Measures for investigating the contextual modulation of information transmission

D Smyth[*], W A Phillips[†‡] and J Kay[§]

Published in *Network: Computation in Neural Systems* 1996, 7(2):307-316

Abstract

The aim of this paper is to show how information theoretic measures can be used to analyze and interpret the results of psychophysical experiments designed to search for conditions under which information from one source modulates the transmission of information from another source. We therefore use measures of mutual and conditional information to analyse systems with two inputs. The information transmitted by such a system can be split into three components depending on whether it is shared between the two inputs or is specific to each. We are here concerned with distinguishing systems that use one input to modulate transmission of information about the other from systems that simply add both inputs, and show how the three components provide evidence for distinguishing between additive and modulatory effects. We also report numerical simulations of the sampling biasses and variances of these measures as a function of the sample size and propose minimum sample sizes that should be used to overcome the bias.

1 Introduction

Functional specialization is a fundamental feature of the mammalian cerebral cortex at both cellular and systems levels of organization. This does not lead to complete independence between the outputs produced at any stage of processing, however, and it has been suggested that this is because it is balanced by functional integration [1] and the search for coherence [2]. One way of integrating activity is by modulating transmission through processors within the system so that they emphasize information that is coherently related to that being concurrently transmitted by other processors. This would be a useful computational strategy because information about one environmental variable is often spread across many different streams of processing, and even when different streams are concerned with different environmental variables they will often be statistically related. These statistical relationships can be used in learning algorithms that discover important environmental variables [3, 4, 5], but our concern here is with how such statistical relations may be used to modulate the transmission of information by the short-term processing dynamics. There is anatomical evidence [6, 7, 8] for long-range cortico-cortical connections that could provide information about activity in other streams and other stages, and physiological evidence [9] suggesting that these connections may have a primarily modulatory or gain-controlling role [10].

The above suggests that what we now need are techniques for detecting whether contextual modulation actually occurs. The purpose of this paper is therefore to show how measures of mutual and conditional information can be used to investigate contextual modulation in biological systems. Together with Ben Craven we are using these measures in psychophysical experiments that explore the interaction between cues in human visual texture segregation. The general approach is intended to be

[*]Institut für Informatik, Universität Leipzig, Pf. 920, Leipzig 04009, Germany.
[†]Department of Psychology, University of Stirling, Stirling FK9 4LA, UK.
[‡]To whom correspondence should be addressed.
[§]Department of Statistics, University of Glasgow, Glasgow G12 8QW, UK.

applicable to many other forms of data, however, including that provided by single-unit electrophysiology.

Here we study the case in which there are two binary input variables and at least one binary output variable. In empirical studies using this paradigm the input variables are chosen such that transmission of information about one of them may be modulated by the other. The input doing the modulation will be called the contextual field (CF), and the other will be called the receptive field (RF). Although not discussed further here the general approach allows for mutual modulatory interactions. It is assumed that the input variables can be presented separately or together, and that performance can be studied under conditions producing transmission of different amounts of information about the target variable.

The investigation of visual texture segregation provides a concrete example. Subjects are shown a square 20 x 20 matrix of small line elements that vary in length and orientation, and they are asked to judge whether the centred boundary separating long from short elements is horizontal or vertical. A CF input is provided by varying the orientation of the small line elements across a vertical or a horizontal boundary. This secondary texture boundary is the same as the target boundary with a specified probability. The question of central theoretical interest in this case is therefore whether the orientation of the secondary texture boundary modulates the transmission of information about the target texture boundary. The strength of the evidence for each of the two texture boundaries is varied by changing the size of the difference in the mean lengths for the RF and the orientation difference for the CF. The data provided by such an experiment will be 2 x 2 x 2 frequency tables at various levels of target detectability, both for context present and for context absent. Here we show how such data can be analyzed and interpreted. We contrast modulatory effects of context with direct linear effects, and we do this for each of the three components of transmitted information.

Important issues that arise when using information theoretic measures to analyze empirical data are those of sampling bias and variance. The measures are defined upon hypothetical probability distributions, and these must be estimated from frequency tables compiled from a sample of limited size (N). This introduces an upward bias on the values calculated for the mutual information between two variables [11, 12, 13, 14, 15, 16, 17]. The bias can be expressed analytically, but only as an asymptotic expansion whose convergence is not guaranteed [17]. Numerical simulations are therefore required to study the biasses and variances under specified conditions. Treves and Panzeri [17] suggested a correction that could be applied to mutual information measures, and their simulations showed it to be a good approximation in the case of two-way mutual information. As we are here concerned with three-way mutual information and with conditional mutual information measures we simulated those cases in order to study the biasses and variances as a function of the sample size. Study of the variances is important for the simple practical reason that in designing such experiments we need to know how the reliability of the estimates increases with N.

