify relevant and disamplify irrelevant signals without corrupting their semantic content. ID spectra can now be used as a way of exploring information processing within biological systems. It will be of particular interest to see whether interactions similar to those produced by our modulatory transfer function can be observed at the cellular level. We have shown that it is possible to use any such assessment to amplify relevant and disamplify irrelevant signals without corrupting their semantic content. Whether biology uses such modulatory interactions can now be explored by applying the ID spectra that we have proposed to biological data. The ID spectra could also be used to enhance our understanding of the information processing performed by local processors within various machine-learning architectures. It will also be possible to build new architectures designed to exploit the computational capabilities made possible by modulatory interactions such as those analysed here.

Acknowledgments: We thank Elena Gheorghiu for help with design of the Gabor patch stimuli, and Eva Kriechbaum & Aimee Lord for assistance with data collection and analysis. William A. Phillips is partially supported by a European Human Brain Project (EU grant 604102) to Lars Muckli. We also thank anonymous reviewers for helpful comments which have resulted in an improved version of the paper.

Author Contributions: William A. Phillips conceived the investigation, designed the computational studies, introduced the concept of an ID spectra, wrote Sections 1 and 9, and contributed to Sections 3, 5 and 8. Benjamin Dering wrote Section 8. Robin A. A. Ince produced all the EID outputs, wrote Sections 4.2 and 7 and contributed to Section 3. Jim W. Kay produced all the PID outputs, wrote Sections 2, 3.1, 4.1 and 6, wrote the appendix and contributed to Sections 3 and 5. All the authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Preliminary Results

Consider the logistic function L, from $\mathbb{R}$ to the open interval $(0,1)$, which is strictly increasing and has the following properties

$$\begin{aligned} &L(x) = 1/(1+\exp(-x)), \quad L(-x) = 1 - L(x), \quad 0 < L(x) < 1, \\ &L(x) > \tfrac{1}{2},\ L(x) = \tfrac{1}{2},\ L(x) < \tfrac{1}{2} \iff x > 0,\ x = 0,\ x < 0, \quad \text{respectively.} \end{aligned} \tag{A1}$$

From (14)–(16), we may use (A1) to write the values of u, v in the form

$$u_M = L[T_M(1,1)],\ v_M = L[T_M(1,-1)],\ u_A = L[T_A(1,1)],\ v_A = L[T_A(1,-1)]. \tag{A2}$$

Now, we write from (10) that

$$\begin{aligned} T_M(-1,-1) &= -\tfrac{1}{2}s_1(1+\exp(s_1 s_2)) = -T_M(1,1), \\ T_M(-1,1) &= -\tfrac{1}{2}s_1(1+\exp(-s_1 s_2)) = -T_M(1,-1) \end{aligned} \tag{A3}$$

and so using (A1) it follows that

$$\begin{aligned} \Pr(Y=1|X_1=-1,X_2=-1) &= L[T_M(-1,-1)] = L[-T(1,1)] = 1 - L[T_M(1,1)] = 1 - u_M \\ \Pr(Y=1|X_1=-1,X_2=1) &= L[T_M(-1,1)] = L[-T(1,-1)] = 1 - L[T_M(1,-1)] = 1 - v_M \end{aligned}$$

A similar argument using the additive transfer function, T_A, shows that

$$\Pr(Y=1|X_1=-1,X_2=-1) = 1 - u_A, \quad \text{and} \quad \Pr(Y=1|X_1=-1,X_2=1) = 1 - v_A.$$

