Information-theoretic sensitivity analysis: a general method for credit assignment in complex networks

Niklas Lüdtke1,2,*, Stefano Panzeri3, Martin Brown4, David S. Broomhead5, Joshua Knowles1,2,6, Marcelo A. Montemurro3 and Douglas B. Kell1,2

1School of Chemistry, 2The Manchester Interdisciplinary Biocentre, 3Faculty of Life Sciences, 4School of Electrical and Electronic Engineering, 5School of Mathematics, and 6School of Computer Science, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

Most systems can be represented as networks that couple a series of nodes to each other via one or more edges, with typically unknown equations governing their quantitative behaviour. A major question then pertains to the importance of each of the elements that act as system inputs in determining the output(s). We show that any such system can be treated as a ‘communication channel’ for which the associations between inputs and outputs can be quantified via a decomposition of their mutual information into different components characterizing the main effect of individual inputs and their interactions. Unlike variance-based approaches, our novel methodology can easily accommodate correlated inputs.

Keywords: sensitivity analysis; information theory; entropy; mutual information; communication channel; NFκB

1. INTRODUCTION

The analysis of networks represents a crucial focus of modern systems biology (Barabási & Oltvai 2004; Hwang et al. 2005; Kitano et al. 2005; Klipp et al. 2005; Wagner 2005; Alon 2006; Davidson 2006; Doyle & Stelling 2006; Kell 2006). For many areas of interest, models of complex networks can be taken to have the form of a deterministic mapping from a set of inputs to one or more output(s) (figure 1). The outputs can be considered separately so that for each output \( Y_k \) there is a map

\[ f_k : (X_1, \ldots, X_n) \mapsto Y_k \]

Usually, the input–output mapping is not available in explicit form but can be evaluated numerically for any given inputs.

Global sensitivity analysis aims to rank the inputs \( X_1, \ldots, X_n \) according to the degree to which they influence the output, individually and conjointly. Here, ‘inputs’ may also refer to intrinsic model parameters whose influence on the output is to be determined as in figure 1b. This type of global sensitivity analysis is commonly performed in a probabilistic manner by evaluating the model for multiple sets of randomly and independently selected input values drawn, for instance, from uniform distributions over suitable intervals. The output, being a function of the randomized inputs, thus also becomes a random variable. If the inputs are sampled independently, the variance of the output distribution can be decomposed into contributions by individual inputs, pairs, triplets and so forth. This procedure is well known in statistics as ‘analysis of variance’ (ANOVA; e.g. Box et al. 1978), and several authors have contributed to improve its computational efficiency for sensitivity analysis (e.g. Rabitz & Alis 1999; Sobol 2001).

Rather than analysing the variance of the output distribution, we take a different route measuring output uncertainty in terms of Shannon’s entropy (Shannon & Weaver 1949). Our starting point is the concept of the ‘communication channel’ (Cover & Thomas 2006), which enables us to view the model as a transmitter of information between inputs and outputs (figure 1b).

The mutual information of two variables is a quantity that measures their mutual dependence (Cover & Thomas 2006). Determining the mutual information \( I(X;Y) \) between random sampling sequences of individual inputs \( X_j \) and their output counterpart can elucidate first-order input–output relations. Mutual information provides a general measure of association that is applicable regardless of the shape of the underlying distributions and—unlike linear- or rank-order correlation—insensitive to non-monotonic dependence among the random variables. Further insight can be obtained by unravelling conditional dependencies among the system inputs. Here, we define novel and general sensitivity measures of second and higher order by evaluating input correlations induced by conditioning on the output.

*Author for correspondence (niklas.ludtke@manchester.ac.uk).
methodology, independent inputs merely simplify the analysis. If input correlations exist (e.g. due to non-orthogonal sampling), their effect can easily be taken into account. We apply the methodology successfully to a model of the NFκB signalling pathway and thereby define how to modify its behaviour to provide a designed maximum effect.

2. METHODS

By randomly sampling the input space, a genuinely deterministic system can be analysed in stochastic terms. Random perturbation of the inputs creates a randomized output \( Y \) with a probability density \( p(y) \). Rather than attempting to find some parametric model of \( p(y) \), the output density is approximated by a histogram, and the output becomes a \textit{discrete} random variable. The corresponding entropy

\[
H(Y) = -\sum_y p(y) \log_2 p(y)
\]

measures a (hypothetical) receiver’s uncertainty about \( Y \) due to input perturbation. For instance, if one input \( X_i \) is fixed, the receiver’s remaining uncertainty can be quantified by the \textit{conditional} entropy

\[
H(Y|X_i) = \sum_x p(x) H(Y|X_i = x),
\]

which is the average uncertainty in \( Y \) over all possible discrete values \( x \) that the input variable \( X_i \) can assume. The discretization of \( X_i \) and \( Y \) is, of course, arbitrary and should be chosen in relation to the number of system evaluations (simulation runs).

The mutual information is defined as the difference in output uncertainty with and without knowledge of \( X_i \), and characterizes the influence \( X_i \) exerts on \( Y \):

\[
I(X_i; Y) = H(Y) - H(Y|X_i).
\]

The link between uncertainty and association established by equation (2.3) is one of the fundamental concepts of Shannon’s information theory and forms the basis of our framework for sensitivity analysis. Calculating the mutual information \( I(X_i; Y) \) for each \( X_i \) constitutes a form of first-order sensitivity analysis, assessing only the influence of individual inputs.