2 Input and Output Variables

Information theory is used here to provide measures of the mutual information between binary input and output variables. Continuous variables are associated with these binary variables, however. The continuous variable can be seen as expressing the strength of evidence for, or the confidence in, a binary decision. Spike rates averaged over some time interval (and/or some input population) have often been interpreted as probabilistic binary codes or as signalling the confidence in, or relevance of, a binary decision [18]. In the psychophysical example outlined in the introduction the binary variables are the orientations of the texture boundaries. The associated continuous variables are those used to vary the strength of the evidence for these boundaries. In the simulations reported here inputs and outputs are continuous positive and negative values, and their sign specifies the value of the associated binary variable. The system we use has two input fields: a receptive field (RF) and a contextual field (CF). The issue is then to determine whether the CF modulates transmission of information about the RF. The RF input, CF input, and system output are described by the binary variables R, C and X respectively. Particular binary values are specified by $\mathbf{r}$, $\mathbf{c}$ and $\mathbf{x}$ (i.e. the lower case letters in bold) and the associated continuous values by r, c and x (i.e. the lower case letters in plain).

3 Information Theoretic Measures

Information theory has been proposed as a useful framework within which one can characterize the efficiency of information transmission [19]. The amount of information in an event is defined as being inversely proportional to the probability of that event. Because the information in two independent events should be additive, information is defined using logarithms:

$$I(\mathbf{e}) = \log \frac{1}{P(\mathbf{e})} \tag{1}$$

where $\mathbf{e}$ is some event and $P(\mathbf{e})$ is the probability of that event as specified by some underlying distribution. The base of the log is arbitrary. Throughout this paper we use base 2 to measure information in bits. The mutual information between two events (the information shared between them) can be expressed as the difference between the information gained by observing one event alone and the information gained by observing the other event given the first. So in our system if we describe an instance of the RF distribution by $\mathbf{r}$ and an instance of the output distribution by $\mathbf{x}$, then the mutual information shared by the systems stimulus ($\mathbf{r}$) and its response ($\mathbf{x}$) is:

$$\begin{aligned} I(\mathbf{x};\mathbf{r}) &= I(\mathbf{x}) - I(\mathbf{x} \mid \mathbf{r}) \\ &= \log \frac{1}{P(\mathbf{x})} - \log \frac{1}{P(\mathbf{x} \mid \mathbf{r})} \\ &= \log \frac{P(\mathbf{x} \mid \mathbf{r})}{P(\mathbf{x})} \end{aligned} \tag{2}$$

Since we are also concerned with a CF, we need to extend this to three variables. Using $\mathbf{c}$ to denote an instance of the CF input distribution, we can quantify the mutual information shared by *both* input stimuli and the output response using:

$$\begin{aligned} I(\mathbf{x};\mathbf{r};\mathbf{c}) &= I(\mathbf{x};\mathbf{r}) - I(\mathbf{x};\mathbf{r} \mid \mathbf{c}) \\ &= I(\mathbf{x}) - I(\mathbf{x} \mid \mathbf{r}) - I(\mathbf{x} \mid \mathbf{c}) + I(\mathbf{x} \mid \mathbf{r},\mathbf{c}) \\ &= \log \frac{1}{P(\mathbf{x})} - \log \frac{1}{P(\mathbf{x} \mid \mathbf{r})} - \log \frac{1}{P(\mathbf{x} \mid \mathbf{c})} + \log \frac{1}{P(\mathbf{x} \mid \mathbf{r},\mathbf{c})} \\ &= \log \frac{P(\mathbf{x} \mid \mathbf{r})P(\mathbf{x} \mid \mathbf{c})}{P(\mathbf{x})P(\mathbf{x} \mid \mathbf{r},\mathbf{c})} \end{aligned} \tag{3}$$

The average of this term can be written as:

$$I(X;R;C) = \sum_{\mathbf{x},\mathbf{r},\mathbf{c}} P(\mathbf{x},\mathbf{r},\mathbf{c}) \log \frac{P(\mathbf{x} \mid \mathbf{r})P(\mathbf{x} \mid \mathbf{c})}{P(\mathbf{x})P(\mathbf{x} \mid \mathbf{r},\mathbf{c})} \tag{4}$$

and was used as a learning objective in the Coherent Infomax algorithm [5] for discovering the RF structure that is statistically related to context.