Therefore, the conditional output probabilities are $\{1-u, 1-v, v, u\}$ when taken in the order $\{--, -+, +-, ++\}$, where (u, v) are replaced by (u_M, v_M) when using the transfer function T_M, and by (u_A, v_A) when using T_A. It follows from (1)–(4) that the joint p.m.f. $p(y, x_1, x_2)$ may be written as

$$\{\lambda u,\ \mu v,\ \mu(1-v),\ \lambda(1-u),\ \lambda(1-u),\ \mu(1-v),\ \mu v,\ \lambda u\}, \tag{A4}$$

where the probabilities are written in the order $\{p_{---}, p_{--+}, p_{-+-}, p_{-++}, p_{+--}, p_{+-+}, p_{++-}, p_{+++}\}$, so, for example, $\Pr(Y=-1, X_1=+1, X_2=-1) = p_{-+-}$. We find the marginal distribution of the output Y. From (A4), we have that

$$\Pr(Y=-1) = \lambda u + \mu v + \mu(1-v) + \lambda(1-u) = \lambda + \mu = \tfrac{1}{2},$$

and so Y, as well as X_1 and X_2 has a uniform binary distribution.

We now calculate the various Shannon entropy terms that will be required in the sequel. Since each of the three variables has a marginal uniform binary distribution, we can say that

$$H(Y) = 1, \quad H(X_1) = 1, \quad \text{and} \quad H(X_2) = 1. \tag{A5}$$

From (2) and (3), and noting that $\lambda + \mu = \frac{1}{2}$, we can write the Shannon entropy of the marginal (X_1, X_2) distribution as

$$H(X_1, X_2) = -2\lambda \log \lambda - 2\mu \log \mu = 1 - (2\lambda)\log(2\lambda) - (1-2\lambda)\log(1-2\lambda) = 1 + h(2\lambda),$$

where the function h is defined in (17). From (A4), we may write the marginal p.m.f.s of (Y, X_1) and (Y, X_2) in the order $\{--, -+, +-, ++\}$.

$$p(y, x_1):\ \{\lambda u + \mu v, \lambda(1-u) + \mu(1-v), \lambda(1-u) + \mu(1-v), \lambda u + \mu v\} = \{\tfrac{1}{2}z, \tfrac{1}{2}(1-z), \tfrac{1}{2}(1-z), \tfrac{1}{2}z\}, \tag{A6}$$

where, as in (17), $z = 2\lambda u + 2\mu v$.

$$p(y, x_2):\ \{\lambda u + \mu(1-v), \lambda(1-u) + \mu v, \lambda(1-u) + \mu v, \lambda u + \mu(1-v)\} = \{\tfrac{1}{2}w, \tfrac{1}{2}(1-w), \tfrac{1}{2}(1-w), \tfrac{1}{2}w\}, \tag{A7}$$

where, as in (17), $w = 2\lambda u + 2\mu(1-v)$.

We now calculate the Shannon entropies of the marginal (Y, X_1) and (Y, X_2) distributions.

$$H(Y, X_1) = -z \log(\tfrac{1}{2}z) - (1-z)\log(\tfrac{1}{2}(1-z)) = 1 + h(z). \tag{A8}$$

$$H(Y, X_2) = -w \log(\tfrac{1}{2}w) - (1-w)\log(\tfrac{1}{2}(1-w)) = 1 + h(w). \tag{A9}$$

Finally, from (A4), we find the Shannon entropy of the joint distribution of (Y, X_1, X_2).

$$\begin{aligned} H(Y, X_1, X_2) &= -2\lambda u \log(\lambda u) - 2\mu v \log(\mu v) - 2\lambda(1-u)\log[(\lambda(1-u)] - 2\mu(1-v)\log[\mu(1-v)], \\ &= -2\lambda \log \lambda - 2\mu \log \mu - 2\lambda[-u \log u - (1-u)\log(1-u)] \\ &\qquad - 2\mu[-v \log v - (1-v)\log(1-v)], \\ &= H(X_1, X_2) + 2\lambda h(u) + 2\mu h(v). \end{aligned} \tag{A10}$$

Appendix B. Proof of Theorem 1

(a) From (6), and using (A5), (A6), (A9) and (A10), we have that

$$\begin{aligned} I[Y;X_1|X_2] &= H(Y,X_2) + H(X_1,X_2) - H(X_2) - H(Y,X_1,X_2) \\ &= 1 + h(w) + H(X_1,X_2) - 1 - H(X_1,X_2) - 2\lambda h(u) - 2\mu h(v) \\ &= h(w) - 2\lambda h(u) - 2\mu h(v). \end{aligned}$$