2.1. An information-theoretic first-order sensitivity index

Critchfield et al. (1986) defined the \textit{mutual information index} (MII), which in our notation is the mutual information normalized by the entropy of the output variable:

\[
s_i = \frac{I(X_i; Y)}{H(Y)}.\]

A first-order sensitivity analysis can be performed by calculating the MII of all inputs, where the mutual information is obtained by computing

\[
I(X_i; Y) = \sum_{x_i} \sum_y p(x_i, y) \log_2 \frac{p(x_i, y)}{p(x_i)p(y)}.
\]
Though $X_i$ and $Y$ are continuous variables, equation (2.5) contains discrete sums, indicating that, in practice, the probability densities are evaluated via the joint histogram and the marginal histograms of the input and output sequences.

### 2.2. Pairwise interactions

If we assume that, by design of the simulation, random input values are drawn independently, there will be no \textit{a priori} correlations among the sequences of input values. However, if inputs interact in their influence on an output, one would expect to find associations in input sequences when conditioning on a particular value of that output. We show that the output-induced conditional dependence among two inputs, characterized by the \textit{conditional mutual information}

$$I(X_i, X_j | Y) = \sum_y p(y) \sum_{x_i, x_j} p(x_i, x_j | y) \log_2 \frac{p(x_i, x_j | y)}{p(x_i | y)p(x_j | y)},$$

(2.6)

provides a measure of the joint influence of the pair $(X_i, X_j)$ on the output $Y$, on average.

To understand why this is indeed an appropriate measure, we consider the degree of association among $(X_i, X_j)$ and $Y$, that is the mutual information $I(X_i, X_j; Y)$. Since this quantity subsumes first- and second-order effects, one has to subtract the influence of the individual inputs, $I(X_i; Y)$ and $I(X_j; Y)$, in order to obtain the pure second-order effect of $X_i$ and $X_j$ on $Y$. Using an auxiliary formula proved in appendix A.3, one obtains

$$I(X_i, X_j; Y) = I(X_i; Y) + I(X_j; Y) - I(X_i, X_j; Y) = I(X_i; X_j | Y) - I(X_i; X_j).$$

(2.7)

The second term on the right-hand side subtracts the effect of any \textit{a priori} input associations due to the applied sampling scheme. If inputs are sampled independently, the term vanishes and the conditional mutual information by itself captures the joint effect of $X_i$ and $X_j$ on $Y$. Note, the simple structure of equation (2.7) makes it possible to apply arbitrary input sampling schemes, without having to be concerned about statistical independence. This reveals a considerable advantage of the information-theoretic approach over variance-based methods, which are not easily extended to non-orthogonal samples (Saltelli et al. 2000).

### 2.3. Higher order interactions

Capturing interactions among three or more inputs in information-theoretic terms requires generalizing the concept of mutual information beyond two variables. To characterize the genuine three-way interaction of input triplets, we apply the same rationale as in §2.2 and consider a decomposition of the mutual information of an input triplet $(X_1, X_2, X_3)$ and the output

(derivation provided in appendix A.3)

$$I(X_1; X_2, X_3; Y) = I(X_1; Y) + I(X_2; Y) + I(X_3; Y) - I(X_1; X_2 | Y) - I(X_1; X_3 | Y) - I(X_2; X_3 | Y) + I(X_1; X_2; X_3 | Y).$$

(2.8)

Having identified the first- and second-order terms on the right-hand side, the decomposition suggests interpreting the remainder as the genuine third-order sensitivity measure. Using the notation $I_3$, we define

$$I_3(X_1; X_2; X_3 | Y) = \frac{I(X_2; X_3 | X_1, Y) - I(X_2; X_3)}{I(X_2; X_3 | Y)}. $$

(2.9)

This quantity is the conditional form of what McGill (1954) called ‘interaction information’. McGill also showed that interaction information is symmetric with respect to permutations of its arguments, meaning for the conditional form that

$$I_3(X_1; X_2; X_3 | Y) = I(X_2; X_3 | X_1, Y) - I(X_2; X_3).$$

(2.10)

Note that interaction information can be negative. By virtue of equation (2.10), this can only happen when all three pairwise interactions have non-zero conditional mutual information. Negative interaction information then indicates an inner redundancy of the triplet $(X_1, X_2, X_3)$, in the sense that the pairs $(X_1, X_2)$, $(X_2, X_3)$ and $(X_1, X_3)$ do not provide entirely independent pieces of information about $Y$. This situation rarely occurs in natural systems, although appendix A.1 presents a contrived and artificial example with negative interaction information.

Generalizing the definition of $I_3$, one can analogously quantify the fourth-order sensitivity via the fourth-order conditional interaction information

$$I_4(X_1; X_2; X_3; X_4 | Y) = I(X_1; X_2; X_3, X_4 | Y) - I(X_1; X_2 | X_3, X_4),$$

(2.11)

Higher order interactions can be defined accordingly.