Equation 4 tells us the amount of mutual information between the RF and CF inputs that is transmitted by the system. There can also be two other components of transmitted information however, i.e. $I(X;R \mid C)$ and $I(X;C \mid R)$. The first is what the output tells us about the target in addition to any information that it shares with the context. The second is what the output tells us about the context in addition to any information that it shares with the target. These can be derived similarly to equation 4:

$$I(X;R \mid C) = \sum_{\mathbf{x},\mathbf{r},\mathbf{c}} P(\mathbf{x},\mathbf{r},\mathbf{c}) \log \frac{P(\mathbf{x} \mid \mathbf{r},\mathbf{c})}{P(\mathbf{x} \mid \mathbf{c})} \tag{5}$$

$$I(X;C \mid R) = \sum_{\mathbf{x},\mathbf{r},\mathbf{c}} P(\mathbf{x},\mathbf{r},\mathbf{c}) \log \frac{P(\mathbf{x} \mid \mathbf{r},\mathbf{c})}{P(\mathbf{x} \mid \mathbf{r})} \tag{6}$$

As we are interested in how the context influences the output, we wish to know how it affects all three components of transmitted information.

4 Transfer Functions

Here we study three activation functions:

$$A_a(r,c) = r + c \tag{7}$$
$$A_m(r,c) = 0.5r\,[1 + \exp(rc)] \tag{8}$$
$$A_b(r,c) = 0.5r\,[1 + \exp(rc)] + c \tag{9}$$

where A_a is an additive activation function, A_m is a modulatory activation function, and A_b combines both additive and modulatory effects into one function. The additive activation function is most commonly used in neural networks for combining inputs from separate sources. The modulatory activation function is that proposed on theoretical and physiological grounds in earlier studies of contextual modulation [5]. The important property is that the RF input determines the sign of the activation and contributes to the magnitude while the CF input can only boost (dampen) the magnitude if the RF and CF inputs agree (disagree) but cannot change the sign. The third activation function combines both additive and modulatory effects. It is likely that both additive and modulatory effects occur together in biological systems so a method of recognizing this is needed. The output of the system is computed using the logistic:

$$y(A) = \frac{1}{1 + \exp(-A)} \tag{10}$$

where A is the activation from equations 7, 8 or 9, and y is a probabilistic binary output.

$$P(X = +1 \mid R = \mathbf{r}, C = \mathbf{c}) = y(A(r,c)) \tag{11}$$
$$P(X = -1 \mid R = \mathbf{r}, C = \mathbf{c}) = 1 - y(A(r,c)) \tag{12}$$

In order to investigate the effects of the context we must specify a correlation between the RF and CF distributions. Throughout this paper we set the single probabilities such that: $\forall \mathbf{r}, \mathbf{c} : P(\mathbf{r}) = P(\mathbf{c}) = 0.5$. We can then describe the correlation simply using the single conditional probability: $P(C = 1 \mid R = 1)$ where 1 denotes the positive class. Using this probability alone we can specify the joint distribution as:

$$P(R = +1, C = +1) = 0.5P(C = 1 \mid R = 1) \tag{13}$$
$$P(R = +1, C = -1) = 0.5[1 - P(C = 1 \mid R = 1)] \tag{14}$$
$$P(R = -1, C = +1) = 0.5[1 - P(C = 1 \mid R = 1)] \tag{15}$$
$$P(R = -1, C = -1) = 0.5P(C = 1 \mid R = 1) \tag{16}$$

5 Simulations

We first show how each of the three measures of transmitted information are related to input strength for each of the transfer functions used. The case where the context has no effect upon the output is included because this can be ensured in empirical investigations by the simple expedient of not presenting the contextual stimuli. Comparison with this case will then provide evidence concerning the nature of any effects that the context has when it is presented. Having shown how the three measures can be used for this purpose we then study their sampling biasses and variances.

5.1 Transfer Functions and Transmitted Information

Figure 1 shows the information transmitted by each of the three components for each of the transfer functions and across a range of RF input strengths but for just a single value of CF input strength ($c = 1$). $P(C = 1 \mid R = 1)$ was set to 0.889972 so that the three components had all the same maximum possible value (0.5 bits). This makes it easier to compare the absolute values but the conclusions drawn below hold for a range of values that is wide but differs across the three components.