(b) From (7), and using (A5), (A6), (A8) and (A10), we have that

$$\begin{aligned} I[Y;X_2|X_1] &= H(Y,X_1) + H(X_1,X_2) - H(X_1) - H(Y,X_1,X_2) \\ &= 1 + h(z) + H(X_1,X_2) - 1 - H(X_1,X_2) - 2\lambda h(u) - 2\mu h(v) \\ &= h(z) - 2\lambda h(u) - 2\mu h(v). \end{aligned}$$

(c) From (9), and using (A5) and (A8), we have

$$I[Y;X_1] = H(Y) + H(X_1) - H(Y,X_1) = 2 - 1 - h(z) = 1 - h(z).$$

(d) From (9), and using (A5) and (A9), we have

$$I[Y;X_2] = H(Y) + H(X_2) - H(Y,X_2) = 2 - 1 - h(w) = 1 - h(w).$$

(e) From (8) and parts (a) and (b), we have that

$$I[Y;X_1;X_2) = 1 - h(z) - (h(w) - 2\lambda h(u) - 2\mu h(v)) = 1 - h(z) - h(w) + 2\lambda h(u) + 2\mu h(v).$$

(f) From (5) and using (A5), (A6) and (A10), we have

$$I[Y;(X_1,X_2)] = H(Y) + H(X_1,X_2) - H(Y,X_1,X_2) = 1 - 2\lambda h(u) - 2\mu h(v).$$

Appendix C. Proof of Theorem 2

From Lemma 6 in [9] a necessary and sufficient condition for UnqX_2 to vanish is that there exists a row stochastic matrix $S = [\sigma(x_1;x_2)]$ such that

$$\Pr(Y = y, X_2 = x_2) = \sum_{x_1 \in B} \Pr(Y = y, X_1 = x_1)\sigma(x_1;x_2). \tag{A11}$$

We first find expressions for the joint p.m.f. in this more general case, but the work involved is very similar to that leading to (A4) above. From (29), (30) and (A1) we note that

$$g = L[F(s_1,s_2)],\ h = L[F(s_1,-s_2)] \text{ and } 0 < g,h < 1.$$

Also, since $\lambda \neq 0$ and $\mu \neq 0$ and $\lambda + \mu = \frac{1}{2}$, we have that $0 < \lambda < \frac{1}{2}$ and $0 < \mu < \frac{1}{2}$. From (28), (29) and (A1) we have that

$$\begin{aligned} \Pr(Y = 1|X_1 = -1, X_2 = 1) &= L[F(-s_1,s_2)] = L[-F(s_1,-s_2)] = 1 - L[F(s_1,-s_2)] = 1 - h \\ \Pr(Y = 1|X_1 = -1, X_2 = -1) &= L[F(-s_1,-s_2)] = L[-F(s_1,s_2)] = 1 - L[F(s_1,s_2)] = 1 - g \end{aligned}$$

It follows that the joint p.m.f. of (Y, X_1, X_2) is

$$\{\lambda g,\ \mu h,\ \mu(1-h),\ \lambda(1-g),\ \lambda(1-g),\ \mu(1-h),\ \mu h,\ \lambda g\}, \tag{A12}$$

and that the p.m.f.s for (Y, X_1) and (Y, X_2), in the order $\{--, -+, +-, ++\}$, are

$$p(y, x_1): \ \{\lambda g + \mu h, \lambda(1-g) + \mu(1-h), \lambda(1-g) + \mu(1-h), \lambda g + \mu v h\}, \tag{A13}$$
$$p(y, x_2): \ \{\lambda g + \mu(1-h), \lambda(1-g) + \mu h, \lambda(1-g) + \mu h, \lambda g + \mu(1-h)\}. \tag{A14}$$

