# Prediction, information, and inference: with application to neuroimaging

Charles Zheng and Yuval Benjamini

November 6, 2016

### Abstract

Neuroscientists have a variety of tools for quantifying multivariate dependence: mutual information, linear correlation-based statistics, Fisher information, and more recently, measures of performance on supervised learning tasks such as classification. We argue that both mutual information and classification accuracy capture intuitive properties of an "information metric" for a channel, and we proceed to develop a general axiomatic characterization of information metrics for channels consisting of a pair of input and output random variables. The key axioms of an information metric are that (i) it is a scalar functional of the joint distribution of the input-output pair, (ii) it is zero for independent variables, and positive for dependent variables, (iii) satisfies a generalized *data-processing* inequality, where transformations of the output can only preserve or reduce the information. We show how prediction tasks can be used to define a general class of information metrics which includes mutual information, as well as a novel information metric, *average Bayes accuracy*, which can be considered an "idealization" of classification accuracy. Furthermore, we consider the possibility of developing a general theory of statistical inference for this class of information metrics. Concretely, we derive a lower confidence bound for average Bayes accuracy as well as a novel lower confidence bound for mutual information.

# 1 Introduction

Historically, neuroscience has largely taken a reductionist approach to understanding the nervous system, proceeding by defining elements and subelements of the nervous system (e.g. neurons), and investigating relationship between two different elements, or the response of an element to external stimulation: say, the response of a neuron's average firing rate to skin temperature. At one level of abstraction, neuroscientists might seek to characterize the functional relationship between elements, but at a higher level of abstraction, it may be sufficient to report scalar measures of dependence. Since neural dynamics are generally both stochastic and nonlinear, it was a natural choice for early neuroscientists to adopt Shannon's *mutual information* as a quantitative measure of dependence. But as new technologies enabled the recording of neural data at larger scales and resolution, the traditional reductionist goals of neuroscience were supplemented by increasingly ambitious attempts within neuroscience to understand the dynamics of neural ensembles, and by efforts originating within psychology and medicine to link the structure and function of the entire human brain to behavior or disease. The larger scope of the data and the questions being asked of the data created an increasing demand for multivariate statistical methods for analyzing neural data of increasingly high dimension. Due to the complexity, variety, and practical difficulties of multivariate statistical analysis of the brain, alternative measures of multivariate dependence such as linear-based correlational statistics, or Fisher information, started to gain traction. For the most part, alternative measures of dependence sacrifice flexibility for a gain in practical convenience: linear-based statistics such as canonical correlation or correlation coefficients fail to capture nonlinear dependencies, and Fisher information requires strong parametric assumptions. Therefore, it was of considerable interest when Haxby (2001) introduced the usage of *supervised learning* (classification tasks) for the purpose of quantifying stimulus information in task fMRI scans. Since then, an entire subfield of neuroimaging, multivariate pattern analysis (MVPA) has been established dedicated to quantifying multivariate information in the brain, and both mutual information and classification accuracy are used by practitioners within the field. Judging from the language used by the practioners themselves, it is intuitively clear to them how classification accuracies can be used to quantify information in brain scans. However, a more thorough examination of the practice raises many questions with regards to the use of classification accuracy as a metric of information: this is one motivation for the current work. But taking a step back, it would seem valuable at this historical juncture to examine the intuitive properties of “information” as a measure of multivariate dependence, and not only consider whether classification accuracy can be considered or used to derive a new information metric, but whether other such metrics might also exist, and whether a unified theory can be developed to account for all of them. This is the larger purpose of the current work, and towards that end we not only propose a general class of information metrics which unifies both information-theoretic and supervised-learning-based approaches, but with an eye toward practical applications, we also examine the question of inferring these quantities from data. An initial result in this direction is the derivation of nonparametric lower confidence bounds for average Bayes accuracy (a novel information metric closely related to classification accuracy,) and an inequality between average Bayes accuracy and mutual information, which, combined with the preceding result, yields a novel lower confidence bound for mutual information.