First consider how these measures indicate an additive contribution by the CF input. If it has an additive effect then $I(X;R;C)$ and $I(X;C \mid R)$ will both be greater than zero when the RF input is zero. This is because even though the RF input has no influence on the output via the activation, it has an effect through the partial correlation with the CF input. When the RF input is greater than zero then evidence that the CF makes only an additive contribution takes two forms: 1) $I(X;R;C)$ is greater with than without context. The function relating it to RF input strength is the same as the curve without context shifted to the left by an amount that is determined by the strength of CF input. This is because the additive activation function does not distinguish between RF and CF inputs, it just adds all inputs together. 2) $I(X;R \mid C)$ is not affected by context, but $I(X;C \mid R)$ is greater with than without context for low values of RF input strength.

Now consider how these measures indicate that the CF input modulates transmission of information about the RF. 1) If there is a modulatory effect then $I(X;R;C)$ grows more rapidly as a function of RF input strength with than without context. This shows the gain-control aspect of contextual modulation since the effect of the context is proportional to the RF strength in equation 8. 2) In this case $I(X;R \mid C)$ will be greater with than without context for low values of RF input strength, and less for high values. 3) $I(X;C \mid R)$ provides a particularly distinctive indication of modulatory effects, because it then shows a positively skewed dependence upon RF input strength, increasing sharply before decreasing. As this may be a particularly useful measure, figure 2 shows the two-dimensional surface of the additive, modulatory, and combined transfer functions for a wide range of both RF and CF input strengths.

If the context affects information transmission in both additive and modulatory ways, then the $I(X;R;C)$ and $I(X;C \mid R)$ curves using the combined transfer function (equation 9) display the individual characteristics of additive and modulatory effects discussed above.

5.2 Sampling Biasses and Variances

Figure 3 compares the mean and standard deviation of the three information measures sampled against the analytical measures for the modulatory activation function. The mean and deviations were computed from 100 trials per sample size and compared with the analytical measure across a range of sample sizes. The sampled and analytical mean values are approximately equivalent for sample sizes above 200 and the standard deviations decrease only very slowly beyond that. This suggests that a sample size of 200 is a good estimate of a lower bound on the sample size to use. The results also show that whereas there is an upward sampling bias on measures of the two-way and three-way mutual information, there is a *downward* bias on estimates of conditional mutual information. The reason for this is not yet clear.

6 Discussion

In this paper we have reported how mutual and conditional information measures between two input sources and the output in a simple system can be used to distinguish various effects of the inputs. In particular we can say whether or not the effects of some contextual stimuli are additive or modulatory or a combination of both. By comparing sampled vs. analytical measures of the information we have found that 200 samples in such a binary input system is sufficient to get an unbiased estimate of the information measures. It is valuable to have three measures because they describe the effects of modulation on the different components of the transmitted information. It is also valuable because the different measures have different requirements. For example $I(X;R;C)$ requires some correlation between the two input distributions whereas the conditional measures require that the two inputs are not perfectly correlated. These measures can all be applied to the psychophysical paradigm described in the introduction and results from such experiments will be reported elsewhere.

Acknowledgements

This work was supported by Network grant CHRX-CT93-0097 from the Human Capital and Mobility Program of the European Community. We thank Peter Cahusac, Ben Craven, Ralf Der, Dario Floreano, Peter Hancock, Ranald Macdonald and Jim Stone for valuable discussions of these issues. We are also very grateful to an anonymous referee for helpful comments.

References

[1] Tononi G, Sporns O and Edelman G M 1994 A measure for brain complexity: Relating functional segregation and integration in the nervous system *Proc. Natl. Acad. Sci. USA* **91** 5033–7

[2] Singer W 1990 Search for coherence: A basic principle of cortical self-organization *Concepts in Neurosci.* **1** 1–26

[3] Becker S and Hinton G E 1992 Self-organizing neural network that discovers surfaces in random-dot stereograms *Nature* **355** 161-3

[4] Stone J and Bray A 1995 A learning rule for extracting spatio-temporal invariances *Network* **6** 429-36

[5] Phillips W A, Kay J and Smyth D 1995 The discovery of structure by multi-stream networks of local processors with contextual guidance *Network* **6** 225–46

[6] Gilbert C D 1992 Horizontal integration and cortical dynamics *Neuron* **9** 1–13

[7] Löwel S and Singer W 1992 Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity *Science* **255** 209–12

[8] Singer W 1993 Synchronization of cortical activity and its putative role in information processing and learning *Annual Rev. Physiol.* **53** 349–74