Note that, since $\lambda + \mu = \frac{1}{2}$, we can write

$$\lambda(1-g) + \mu(1-h) = \tfrac{1}{2} - \lambda g - \mu h, \text{ and } \lambda(1-g) + \mu h = \tfrac{1}{2} - \lambda g - \mu(1-h).$$

From (A11), we now write out the system of equations that we will use to find a stochastic matrix.

$$\lambda g + \mu(1-h) = (\lambda g + \mu h)\sigma_{--} \ + \ (\tfrac{1}{2} - \lambda g - \mu h)\sigma_{+-} \tag{A15}$$
$$\tfrac{1}{2} - \lambda g - \mu(1-h) = (\lambda g + \mu h)\sigma_{-+} \ + \ (\tfrac{1}{2} - \lambda g - \mu h)\sigma_{++} \tag{A16}$$
$$\tfrac{1}{2} - \lambda g - \mu(1-h) = (\tfrac{1}{2} - \lambda g - \mu h)\sigma_{--} \ + \ (\lambda g + \mu h)\sigma_{+-} \tag{A17}$$
$$\lambda g + \mu(1-h) = (\tfrac{1}{2} - \lambda g - \mu h)\sigma_{-+} \ + \ (\lambda g + \mu h)\sigma_{++} \tag{A18}$$

Using (A15) and (A17), we first solve for σ_{--} and σ_{+-} and obtain

$$\begin{bmatrix} \lambda g + \mu(1-h) \\ \frac{1}{2} - \lambda g - \mu(1-h) \end{bmatrix} = \begin{bmatrix} \lambda g + \mu h & \frac{1}{2} - \lambda g - \mu h \\ \frac{1}{2} - \lambda g - \mu h & \lambda g + \mu h \end{bmatrix} \begin{bmatrix} \sigma_{--} \\ \sigma_{+-} \end{bmatrix}$$

Hence, inverting the matrix, we can write

$$\begin{bmatrix} \sigma_{--} \\ \sigma_{+-} \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} \lambda g + \mu h & \lambda g + \mu h - \frac{1}{2} \\ \lambda g + \mu h - \frac{1}{2} & \lambda g + \mu h \end{bmatrix} \begin{bmatrix} \lambda g + \mu(1-h) \\ \frac{1}{2} - \lambda g - \mu(1-h) \end{bmatrix},$$

where the determinant $\Delta = \lambda g + \mu h - \frac{1}{4}$. Now, $\Delta > 0$ provided that $g \geq \frac{1}{2}, h \geq \frac{1}{2}$ and g, h are not both equal to $\frac{1}{2}$. After some manipulation we obtain

$$\begin{bmatrix} \sigma_{--} \\ \sigma_{+-} \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} \lambda g + \frac{1}{2}\mu - \frac{1}{4} \\ \mu(h - \frac{1}{2}) \end{bmatrix}$$

and so when $g \geq \frac{1}{2}$ and $h \geq \frac{1}{2}$, but both are not equal to $\frac{1}{2}$ then σ_{--} and σ_{+-} are both non-negative and they sum to 1.

Very similar calculations for solving (A16) and (A18) give that

$$\begin{bmatrix} \sigma_{-+} \\ \sigma_{++} \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} \mu(h - \frac{1}{2}) \\ \lambda g + \frac{1}{2}\mu - \frac{1}{4} \end{bmatrix}$$

and the same reasoning as above shows that σ_{-+} and σ_{++} are both non-negative and they also sum to 1. Hence we have found a row stochastic matrix

$$S = \begin{bmatrix} \sigma_{--} & \sigma_{+-} \\ \sigma_{-+} & \sigma_{++} \end{bmatrix}$$

which satisfies (A11), and we conclude that UnqX_2 =0.