### 2.4. The information balance: a summation theorem for sensitivity indices

Having identified measures for first-, second- and higher order sensitivities, we consider a decomposition of the total mutual information for arbitrary number of inputs. Generalizing equation (2.3), one obtains the general form of the information balance for a system.
with $n$ inputs

$$I(X_1, \ldots, X_n; Y) = H(Y) - H(Y|X_1, \ldots, X_n).$$  \hfill (2.12)

In addition to $H(Y)$, which is straightforward to compute, the ‘noise entropy’ $H(Y|X_1, \ldots, X_n)$ has to be evaluated. In a deterministic system, this quantity vanishes for continuous random variables. However, computing information-theoretic quantities in a continuous fashion would require parametric models of all random variables. Since the input–output mapping is often not given in closed form, it will generally be impossible to derive such models analytically. Moreover, there is no general parameterization scheme that can be fitted to the multitude of possible empirical distributions arising in various systems. Rather, the parametric distributions must be selected on a case-by-case basis, with no guarantee of obtaining a close fit. In our experience, it proved very difficult to match the heavy-tailed output histograms arising in our particular application (cf. §3) with any standard distribution function.

Hence, to make the information-theoretic quantities measurable, we choose to discretize all variables. The noise entropy then takes the role of the residual uncertainty in $Y$ that persists, given all inputs with a finite precision determined by the imposed discretization. We shall refer to this residual uncertainty as the discretization entropy, denoted by $H_\Delta$. In §3.4, we show how $H_\Delta$ can be estimated via Monte Carlo simulation.

Normalizing by $H(Y) - H_\Delta$—the maximum amount by which the output uncertainty can be reduced by the parameters—one can rewrite the information balance for a discretized deterministic system as

$$\frac{I(X_1, \ldots, X_n; Y)}{H(Y) - H_\Delta} = 1.$$  \hfill (2.13)

Equation (2.13) provides the basis for a summation theorem, since it is possible to express the left-hand side in terms of the previously defined sensitivity indices, as shown in appendix A.3. Decomposition up to third order yields

$$\frac{1}{H(Y) - H_\Delta} \left\{ \sum_i I(X_i; Y) + \sum_{i<j} I(X_i; X_j|Y) + \sum_{i<j<k} I_3(X_i; X_j; X_k|Y) + \Delta I \right\} = 1.$$  \hfill (2.14)

Here, summations extend over all index combinations excluding permutations. While related decompositions of the information of an ensemble of variables have been considered previously (Watanabe 1960; Fano 1961; Panzeri et al. 1999; Amari 2001; Schneidman et al. 2006), they have never been applied in the context of sensitivity analysis (see appendix A.1 for a discussion of alternative decompositions).

Calculation of the conditional mutual information requires knowledge of the underlying marginal and joint probability density functions. In practice, these densities must be estimated empirically by means of marginal and joint histograms. Particularly, the empirical estimate of a joint density can be problematic when the amount of available data is insufficient to populate all the bins in its joint histogram. This leads to a systematic error (‘bias’) in the limited-sampling estimation of information; the higher the dimensionality of the histograms to be sampled, the larger the bias. Thus, the estimation of higher order interactions is particularly difficult. However, reliable correction of the sampling bias is possible using advanced statistical techniques (Panzeri & Treves 1996; Nemenman et al. 2004; Montemurro et al. in press). Given the amount of simulations we could produce in the particular application presented below, these techniques allowed an accurate elimination of the bias for up to third-order interactions. Only the first-order quantification would have been possible without using such bias reduction techniques.

Even though, for the practical reasons described above, sensitivity indices can only be evaluated up to a certain order, the remainder $\Delta I$—the combined effect of all higher order interactions—can be assessed since all other terms in the equation are known. If the lower order sensitivity indices capture the essence of the dependence structure, the remainder will be a small fraction of $H(Y) - H_\Delta$. A significant value of $\Delta I$ would indicate that important higher order interactions exist, which is generally not expected in most simple systems (Rabitz & Aliş 1999). In large networks, higher order interactions require an extreme number of connections, unless the degree of connectivity varies strongly across the network. Hence, one would expect to find a small number of local ‘hubs’ forming highly connected subnetworks. While this is still the subject of debate, we note that the complex networks arising in biological systems do indeed tend to have sparse intrinsic connectivity patterns (Wagner & Fell 2001; Barabási & Oltvai 2004; Csete & Doyle 2004).

### 2.5. Total sensitivity indices

A very useful concept in variance-based sensitivity analysis is the so-called total sensitivity index (Saltelli et al. 2005), which measures the overall influence that a particular input exerts on the output, comprising main effects and all interactions. In the ANOVA framework, the total sensitivity expresses the remaining output variance when all other inputs are kept fixed. The idea is to calculate this quantity without relying on the other sensitivity indices (first, second, third-order and so forth). If a total sensitivity index is zero, the corresponding input is irrelevant; if not, it is interesting to relate it to the other indices. For instance, comparing the total sensitivity index of an input with its first-order index reveals the degree to which the input is interacting with others.

This concept can be readily applied to information-based sensitivity analysis. The information-theoretic
total sensitivity index for variable $X_i$ is given by

$$s_{\text{total},i} = \frac{H(Y|\{X_1, \ldots, X_n\}) - H_{\Delta}}{H(Y) - H_{\Delta}}. \quad (2.15)$$

The total sensitivity index can also be expressed as the sum of all sensitivity indices involving $X_i$:

$$s_{\text{total},i} = \frac{1}{H(Y) - H_{\Delta}} \left\{ I(X_i; Y) + \sum_{j \neq i} I(X_i; X_j | Y) + \sum_{j \neq i, j \neq k} I_i(X_i; X_j; X_k | Y) + \ldots \right\}. \quad (2.16)$$

Note, the sum of all total sensitivity indices is generally greater than 1 since expansions for different input variables will share certain sensitivity indices if the variables interact.