## 1.1 Organization

The rest of the paper is organized as follows. Section 2 plays the role of a “background” section that gives the basics of mutual information and supervised learning as they are used in neuroscience, but also introduces our axiomatization of information. In section 3, we further explore some of the themes developed in the preceding section relating information, uncertainty, and prediction, and introduce a general class of information metrics which satisfies our axioms. We define a new information metric belonging to this class, average Bayes accuracy, and we also show how mutual information can be considered an “extended member” of the class. In section 4 we develop the basic theory of what kinds of inferences about our information metrics are possible discuss the kinds of experimental designs and supervised learning pipelines which are needed to enable such inference. Concretely, we develop a lower confidence bound for average Bayes accuracy. In section 5 we outline a comparative theory for different metrics within our framework: how are the different information metrics related? We discuss the calculus of variations as a possible general technique for establishing inequalities between different information metrics, and in particular we derive a “randomized Fano’s inequality”: a lower bound for mutual information as a function of average Bayes accuracy. Combined with our lower confidence bound for average

Bayes accuracy, this yields a novel lower confidence bound for mutual information. We provide a practical data analysis example in section 6. A discussion section includes future directions and loose ends are treated, and most of the technical proofs and lemmas are found in the appendix.

# 2 Measures of information

## 2.1 Mutual information and its usage

While Shannon's theory of information was motivated by the problem of designing communications system, the applicability of mutual information was quickly recognized by neuroscientists. Only four years after Shannon's seminal paper in information theory (1948), McKay and McCullough (1952) inaugurated the application of mutual information to neuroscience. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ have joint density $p(\boldsymbol{x}, \boldsymbol{y})$ with respect to the product measure $\mu_x \times \mu_y$, then the mutual information is defined as

$$\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y}) = \int p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} d\mu(\boldsymbol{x}) d\mu(\boldsymbol{y}).$$

where $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ are the marginal densities with respect to $\mu_x$ and $\mu_y$[1]. Since then, mutual information has enjoyed a celebrated position in both experimental and theoretical neuroscience. Experimentally, mutual information has been used to detect strong dependencies between stimulus features and features derived from neural recordings, which can be used to draw conclusions about the kinds of stimuli that a neural subsystem is designed to detect, or to distinguish between signal and noise in the neural output. Theoretically, the assumption that neural systems maximize mutual information between salient features of the stimulus and neural output has allowed scientists to predict neural codes from signal processing models: for instance, the center-surround structure of human retinal neurons matches theoretical constructions for the optimal filter based on correlations found in natural images [cite].

The mutual information measures the information "capacity" of a channel consisting of an input $\boldsymbol{X}$ and an output $\boldsymbol{Y}$, and satisfies a number of important properties.

[1] Note that the mutual information is invariant with respect to change-of-measure.

1. The channel input $\boldsymbol{X}$ and output $\boldsymbol{Y}$ can be random vectors of arbitrary dimension, and the mutual information remains a scalar functional of the joint distribution $p(\boldsymbol{X}, \boldsymbol{Y})$.

2. When $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, $\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y}) = 0$; otherwise, $\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y}) > 0$.

3. The data-processing inequality: for any vector-valued function $\vec{f}$ of the output space,
$$\mathrm{I}(\boldsymbol{X}; \vec{f}(\boldsymbol{Y})) \leq \mathrm{I}(\boldsymbol{X}; \boldsymbol{Y}).$$

4. Symmetry: $\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y}) = \mathrm{I}(\boldsymbol{Y}; \boldsymbol{X})$.

5. Additivity: if $(\boldsymbol{X}_1, \boldsymbol{Y}_1)$ is independent of $(\boldsymbol{X}_2, \boldsymbol{Y}_2)$, then
$$\mathrm{I}((\boldsymbol{X}_1, \boldsymbol{Y}_1); (\boldsymbol{X}_2, \boldsymbol{Y}_2)) = \mathrm{I}(\boldsymbol{X}_1; \boldsymbol{Y}_1) + \mathrm{I}(\boldsymbol{X}_2; \boldsymbol{Y}_2).$$

Two additional consequences result from the data-processing inequality:

- *Invariance under bijections.* If $\vec{f}$ has an inverse $\vec{f}^{-1}$, then
$$\mathrm{I}(\boldsymbol{X}; \vec{f}(\boldsymbol{Y})) \leq \mathrm{I}(\boldsymbol{X}; \boldsymbol{Y}) = \mathrm{I}(\boldsymbol{X}; \vec{f}^{-1}(\vec{f}(\boldsymbol{Y}))) \leq \mathrm{I}(\boldsymbol{X}; \vec{f}(\boldsymbol{Y})),$$
therefore, $\mathrm{I}(\boldsymbol{X}; \vec{f}(\boldsymbol{Y})) = \mathrm{I}(\boldsymbol{X}; \boldsymbol{Y})$.