Appendix D. Proof of Corollary 1

It follows from (A3) that T_M satisfies the properties of F in (28). We now show that $u_M > \frac{1}{2}$ and that $v_M > \frac{1}{2}$. From (A2) and (11) we have that

$$u_M = L[\tfrac{1}{2}s_1(1 + \exp(s_1 s_2))] \text{ and } v_M = L[\tfrac{1}{2}s_1(1 + \exp(-s_1 s_2))].$$

Since $s_1 > 0, s_2 > 0$, we conclude from (A1) that u_M and v_M are both greater than $\frac{1}{2}$. Hence, from Theorem 2, $\mathrm{Unq}X_2 = 0$.

Appendix E. Proof of Corollary 2

Using (11) we know that

$$\begin{aligned} T_A(-1,-1) &= -s_1 - s_2 = -(s_1 + s_2) = -T_A(1,1), \\ T_A(-1,1) &= -s_1 + s_2 = -(s_1 - s_2) = -T_A(1,-1) \end{aligned}$$

and so T_A has the properties of F defined in (28). Also

$$u_A = L[s_1 + s_2] \text{ and } v_A = L[s_1 - s_2],$$

and so from (A1), and the assumption that $s_1 > 0, s_2 > 0$, we have that $u_A > \frac{1}{2}$ and also that $v_A \geq \frac{1}{2}$ if and only if $s_1 \geq s_2$. Hence, from Theorem 2 it follows that $\mathrm{Unq}X_2 = 0$.

Appendix F. Proof of Theorem 3

From Lemma 6 in [9], a necessary and sufficient condition for $\mathrm{Unq}X_1$ to vanish is that there exists a row stochastic matrix $T = [\tau(x_2; x_1)]$ such that

$$\Pr(Y = y, X_1 = x_1) = \sum_{x_2 \in B} \Pr(Y = y, X_2 = x_2)\tau(x_2; x_1). \tag{A19}$$

Since we are using T_A, here $g = u_A$ and $h = v_A$. From (A19), we now write out the system of equations that we will use to find a stochastic matrix.

$$\begin{aligned} \lambda g + \mu h &= (\lambda g + \mu(1-h))\tau_{--} + (\tfrac{1}{2} - \lambda g - \mu(1-h))\tau_{+-} \\ \tfrac{1}{2} - \lambda g - \mu h &= (\lambda g + \mu(1-h))\tau_{-+} + (\tfrac{1}{2} - \lambda g - \mu(1-h))\tau_{++} \\ \tfrac{1}{2} - \lambda g - \mu h &= (\tfrac{1}{2} - \lambda g - \mu(1-h))\tau_{--} + (\lambda g + \mu(1-h))\tau_{+-} \\ \lambda g + \mu h &= (\tfrac{1}{2} - \lambda g - \mu(1-h))\tau_{-+} + (\lambda g + \mu(1-h))\tau_{++} \end{aligned}$$

We note that the only difference in this system of equations, as compared with (A15)–(A18) is that h has been replaced by $1 - h$, and so one would expect that the result will hold when $u_A > \frac{1}{2}$ and $v_A \leq \frac{1}{2}$, and this turns out to be the case.

Following the same argument used in the proof of Theorem 2 it turns out that

$$T = \frac{1}{\lambda g + \mu(1-h) - \frac{1}{4}} \begin{bmatrix} \lambda g + \frac{1}{2}\mu - \frac{1}{4} & \mu(\frac{1}{2} - h) \\ \mu(\frac{1}{2} - h) & \lambda g + \frac{1}{2}\mu - \frac{1}{4} \end{bmatrix},$$

and we see that T is a row stochastic matrix provided that $g \geq \frac{1}{2}$ and $h \leq \frac{1}{2}$ and g and h cannot both be equal to $\frac{1}{2}$. From the proof of Corollary 2, we know that $u_A > \frac{1}{2}$ and from (A1) we know that $v_A \leq \frac{1}{2}$ if and only if $s_1 \leq s_2$.