### 3. INFORMATION-THEORETIC SENSITIVITY ANALYSIS OF A MODEL OF THE NFkB SIGNALLING PATHWAY

As an example, we apply our methodology to parameter sensitivity analysis in systems biology. We consider a model of the (IκB)/NFkB signalling pathway (Hoffmann et al. 2002) and investigate the interdependencies among intrinsic parameters (in this case 64 reaction rate constants) with respect to their influence on the time course of the concentration of a particular metabolite, the nuclear transcription factor NFkB, which is a key component in early immune response.

In a nutshell, the pathway model works as follows. There are three main components: NFkB, IκB inhibitory proteins (IκBz and its isoforms IκBβ and IκBα) and the IκB kinase (IKK). The model describes the kinetics of interaction between these components, their transport between nucleus and cytoplasm, the inhibitor IκB degradation, as well as the NFkB-regulated gene expression and subsequent resynthesis of the inhibitors (figure 2). NFkB is normally bound in an IκB–NFkB complex. Following a step increase in the concentration of IKK, which models the effect of an extracellular stimulus (e.g. tumour necrosis factor (TNFα)), NFkB is released from the IκB–NFkB complex and enters the nucleus. The IκBs are rapidly degraded. In the nucleus, NFkB regulates the expression of genes leading to a resynthesis of the IκB inhibitor proteins. The newly synthesized IκB binds to the nuclear NFkB forming an IκB–NFkB complex and subsequently shuttles NFkB back to the cytoplasm, thus initiating a negative feedback loop. The cycle is repeated until all IKK has decayed. As a result of the delayed negative feedback, the concentration of nuclear NFkB exhibits an oscillatory behaviour (figure 3) that can be characterized in terms of features such as peak amplitude, frequency or phase (Nelson et al. 2004; Kell 2006).

To exemplify the sensitivity analysis, we select one feature as output $Y$, namely the time difference between the first two peaks of the nuclear NFkB oscillation, $P_1 = T_2 - T_1$. Evaluating the input–output function thus involves numerically solving a system of 24 ordinary nonlinear differential equations (ODEs) corresponding to the reaction equations and subsequently determining the first two maxima in one component of the solution. The entire analysis is performed with respect to the selected feature, based on 680 000 simulations, which yield sufficiently accurate information measures up to third order. All parameters were varied simultaneously, each drawn independently from a uniform distribution over the interval from 0.9 to 2.0 times the nominal value, and discretized into 15 bins. The lower bound of the sampling interval is dictated by the empirical observation that oscillations are only guaranteed to occur for parameter values above about 0.9 of the nominal values given in Hoffmann et al. (2002).

#### 3.1. First-order sensitivity indices

The first-order analysis reveals that only a small subset of about eight parameters out of 64 (or at most 11, if one includes parameters 1, 19 and 37) significantly influence the output feature $P_1$ (figure 4). The parameters thus identified either directly affect the amount of available NFkB (9, cytoplasmic release; 19, nuclear import; 1, IκBα–NFkB association) or control the strength of the
negative feedback, which depends on the production of the inhibitor protein IκBα (28, transcription; 36, constitutive translation), its transport into nucleus (38, nuclear import of IκBα), its destruction (37, degradation; 29, IκBα mRNA degradation; 62, IKK–IκBα catalysis). In addition, the feedback is indirectly affected by the IκB kinase (IKK). Therefore, its decay rate (61) and the rate of binding IKK to the IκBα–NFκB complex (52), a step prior to the release of NFκB, are also important. Parameters specifically relating to the inhibitor isoforms IκBβ or IκBε are insignificant, which is in accordance with experiments indicating that only the knockout of IκBα is lethal (Gerondakis et al. 1999).

The result is consistent with previous local sensitivity analyses of a different output feature (Ihekwaba et al. 2004, 2005) and in accordance with global sensitivity analysis of the overall time course of the nuclear NFκB concentration (Yue et al. 2006).

3.2. Second-order sensitivity indices

At the level of pairwise interactions, a rather small number of relevant pairs emerge (figure 5). No significant synergies are observed, meaning that only those pairs wherein at least one partner is individually relevant have significant interactions. The predominant contributions are by pairs where both partners have significant individual impact. The degree of interactivity, that is the number of relevant pairs in which a parameter appears, varies strongly. For instance, parameter 29 (the IκBα mRNA degradation rate) seems to play the role of a ‘super parameter’, in the sense that it is involved in most, and the strongest, interactions. Note that pairs with very small yet statistically significant sensitivities are not visualized in the interaction matrix (figure 5), due to limited diagram resolution.

Statistical significance can be assessed via a bootstrap test, which involves repeated random shuffling of the parameter sampling sequences and recalculation of the conditional mutual information from these shuffled sequences. Several hundred repetitions produce a bell-shaped distribution of random sensitivity values, the mean and standard deviation of which characterize the range of ‘chance values’ of the particular sensitivity index under consideration. If the index calculated from the original non-shuffled data is two or three standard deviations above its bootstrap mean, it can be considered statistically significant.