- *Monotonicity with respect to inclusion of outputs.* Suppose we have an output ensemble $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$. Then the individual component $\boldsymbol{Y}_1$ can be obtained as a projection of the ensemble. By the data-processing inequality, we therefore have
$$\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y}_1) \leq \mathrm{I}(\boldsymbol{X}; (\boldsymbol{Y}_1, \boldsymbol{Y}_2)).$$
Intuitively, if we observe both $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, this can only *increase* the information we have about $\boldsymbol{X}$ compared to the case where we only observe $\boldsymbol{Y}_1$ by itself.

And it is the property of *invariance under bijections*, inclusive of non-linear bijections, which qualifies mutual information as a *non-linear measure of dependence.* Linear measures, such as Pearson correlation, are not invariant under bijections.

Besides the formal definition, there are a number of well-known alternative characterizations of mutual information in terms of other information-theoretic quantities: the *entropy* H:

$$\mathrm{H}_\mu(\boldsymbol{X}) = -\int p(\boldsymbol{X}) \log p(\boldsymbol{X}) d\mu(\boldsymbol{X}),$$

and the *conditional entropy*:

$$\mathrm{H}_\mu(\boldsymbol{X}|\boldsymbol{Y}) = -\int p(\boldsymbol{Y}) d\mu_y(\boldsymbol{Y}) \int p(\boldsymbol{X}|\boldsymbol{Y}) \log p(\boldsymbol{X}|\boldsymbol{Y}) d\mu_x(\boldsymbol{X}).$$

Some care needs to be taken with entropy and conditional entropy since they are not invariant with respect to change-of-measure: hence the use of the subscript in the notation $\mathrm{H}_\mu$. In particular, there is a difference between *discrete entropy* (for counting measure) and *differential entropy* (for Lesbegue measure.) Intutively, entropy measures an observer's uncertainty of the random variable $\boldsymbol{X}$, supposing the observer has no prior information other than the distribution of $\boldsymbol{X}$. Conditional entropy measures the *expected uncertainty* of $\boldsymbol{X}$ supposing the observer observes $\boldsymbol{Y}$.

However, regardless of the base measure, the following identities hold:

$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{H}_{\mu_x \times \mu_y}((\boldsymbol{X},\boldsymbol{Y})) - \mathrm{H}_{\mu_x}(\boldsymbol{X}) - \mathrm{H}_{\mu_y}(\boldsymbol{Y}).$$

$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{H}_\mu(\boldsymbol{Y}) - \mathrm{H}_\mu(\boldsymbol{Y}|\boldsymbol{X}). \tag{1}$$

The second identity (1) is noteworthy as being practically important for estimation of mutual information. Since the entropies in question only depend on the marginal and conditional distributions of $\boldsymbol{Y}$, the problem of estimating $\mathrm{I}(\boldsymbol{X};\boldsymbol{Y})$ can be reduced from a $\dim(\boldsymbol{X}) + \dim(\boldsymbol{Y})$-dimensional nonparametric estimation problem to a $\dim(\boldsymbol{Y})$-dimensional problem: hence this identity is a basis of several methods of estimation used in neuroscience, such as Gastpar (2014).