For the last part, we know from (A1) that, when $s_1 = s_2$,

$$v_A = L[s_1 - s_2] = L[0] = \tfrac{1}{2}.$$

From (A14), with $g = u_A$ and $h = v_A = \frac{1}{2}$, we have that the marginal distributions of (Y, X_1) and (Y, X_2) are identical. Hence since both marginals have the same range space, B^2, it follows from [9] (Corollary 8) that $\mathrm{Unq}X_1 = 0$ and $\mathrm{Unq}X_2 = 0$. This completes the proof.

Appendix G. Proof of Theorem 4

(a) We saw in Theorem 2, Corollary 1, that $\mathrm{Unq}X_2 = 0$. It follows from (21) and Theorem 1(b) that

$$Syn = I[Y; X_2|X_1] = h(z_M) - 2\lambda h(u_M) - 2\mu h(v_M).$$

From Theorem 1(a) and (20) we have that

$$UnqX_1 = I(Y; X_1|X_2) - I(Y; X_2|X_1) = h(w_M) - h(z_M),$$

and from (19) we deduce that

$$Shar_{\mathrm{S+M}} = I(Y; X_2) = 1 - h(w_M).$$

(b) In this case, $v_A = \frac{1}{2}$, so $h(v_A) = 1$, and $z_A = w_A$. From Theorem 3, we know that $\mathrm{Unq}X_1 = 0$ and $\mathrm{Unq}X_2 = 0$. From (20), (21) we have

$$Syn = I[Y; X_1|X_2] = I[Y; X_2|X_1] = h(z_A) - 2\lambda h(u_A) - 2\mu.$$

and from (18) and (19) it follows that

$$Shar_{\mathrm{S+M}} = I[Y; X_1] = I[Y; X_2] = 1 - h(z_A).$$

(c) From Theorem 3, $\mathrm{Unq}X_1 = 0$, and using (18), (20) and (21) we obtain

$$Syn = I[Y; X_1|X_2] = h(w_A) - 2\lambda h(u_A) - 2\mu h(v_A),$$

$$UnqX_2 = I[Y; X_2|X_1] - I[Y; X_1|X_2] = h(z_A) - h(w_A),$$

and

$$Shar_{\mathrm{S+M}} = I[Y; X_1] = 1 - h(z_A).$$

(d) From Theorem 2, Corollary 2, $\mathrm{Unq}X_2 = 0$. Using the same deductions as in part (a), we find that

$$\begin{aligned} Syn &= I[Y; X_2|X_1] = h(z_A) - 2\lambda h(u_A) - 2\mu h(v_A), \\ UnqX_1 &= I[Y; X_1|X_2] - I[Y; X_2|X_1] = h(w_A) - h(z_A), \\ Shar_{\mathrm{S+M}} &= I[Y; X_2] = 1 - h(w_A). \end{aligned}$$

Appendix H. Proof of Theorem 5

For part (a), the correlation between inputs is $+1$, and so we know from (2), (3) and (17) that

$$\lambda = \tfrac{1}{2}, \mu = 0, z = v, w = v.$$

Hence, from Theorem 1(a, b)

$$I[Y : X_1|X_2] = h(v) - h(v) = 0, \quad \text{and} \quad I[Y; X_2|X_1] = h(u) - h(u) = 0.$$

From (20) and (21) it follows that $\mathrm{Unq}X_1 = \mathrm{Unq}X_2 = \mathrm{Syn} = 0$. Then from Theorem 1 and (18) it follows that $\mathrm{Shar}_{\mathrm{S+M}} = I[Y; X_1] = 1 - h(u)$.

In (b), the correlation between inputs is -1, and so we know from (2), (3) and (17) that

$$\lambda = 0, \mu = \tfrac{1}{2}, z = v, w = 1 - v.$$

Hence, from Theorem 1(a,b), and noting that $h(1-v) = h(v)$,

$$I[Y : X_1|X_2] = h(1-v) - h(v) = 0, \quad \text{and} \quad I[Y; X_2|X_1] = h(v) - h(v) = 0.$$

From (20) and (21) it follows that UnqX_1 = UnqX_2 = Syn =0. Then from Theorem 1 and (18) it follows that Shar$_{\text{S+M}}$ = $I[Y; X_1] = 1 - h(v)$.