[9] Hirsch J A and Gilbert C D 1991 Synaptic physiology of horizontal connections in cat's visual cortex *J. Neurosci.* **11** 1800–9

[10] Fox K, Sato H and Daw N 1990 The effect of varying stimulus intensity on NMDA-receptor activity in cat visual cortex *J. Neurophysiol.* **64** 1413–28

[11] Miller G A 1955 Note on the bias of information estimates *Information Theory in Psychology: Problems and Methods* ed H Quastler (Glencoe Illinois: Free Press)

[12] Carlton A G 1969 On the bias of information estimates *Psychological Bulletin* **71** 108–9

[13] MacRae A W 1970 Channel capacity in absolute judgement tasks: An artifact of information bias? *Psychological Bulletin* **73** 112–21

[14] Eckhorn R and Pöpel B 1975 Rigorous and extended application of information theory to the afferent visual system of the cat. II: Experimental results *Kybernetik* **17** 7–17

[15] Optican L M, Gawne T J, Richmond B J and Joseph P J 1991 Unbiased measures of transmitted information and channel capacity from multivariate neuronal data *Biological Cybernetics* **65** 305–10

[16] Kjær T W, Hertz J A and Richmond B J 1994 Decoding cortical neuronal signals: Network models, information estimation and spatial tuning *J. Computational Neurosci.* **1** 109–39

[17] Treves A and Panzeri S 1995 The upward bias in measures of information derived from limited data samples *Neural Comput.* **7** 399–407

[18] Krüger J and Becker J D 1991 Recognizing the visual stimulus from neuronal discharges *Trends in the Neurosciences* **14** 282–6

[19] Hamming R W 1980 *Coding and Information Theory* (New Jersey: Prentice Hall)

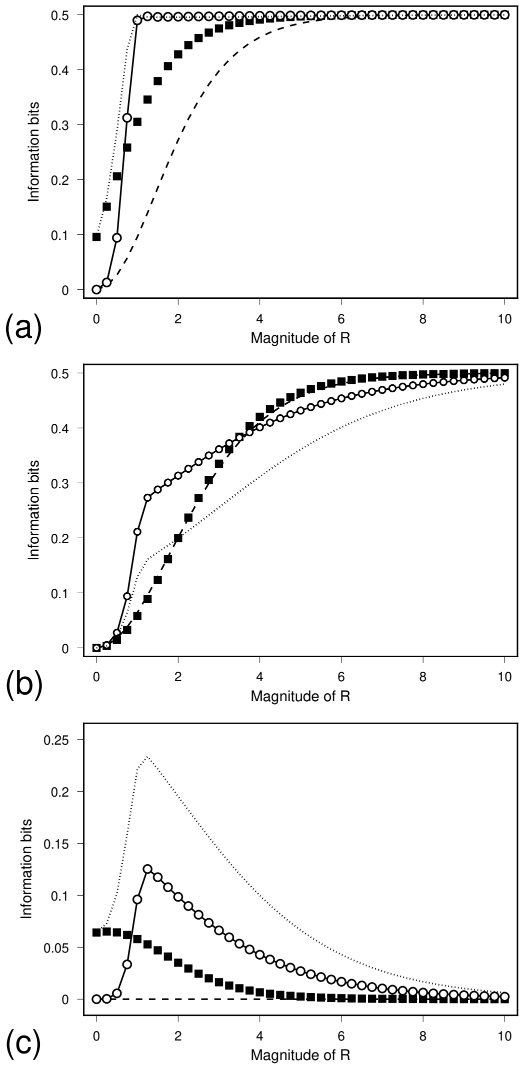

Figure 1: Comparison of the activation functions for $I(X;R;C)$ in subfigure **a**, $I(X;R \mid C)$ in subfigure **b**, and $I(X;C \mid R)$ in subfigure **c**. The activation functions are zero context (dashed), additive context (squares), modulatory context (circles), and both additive and modulatory combined (dotted). The CF input strength was fixed at one while the input conditional probability $P(C=1 \mid R=1) = 0.889972$. All information measures are analytical.