While the particular parameter set identified as most relevant is very biologically plausible, the strongly varying degree of parameter interactivity is surprising. For instance, it is not obvious why the...
3.3. Third-order interactions

The sparseness of the interaction structure continues at the third order (figure 6). It is again the combinations of individually relevant parameters that exhibit the strongest tripletwise interactions. The assessment of statistical significance is analogous to the procedure described in §3.2. All third-order indices are positive.

3.4. Monte Carlo estimation of discretization entropy

Let \( n \) be the number of system inputs. Assume a discretization scheme where each input range is partitioned into the same number of bins, denoted by \( n_{\text{bins}} \). Let \( j_1, \ldots, j_n \) be the bin indices of the inputs \( X_1, \ldots, X_n \) with \( j_1 = 1, \ldots, n_{\text{bins}} \), \( j_2 = 1, \ldots, n_{\text{bins}} \) and \( j_n = 1, \ldots, n_{\text{bins}} \); and let \( \Delta x_1, \ldots, \Delta x_n \) be the corresponding bin widths. Then the bins are defined as

\[
B_{j_1} = x_{i,\text{min}} + [(j_1 - 1)\Delta x_1, j_1 \Delta x_1] \\
\vdots \\
B_{j_n} = x_{i,\text{min}} + [(j_n - 1)\Delta x_n, j_n \Delta x_n],
\]

(3.1)

The discretization entropy is the averaged conditional entropy of \( Y \), for which uniformly distributed input values is simply

\[
H_\Delta = \left( \frac{1}{n_{\text{bins}}} \right)^n \sum_{j_1, \ldots, j_n=1}^{n_{\text{bins}}} \frac{H(Y|X_1 = x_1, \ldots, X_n = x_n \in B_{j_1})}{n_{\text{bins}}}.
\]

(3.2)

The summation extends over the total number of input bin combinations, which is \((n_{\text{bins}})^n\). Since this number can be very large, equation (3.2) cannot be generally evaluated. However, Monte Carlo estimates can provide excellent approximations. Figure 7 shows the estimated discretization entropy as a function of the number of bin combinations used to compute the average. Only several hundred bin combinations are required to obtain a reliable estimate. For each bin combination, about 100 evaluations of the feature \( Y \) were performed to estimate the local conditional histogram from which the conditional entropy \( H(Y|X_1 = x_1, \ldots, X_n = x_n \in B_{j_1}) \) in equation (3.2) is computed. The figure can be understood as a consequence of the output/feature uncertainty being small, for reasonably fine input discretization. Hence, the conditional histograms approximating \( p(Y|X_1 = x_1, \ldots, X_n = x_n \in B_{j_1}) \) tend to have only very few bins with non-vanishing probability. In our case, the output is discretized into 15 bins out of which usually only one or two have non-zero counts, which can therefore be estimated properly with about 100 data. Thus, \( 600 \times 100 = 60,000 \) evaluations provide a reasonably accurate estimate of the discretization entropy.

3.5. Information balance

The block diagram of the information balance (figure 8) shows that higher order interactions do contribute significantly to the total sensitivity. Moreover, only a small subset of parameter pairs and triplets interact significantly (figures 5 and 6), and we expect such sparse connectivity to continue at higher orders.

3.6. Total sensitivity indices

Total sensitivity indices consist of conditional entropies of the type \( H(Y|\{X_1, \ldots, X_n\} \backslash X_i) \), which therefore can
4. CONCLUSIONS

With the advent of advanced estimation techniques, mutual information has become a viable means of characterizing input-output interactions in complex networks. The framework developed in this paper lays

the theoretical foundations for an information-theoretic sensitivity analysis that assigns credit or influence to input variables in terms of their overall contribution to a system’s output entropy. However, our method is far more than a replacement of analysis of variance by analysis of entropy. The information-theoretic approach does not rely on implicit assumptions of normally distributed outputs and is easily generalized to include non-orthogonal input sequences. Moreover, the information-theoretic approach lends itself well to the analysis of systems with an intrinsically stochastic structure, such as biochemical reactions with small numbers of molecules (Wilkinson 2006). In this case, the noise entropy provides a combined representation
of the output uncertainty due to discretization and all sources of intrinsic stochasticity.

We thank the UK BBSRC for financial support, and Mike White, Pawel Paszek and Caroline Horton for useful discussion. This is a contribution from the Manchester Centre for Integrative Systems Biology (www.mcisb.org).