However, by symmetry, we also have the flipped identity

$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{H}_\mu(\boldsymbol{X}) - \mathrm{H}_\mu(\boldsymbol{X}|\boldsymbol{Y}). \tag{2}$$

In neuroscience studies, where $\boldsymbol{X}$ is the controlled stimulus, and $\boldsymbol{Y}$ is the neural activity, the two mirror pairs (1) and (2) have different interpretations. Rather than providing a basis for practical estimation, (2) provides an *interpretation* of the mutual information. Loosely speaking, $\mathrm{H}_\mu(\boldsymbol{X})$ is the uncertainty of $\boldsymbol{X}$ before having observed $\boldsymbol{Y}$, and $\mathrm{H}_\mu(\boldsymbol{X}|\boldsymbol{Y})$ is the uncertainty of $\boldsymbol{X}$ after having observed $\boldsymbol{Y}$, hence $\mathrm{H}_\mu(\boldsymbol{X}) - \mathrm{H}_\mu(\boldsymbol{X}|\boldsymbol{Y})$ is how much the observation of $\boldsymbol{Y}$ has *reduced* the uncertainty of $\boldsymbol{X}$. Stated in words,

$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \text{average reduction of uncertainty about } \boldsymbol{X} \text{ upon observing } \boldsymbol{Y}.$$

We explore this concept of "information as reduction of uncertainty" much further when we propose our general class of information metrics: there, we see that properties (i)-(iii) of the mutual information are consequences of this second characterization of mutual information.

But how what is the practical import of these properties of mutual information? Supposing we identify a number of properties as being essential for scientific purposes (and others as non-essential), this then suggests that alternative measures of information, also satisfying the same essential properties, could be just as effective for scientific work as mutual information. And some might have additional advantages.

Towards our goal of formulating a general theory of information metrics, it is important to evaluate exactly *how important* each of the listed properties.

[to be continued: applications of mutual information]

- Example: Comparison of decoders in Nelken. Property (i) is important to enable model comparison. Property (iii) is needed because relationships may be nonlinear.

- Example: Redundancy in population code of retina. Property (i)-(iii) and (v) are needed to obtain a meaningful measure of redundancy.

- In general, symmetry not important, but additivity is desirable for measures of redundancy. Property (ii) can usually be enforced since any measure needs to have a unique "minimum" value for the case of independence.

## 2.2 Supervised learning

- Supervised learning task is defined using a prediction task.

- 1. A predictive model is learned using training data

- 2. The performance of the model on the prediction task is estimated using independent test data

- Classical examples of prediction tasks: regression and classification
- Third example: identification
- Definition of Bayes prediction model
- General definition of supervised learning task
- SL performance can be a scalar
- SL can be used to test for independence
- Bayes performance satisfies data-processing inequality
- How SL is interpreted in MVPA as information
- Section 3, we'll see how Bayes performance can be legitimately considered a measure of information

## 2.3 Axiomatic characterization of information

Let $\mathcal{I}(\boldsymbol{X};\boldsymbol{Y})$ denote a general measure of information.

1. Information is a scalar functional of the joint distribution $p(\boldsymbol{X},\boldsymbol{Y})$.
2. When $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, $\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) = 0$; otherwise, $\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) > 0$.
3. The data-processing inequality: for any vector-valued function $\vec{f}$ of the output space,
$$\mathcal{I}(\boldsymbol{X};\vec{f}(\boldsymbol{Y})) \leq \mathcal{I}(\boldsymbol{X};\boldsymbol{Y}).$$

Mutual information satisfies the property. Bayes performance satisfies it in regression case, but not in identification case due to dependence on exemplars.

In the next section, we see how both MI and Bayes perf are examples of a large class of information measures.

# 3 Information, uncertainty and prediction

The general idea is to look at reduction of uncertainty.

- Prediction loss is a measure of uncertainty. Mean-square loss corresponds to variance.
- Consider two prediction problems: in one case you are given no side information, and in the second case you are given $\boldsymbol{Y}$. The difference in risk between the two problems gives a measure of uncertainty reduction.
- In regression, this gives “variance explained” as a measure of information.
- In classification, we get a normalized form of Bayes accuracy.
- In randomized classification, we get (normalized) average Bayes accuracy.
- We can prove that any prediction task yields an information measure i.e. satisfying the three axioms.
- If the no-information risk is known in advance, we say the information measure is estimable.
- Mutual information can be characterized this way, but it is not estimable.
- Mutual information can be derived as a limit of linear combinations of normalized ABA, due to channel coding theorem. Still not estimable due to limit property.

# 4 Statistical inference

- Only lower bounds are possible, because...

## 5 Connections

## 6 Applications

## 7 Discussion