References

1. Gilbert, C.D.; Sigman, M. Brain States: Top-Down Influences in Sensory Processing. *Neuron* **2007**, *54*, 677–696.
2. Phillips, W.A.; Singer, W. In search of common foundations for cortical computation. *Behav. Brain Sci.* **1997**, *20*, 657–722.
3. Phillips, W.A.; Silverstein, S.M. Convergence of biological and psychological perspectives on cognitive coordination in schizophrenia. *Behav. Brain Sci.* **2003**, *26*, 65–138.
4. Lamme, V.A.F. Beyond the classical receptive field: Contextual modulation of V1 responses. In *The Visual Neurosciences*; Werner, J.S., Chalupa, L.M., Eds.; MIT Press: Cambridge, MA, USA, 2004; pp. 720–732.
5. Kay, J.; Floreano, D.; Phillips, W.A. Contextually guided unsupervised learning using local multivariate binary processors. *Neural Netw.* **1998**, *11*, 117–140.
6. Larkum, M. A cellular mechanism for cortical associations: An organizing principle for the cerebral cortex. *Trends Neurosci.* **2013**, *36*, 141–151.
7. Phillips, W.A.; Larkum, M.E.; Harley, C.W.; Silverstein, S.M. The effects of arousal on apical amplification and conscious state. *Neurosci. Conscious.* **2016**, 1–13, doi:10.1093/nc/niw015.
8. Williams, P.L.; Beer, R.D. Nonnegative Decomposition of Multivariate Information. *arXiv* **2010**, arXiv:1004.2515.
9. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying Unique Information. *Entropy* **2014**, *16*, 2161–2183.
10. Griffith, V.; Koch, C.; Griffith, V. Quantifying synergistic mutual information. In *Guided Self-Organization: Inception. Emergence, Complexity and Computation*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 9, pp. 159–190.
11. James, R.G.; Emenheiser, J.; Crutchfield, J.P. Unique Information via Dependency Constraints. *arXiv* **2017**, arXiv:1709.06653.
12. Ince, R.A.A. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy* **2017**, *19*, 318.
13. Ince, R.A.A. The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv* **2017**, arXiv:1702.01591.
14. Phillips, W.A.; Kay, J.; Smyth, D. The discovery of structure by multi-stream networks of local processors with contextual guidance. *Netw. Comput. Neural Syst.* **1995**, *6*, 225–246.
15. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley-Interscience: New York, NY, USA, 1991.
16. Schneidman, E.; Bialek, W.; Berry, M.J. Synergy, Redundancy, and Population Codes. *J. Neurosci.* **2003**, *23*, 11539–11553.
17. Kay, J. Neural networks for unsupervised learning based on information theory. In *Statistics and Neural Networks: Advances at the Interface*; Kay, J.W., Titterington, D.M., Eds.; Oxford University Press: Oxford, UK, 1999; pp. 25–63.
18. Kay, J.; Phillips, W.A. Activation functions, computational goals and learning rules for local processors with contextual guidance. *Neural Comput.* **1997**, *9*, 895–910.
19. Kay, J.W.; Phillips, W.A. Coherent infomax as a computational goal for neural systems. *Bull. Math. Biol.* **2011**, *73*, 344–372.
20. James, R.G.; Crutchfield, J.P. Multivariate Dependence beyond Shannon Information. *Entropy* **2017**, *19*, 530.
21. Wibral, M.; Priesemann, V.; Kay, J.W.; Lizier, J.T.; Phillips, W.A. Partial information decomposition as a unified approach to the specification of neural goal functions. *Brain Cognit.* **2017**, *112*, 25–38.
22. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Phys. Rev. E* **2013**, *87*, doi:10.1103/PhysRevE.87.012130.
23. Pica, G.; Piasini, E.; Chicharro, D.; Panzeri, S. Invariant components of synergy, redundancy, and unique information. *Entropy* **2017**, *19*, 451, doi:10.3390/e19090451.