APPENDIX A

A.1. The sign of interaction information

In this section, we discuss the possibility of negative conditional interaction information (CII) and examine the conditions under which this can occur. Most maps do not seem to have this property, but we provide a carefully constructed example of a simple system with three parameters that can exhibit negative conditional interaction information. Consider the function

\[ Y = f(X_1, X_2, X_3) = g_1(X_1)g_2(X_2)g_3(X_3) + h_1(X_1)h_2(X_2)h_3(X_3), \]

where \( X_1, X_2, X_3 \in [0,2] \), the components \( g_i \) and \( h_i \) are the functions

\[ g_i(x) = \frac{1}{2} \{ \tanh[\lambda x] - \tanh[\lambda(x - 1)] \}, \]

and

\[ h_i(x) = \frac{1}{2} \{ \tanh[\lambda(x - 1)] - \tanh[\lambda(x - 2)] \}. \]

Figure 10 shows a plot of the auxiliary functions. The control parameter \( \lambda \) determines the ‘steepness’ of the sigmoid components. For \( \lambda \to \infty \), \( g_i \) and \( h_i \) become ‘square-wave pulses’. Symbolically, one can write

\[ r = x_1 \prod^* x_1 \prod^* x_2 \prod^* x_3 \prod^* x_1, \prod^* x_2, \prod^* x_3, \prod^*. \]

Hence, the random variables are weighted by indicator functions that respond to their values being either in the lower or upper half of their interval.

Figure 11 shows a schematic visualization of the two possible scenarios (different \( \lambda \)). If all the information-theoretic sensitivity measures capturing first-, second- and third-order effects have positive sign, these contributions—together with the discretization entropy—sum up to the output entropy (figure 11a). Under certain circumstances, the sum of first- and second-order indices exceeds the output entropy (figure 11b), where the pairwise interactions are not providing independent pieces of information. In this case, the excess information is compensated by a negative third-order contribution. Although the sum of all sensitivity indices does equal the output entropy, a meaningful interpretation of the components as second- or third-order sensitivity measures is no longer justified.

Table 1 shows the information balance of the system for two choices of \( \lambda \). Owing to symmetry in the inputs, the information measures within each order of interaction are the same. Therefore, only the sums of first- and second-order terms are provided.

Figure 10. Graphs of the sigmoid pulse functions \( g_\lambda \) and \( h_\lambda \).

Apparently, the combination of non-monotonicity and point symmetry of the system (A 1) leads to the negative information measure. Choosing a smaller value of the control parameter \( \lambda \) destroys the symmetry, and consequently the CII becomes positive. Exact point symmetry is not likely to be a feature of natural systems such as our example from systems biology, since it would require an extreme degree of regularity. Therefore, we conclude that the example presented is a rare exception.

One potential alternative scheme that could provide an information decomposition based solely on non-negative quantities is the maximum entropy approach (Jaynes 1957; Amari 2001; Schneidman et al. 2006). For instance, Amari’s elegant method of information geometry (Amari 2001) requires the calculation of surrogate maximum entropy distributions, referred to as \( p^{(2)}, p^{(3)}, \) etc. Here, \( p^{(2)} \) has the same pairwise marginals as the joint density \( p^{(3)} \), the same tripletwise marginals) but contains no higher order correlations. The Kullback–Leibler divergence of \( p^{(2)} \) and \( p^{(1)} \) can be taken to represent the total entropy attributable to second-order interaction and so on. However, at present, the maximum entropy approach is only fully developed with respect to a decomposition of the joint entropy of a set of variables (Amari 2001; Schneidman et al. 2006). We are currently investigating how to extend this approach to obtain a decomposition of the total information for an input–output relation. One practical concern in using a maximum entropy formalism for sensitivity analysis of complex systems with tens of parameters is that the maximum entropy method is (at present) computationally prohibitive and data-intensive. The maximum entropy densities are high-dimensional and suffer from a larger sampling bias than our approach, which is based on pair- and tripletwise contributions that can easily be corrected for the bias. In addition, in order to become useful for sensitivity analysis, the maximum entropy methodology needs to be extended to provide an information decomposition that
relevant terms have been added. (a) Block diagram of an information balance with positive conditional interaction information (CII). Its components are the discretization entropy $H_a$, the first-order sensitivity indices, the second-order indices and the third-order index given by the conditional interaction information. In this example, the CII is positive and hence an integral component of the output entropy, i.e. the sum of sensitivity indices is less or equal to the output entropy. The sum of discretization entropy plus first- and second-order terms exceeds the output entropy. The negative CII. Here, the sum of discretization entropy plus first and second-order terms exceeds the output entropy. The magnitude of the negative CII equals the information excess. The sum of all indices does equal the output entropy, but the sensitivity indices have different signs. In this case, an interpretation of the CII as a third-order sensitivity index is not meaningful, since the input pairs do not contribute independent pieces of information, yielding non-orthogonal information decomposition.

explicitly identifies the particular most relevant parameter interactions.

Future research will also be directed at further illuminating the relation between properties of the input–output map (e.g. monotonicity) and the sign of higher order sensitivity indices.

A.2. Bias corrected estimates of mutual information

Mutual information $I(X;Y)$ between an input variable $X$ and an output variable $Y$ is a functional of the probability densities of input and output. In practice, these probabilities are usually not known a priori and can only be estimated empirically from a limited number $N$ of independent joint observations (‘trials’) of $X$ and $Y$. The statistical errors made in measuring the probabilities owing to limited sampling leads to a severe systematic error (bias) in the information measures. This section is devoted to explain how we corrected the bias problem. For brevity, we focus on $I(X;Y)$, which we shall simply refer to as $I$, but these considerations would straightforwardly apply to other information quantities used in this paper, such as the conditional mutual information. For the sake of explanation, we suppose that $X$ is an $n$-dimensional variable $X = \{X_1, \ldots, X_n\}$.