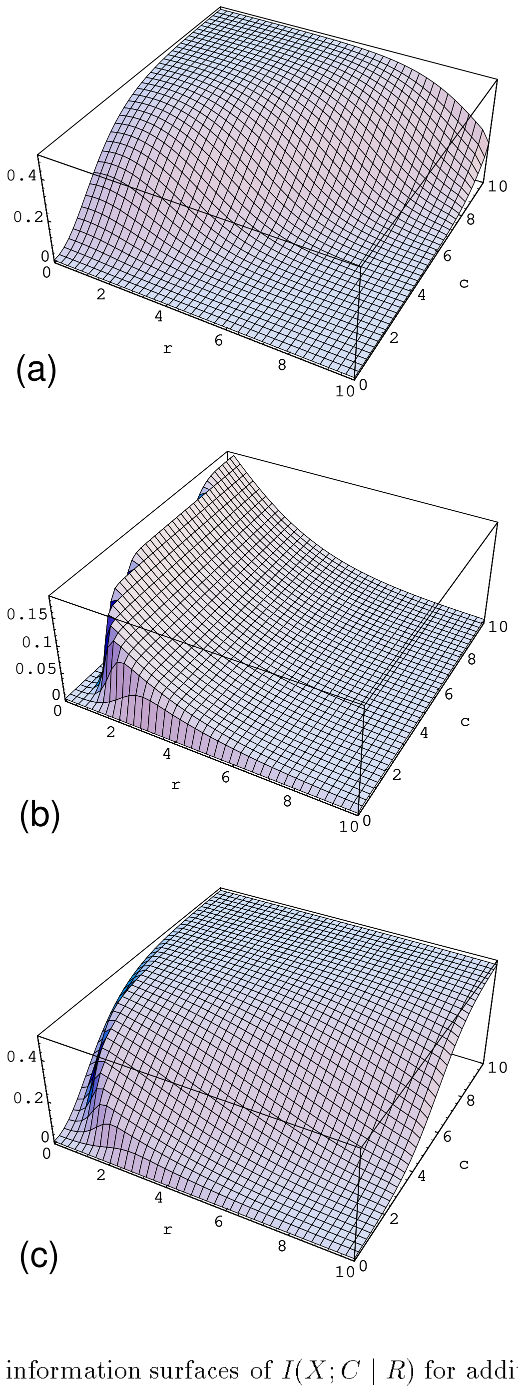

Figure 2: Comparison of the information surfaces of $I(X;C \mid R)$ for additive (**a**), modulatory (**b**) and both (**c**) over all RF and CF input strengths. The conditional probability $P(C=1 \mid R=1) = 0.889972$. All information measures are analytical.

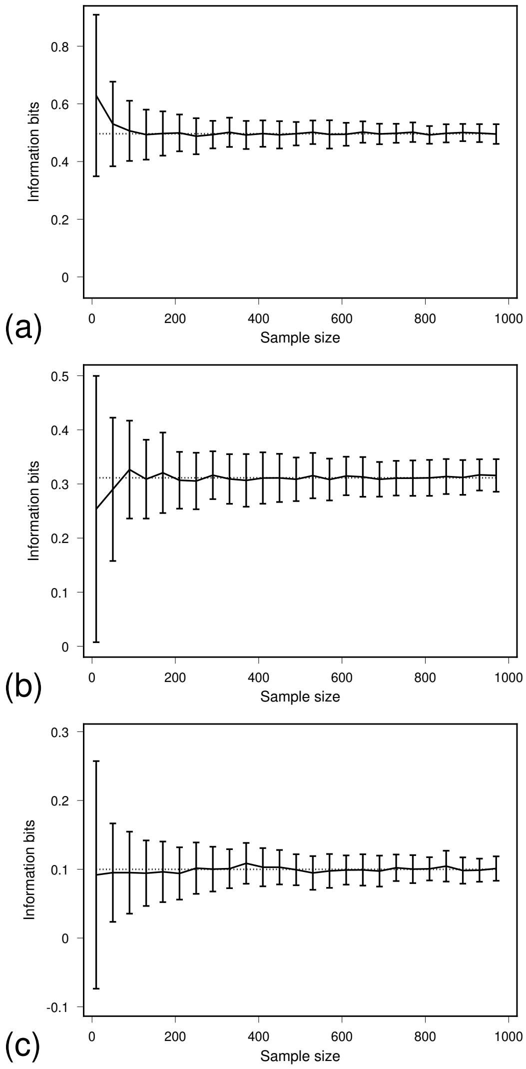

Figure 3: Comparison of the analytical (dotted) and sampled (error bars) information measures for $I(X;R;C)$ in subfigure **a**, $I(X;R \mid C)$ in subfigure **b**, and $I(X;C \mid R)$ in subfigure **c**. The error bars represent standard deviation about the mean information computed from 100 trials for each sample size. The RF and CF input magnitudes were set at 2, and $P(C = 1 \mid R = 1) = 0.889972$.