24. Wibral, M.; Lizier, J.T.; Vögler, S.; Priesemann, V.; Galuske, R. Local active information storage as a tool to understand distributed neural information processing. *Front. Neuroinf.* **2014**, *8*, doi:10.3389/fninf.2014.00001.
25. Lizier, J.T.; Prokopenko, M.; Zomaya, A. Local information transfer as a spatiotemporal filter for complex systems. *Phys. Rev. E* **2008**, *77*, doi:10.1103/PhysRevE.77.026110.
26. Wibral, M.; Lizier, J.T.; Priesemann, V. Bits from brains for biologically inspired computing. *Front. Robot. AI* **2015**, doi:10.3389/frobt.2015.00005.
27. Van de Cruys, T. Two Multivariate Generalizations of Pointwise Mutual Information. In Proceedings of the Workshop on Distributional Semantics and Compositionality, Portland, Oregon, 24 June 2011; pp. 16–20.
28. Church, K.W.; Hanks, P. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.* **1990**, *16*, 22–29.
29. James, R.G.; Ellison, C.J.; Crutchfield, J.P. Anatomy of a bit: Information in a time series observation. *Chaos* **2011**, 037109, doi:10.1063/1.3637494
30. Olbrich, E.; Bertschinger, N.; Rauh, J. Information decomposition and synergy. *Entropy* **2015**, *17*, 3501–3517.
31. Barrett, A.B. An exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Phys. Rev. E* **2015**, *91*, doi.org/10.1103/PhysRevE.91.052802
32. Chen, C.C.; Kasamatsu, T.; Polat, U.; Norcia, A.M. Contrast response characteristics of long-range lateral interactions in cat striate cortex. *Neuroreport* **2001**, *12*, 655–661.
33. Polat, U.; Mizobe, K.; Pettet, M.W.; Kasamatsu, T.; Norcia, A.M. Collinear stimuli regulate visual responses depending on cell's contrast threshold. *Nature* **1998**, *391*, 580–584.
34. Ince, R.A.A.; Giordano, B.L.; Kayser, C.; Rousselet, G.A.; Gross, J.; Schyns, P.G. A Statistical Framework for Neuroimaging Data Analysis Based on Mutual Information Estimated via a Gaussian Copula. *Hum. Brain Mapp.* **2017**, *38*, 1541–1573.
35. Panzeri, S.; Senatore, R.; Montemurro, M.A.; Petersen, R.S. Correcting for the Sampling Bias Problem in Spike Train Information Measures. *J. Neurophys.* **2007**, *98*, 1064–1072.
36. Ince, R.A.A.; Mazzoni, A.; Bartels, A.; Logothetis, N.K.; Panzeri, S. A Novel Test to Determine the Significance of Neural Selectivity to Single and Multiple Potentially Correlated Stimulus Features. *J. Neurosci. Methods* **2012**, *210*, 49–65.
37. Stramaglia, S.; Angelini, L.; Wu, G.; Cortes, J.; Faes, L.; Marinazzo, D. Synergistic and redundant information flow detected by unnormalized Granger causality: Application to resting state fMRI. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 2518–2524.
38. Timme, N.M.; Ito, S.; Myroshnychenko, M.; Nigam, S.; Shimono, M.; Yeh, F.-C. High-Degree Neurons Feed Cortical Computations. *PLoS Comput. Biol.* **2016**, *12*, e1004858, doi:10.1371/journal. pcbi.1004858.
39. Phillips, W.A.; Clark, A.; Silverstein, S.M. On the functions, mechanisms, and malfunctions of intracortical contextual modulation. *Neurosci. Biobehav. Rev.* **2015**, *52*, 1–20.