![Figure 11. Schematic visualization of information-theoretic sensitivity indices by a block diagram. (a) Block diagram of an information balance with positive conditional interaction information (CII). Its components are the discretization entropy $H_a$, the first-order sensitivity indices, the second-order indices and the third-order index given by the conditional interaction information. In this example, the CII is positive and hence an integral component of the output entropy, i.e. the sum of sensitivity indices is less or equal to the output entropy. The sum of discretization entropy plus first- and second-order terms exceeds the output entropy. The negative CII. Here, the sum of discretization entropy plus first and second-order terms exceeds the output entropy. The magnitude of the negative CII equals the information excess. The sum of all indices does equal the output entropy, but the sensitivity indices have different signs. In this case, an interpretation of the CII as a third-order sensitivity index is not meaningful, since the input pairs do not contribute independent pieces of information, yielding non-orthogonal information decomposition.](image)

<table>
<thead>
<tr>
<th>absolute information</th>
<th>fraction of $H(F) - H_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$ $\lambda = 10$ (15 bins)</td>
<td></td>
</tr>
<tr>
<td>$H_a$</td>
<td>0.7257</td>
</tr>
<tr>
<td>first order</td>
<td>0.3238</td>
</tr>
<tr>
<td>second order</td>
<td>1.0338</td>
</tr>
<tr>
<td>third order</td>
<td>0.0402</td>
</tr>
<tr>
<td>sum</td>
<td>2.135 bits</td>
</tr>
<tr>
<td>$H(F)$</td>
<td>2.1235 bits</td>
</tr>
<tr>
<td>$b$ $\lambda = 1000$ (15 bins)</td>
<td></td>
</tr>
<tr>
<td>$H_a$</td>
<td>0.1191</td>
</tr>
<tr>
<td>first order</td>
<td>0.0067</td>
</tr>
<tr>
<td>second order</td>
<td>0.8178</td>
</tr>
<tr>
<td>third order</td>
<td>-0.1031</td>
</tr>
<tr>
<td>sum</td>
<td>0.8405 bits</td>
</tr>
<tr>
<td>$H(F)$</td>
<td>0.8450 bits</td>
</tr>
</tbody>
</table>

The bias due to sampling with $N$ trials is defined as the difference between the limited sampling average value of information $I_N$ ($I_N$ being a probability-weighted average over all possible $(X,Y)$ outcomes with $N$ trials) and the true value of information $I$. Subtracting the bias from the limited sampling estimate allows for a much more accurate estimation of the true information $I$. We observe that $I(X;Y)$ can be written as the difference between two entropies

$$ I(X;Y) = H(X) - H(X|Y). $$

The bias of $I$ is the difference between the biases of the two entropies. It is well known (Miller 1955) that entropies are biased downward (i.e. their bias is negative). $H(X)$ depends only on the marginal distribution $p(X)$. Its bias is much lower than that of $H(X|Y)$, which depends on $p(x, y)$, which is obviously much harder to sample than $p(x)$. As a result, $I$ is biased upward. Analytical considerations (Panzeri & Treves 1996) show that the bias decreases approximately linearly when increasing the number of trials, and increases approximately exponentially when increasing the number of dimensions $n$ of the $X$-space. This makes it difficult to estimate $I$ for large $n$.

Fortunately, the bias of the entropy can be computed approximately and eliminated by means of a number of techniques. In our experience, one of the most effective is the Bayesian technique of Nemenman et al. (2004), which uses a family of prior distributions that are weighted to produce a uniform expectation of entropy before any data are sampled. As $(X,Y)$ data become available, the entropy estimation is updated as an average over all the possible hypothetical probability distributions weighted by their conditional probability given the data. The algorithm converges to the true value of entropy quite rapidly with the
number of trials $N$. Unless the dimensionality of the space $n$ is too large, residual errors left after this bias reduction are small. Since the uncorrected estimate of $I$ is biased upward, residual errors in the estimation of $I$ tend to be upward as well (Montemurro et al. in press).

To check whether any residual error is small, we have developed a different way to estimate information that is biased downward rather than upward (Montemurro et al. in press). This allows one to check the reliability of information-based sensitivity measures by assessing the proximity of the upper and lower bounds. To produce a downward-biased estimate of $I$, we used the following procedure (Montemurro et al. in press). We considered the entropy that would be obtained if the input was independent at fixed output, that is $p(x|y) = P_p(x|y)$. The corresponding entropy of this ‘independent’ distribution is called $H_{ba}(X|Y)$ and typically has a very small bias because only marginal probabilities have to be sampled.

Alternatively, correlations between input variables can be removed by ‘shuffling’ the data at fixed $Y$, thus creating pseudo $X$-vectors obtained by randomly combining $x_i$ values from different trials in which the value $Y$ was observed. The resulting entropy, called $H_{ba}(X|Y)$, has the same asymptotic value of $H_{ba}(X|Y)$ for an infinite number of trials, but has a much higher bias than $H_{ba}(X|Y)$ for finite $N$. Following the mathematical analysis of Panzeri & Treves (1996), Montemurro et al. showed that the bias of $H_{ba}(X|Y)$ is of the same order of magnitude as the bias of $H(X|Y)$ but typically slightly larger. This observation suggests computing $I$ in the following way:

$$I_{ba} = H(X) - H_{ba}(X|Y) + H_{ba}(X|Y) - H(X|Y).$$

Owing to the bias cancellation created by the entropy terms added to the r.h.s., $I_{ba}$ has the same value of $I$ for an infinite number of trials, but a much smaller bias for finite $N$. Moreover, since $H_{ba}(X|Y)$ is more downward biased than $H(X|Y)$, the resulting bias of $I_{ba}$ is negative (Montemurro et al. in press). Thus, in cases when the upward-biased estimator $I$ and the downward-biased estimator $I_{ba}$ coincide, we can be confident that our information estimate is unbiased. If they do not coincide, their difference provides an idea of the order of magnitude of our uncertainty in the information estimation.

### A.3. Decomposition of the total mutual information

The following three equations are elementary formulae, proofs of which can be found in standard textbooks (e.g. Cover & Thomas 2006):

\[
H(X, Y) = H(X) + H(Y|X),
\]

\[
I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X),
\]

\[
I(X; Y) = H(X) + H(Y) - H(X, Y).
\]

We next derive a decomposition of the total mutual information of three variables, which will serve as an auxiliary formula in the general decomposition with an arbitrary number of variables.

### A.3.1. Decomposition with three variables

By means of theorem (A 4), the total mutual information of a pair of random variables $(X, Y)$ and a third variable $Z$ can also be expressed in terms of entropies:

\[
\begin{align*}
I(X; Y, Z) &= H(X, Y) + H(Z) - H(X, Y, Z) \\
&= H(X) + H(Y|X) + H(Z) - H(X, Y, Z) \\
&= H(X) + H(Y) - I(X; Y) + H(Z) \\
&= H(X, Y, Z),
\end{align*}
\]

where we have applied (A 2) in the second step and (A 3) in the third. Expressing $H(X)$ and $H(Y)$ in terms of mutual information with respect to $Z$ by means of (A 3) we obtain

\[
\begin{align*}
I(X; Y, Z) &= I(X; Z) + H(Y|Z) + I(Y; Z) + H(Y|Z) \\
&= H(Z) - H(X, Y, Z) - I(X; Y).
\end{align*}
\]

Rearranging terms and using the relation $H(Z) - H(X, Y, Z) = -H(X, Y|Z)$ yields

\[
I(X; Y, Z) = I(X; Z) + I(Y; Z) + H(Y|Z) - H(X, Y|Z) - I(X; Y),
\]

where by virtue of (A 4) the remaining entropy terms amount to the conditional mutual information of $X$ and $Y$ given $Z$.

Hence,

\[
I(X; Y, Z) = I(X; Z) + I(Y; Z) + I(X, Y|Z) - I(X; Y).
\]

By conditioning on a fourth variable $W$, one also obtains

\[
\begin{align*}
I(X, Y; Z|W) &= I(X; Z|W) + I(Y; Z|W) \\
&= I(X, Y|Z, W) - I(X; Y|W).
\end{align*}
\]

The decomposition (A 5) has previously been applied in computational neuroscience (Adelman et al. 2003).

### A.3.2. Decomposition with arbitrary number of variables

The general decomposition with an arbitrary number of input variables can be derived by repeated application of (A 5) and (A 6). Assuming independent inputs $X_1, \ldots, X_n$, the total mutual information can be expanded using (A 5) by setting $X=X_1$ and $Y=X_2, \ldots, X_n$:

\[
I(X_{1}, \{X_{2}, \ldots, X_{n}\}; F) = I(X_{1}; F) + I(X_{2}, \ldots, X_{n}; F) + I(X_{1}; X_{2}, \ldots, X_{n}|F) - I(X_{2}, \ldots, X_{n}; F) = 0, \text{ if independent first-order term}
\]

\[
+ I(X_{2}, \{X_{3}, \ldots, X_{n}\}; X_{1}|F).
\]
Thus, a first-order term has been separated. Applying (A 6) to the last term separates a second-order term

$$I(X_1, X_2, \ldots, X_n; F) = I(X_1; F) + \frac{I(X_1, X_2; F)}{I(X_1, X_2; F)} + I(X_2, \ldots, X_n; F)$$

second-order term

$$+ I(X_3, \ldots, X_n; X_1|F) + I(X_2, X_3, \ldots, X_n|X_1, F)$$

$$- I(X_2; X_3, \ldots, X_n|F).$$

Expanding the last two terms, again using (A 6), produces a third-order term

$$I(X_1, X_2, \ldots, X_n; F) = I(X_1; F) + I(X_1, X_2; F) + I(X_2, \ldots, X_n; F)$$

$$+ I(X_3, \ldots, X_n; X_1|F) + I(X_3, X_2; F) - I(X_3, X_2; X_1|F)$$

$$= I(X_1; X_1|F),$$

third-order term

$$+ I(X_4, \ldots, X_n; X_2, F) - I(X_4, \ldots, X_n; X_2|F)$$

$$+ I(X_3, X_4, \ldots, X_n; X_1, F) - I(X_3, X_4, \ldots, X_n; X_1|F)$$

$$- I(X_3; X_4, \ldots, X_n|X_1, F) + I(X_3; X_4, \ldots, X_n|F).$$

Similarly, the remaining terms can be expanded further using auxiliary formulae (A 5) and (A 6), thus splitting up more and more terms of ever higher order, and yielding all possible combinations of input variables for each order.

REFERENCES


