Estimating Mutual Information from Misclassification Rates

Charles Zheng and Yuval Benjamini

December 8, 2015

Abstract

Mutual information is a useful measure of dependence between the input of a neural subsystem, X, and its output, or measured output, Y, due to its flexibility for capturing nonlinear associations, and its rich information-theoretic context. However, since existing nonparametric methods for estimating mutual information scale so poorly for high-dimensional stimulus and response spaces, researchers often rely on supervised learning techniques to characterize nonlinear dependence in such settings. We exploit the relationship between mutual information and Bayes error to obtain an estimate of mutual information from misclassification rate: such a procedure can be viewed as an indirect way of leveraging prior information about the dependence structure of X and Y to obtain a better estimate of mutual information. However, the quality of the resulting estimate of mutual information depends on the accuracy of the predictive model, as well as the availability of data which which to train and test the predictive model. Under a particular high-dimensional, low-SNR regime, we obtain an upper bound on the estimation error $|\hat{I}(X;Y) - I(X;Y)|$ of order $O(d/n)$, where d is the dimensionality of the model and n is the number of observations. In simulated data, we compare our proposed approach to existing nonparametric methods for estimating $I(X;Y)$, both when the model is correct, and when the model is mildly misspecified.

1 Introduction

In neuroscience, one is often interested in measuring depending between the input of a neural subsystem, X, and its output, or measured output, Y. A variety of measures of dependence–including as correlation, mutual information, and prediction error, tend to be used in different settings because of their respective tradeoffs between interpretability, flexibility, and high-dimensional scalability. Correlation is highly interpretable and scalable, but is inflexible: it fails to capture many forms of nonlinear dependence. Mutual information is interpretable and flexible, but it is generally difficult to estimate in high-dimensional settings. Hence, in such settings, when a nonlinear measure of dependence is desired, a powerful approach is to apply supervised learning methods to either predict X based on Y (decoding), or Y based on X (encoding). Of course, predictive models are extremely interesting beyond the measure of prediction error: one commonly examines the fitted model to find clues to the underlying dynamics of the system. However, one is still often interested in a one-dimensional summary of the dependence structure: in that regard, while prediction error and mutual information are both interpretable, mutual information has the added advantage of its rich context in information theory, while prediction error has the disadvantage of the arbitrariness of the loss function. Even when considering misclassification error, one often faces the problem of how to partition a high-dimensional space into discrete classes. Furthermore, the ideal definition of prediction error is the *Bayes error*: the prediction error of the optimal rule, but obtaining the Bayes error depends on being able to learn the correct model, as well as having infinite data to fit the model. That said, in many problems it may be more feasible to estimate the Bayes error than to obtain a fully nonparametric estimate of the mutual information, since we can easily exploit prior knowledge about the dependence structure between x and Y (for instance, a generalized linear model) to train the predictive model, while nonparametric estimators of mutual information fail to exploit this prior knowledge.

In fact, one could exploit the strong relationship between mutual information and the prediction error to obtain an estimate of the mutual information from the observed classification rate. For example, using generalizations of Fano's inequality, one can obtain a lower bound on mutual information in relation to the optimal prediction error, or Bayes error. Such a technique for obtaining estimates of mutual information from classification rates can be understood as a way to leverage the prior information about X and Y

implied by the prediction model in order to obtain a *model-based* estimate of mutual information, $\hat{I}(X;Y)$. However, a prominent challenge to such an approach is the *finite sample bias* resulting from having a limited number of observations N for training and testing the prediction rule.

However, in low-SNR settings, which are commonly encountered in applications, we find that the connection between mutual information and prediction error in the form of misclassification rate, can be made even stronger than the lower bound implied by Fano's inequality. In a particular low-SNR regime, we find an exact asymptotic relationship between the Bayes misclassification probability and the mutual information. Furthermore, our framework allows us to characterize the discrepancy between the observed misclassification rate, and the Bayes error, which allows us to derive that the sample complexity of estimating the mutual information: $|\hat{I}(X,Y) - I(X;Y)|$ is of the order $O(1/N)$ in the number of combined training and testing observations. Our approach depends on an assumption of restricted dependence between components (X_i, Y_i) and (X_j, Y_j), but the assumption holds in many cases of practical interest.

1.1 Motivation

The specific setup we consider was motivated by a number of studies

- Face recognition in monkeys
- Identification of natural images

1.2 Setup

Assume X and Y are real random vectors with the same dimensionality, d. Our results are derived under a model where X has a continuous density $p(x)$, but in which the experimenter observes multiple repeats of Y conditional on a common X. The data therefore consists of tuples $(x^{(i)}, y^{(i,1)}, \ldots, y_i^{(i,r)})$, where $x^{(i)}$ is the ith unique stimulus, and $y^{(i,1)}, \ldots, y^{(i,r)}$ are the repeats of Y given $X = x^{(i)}$, which are assumed to be conditionally independent given X.

Let K denote the number of unique stimuli. The data therefore consists of $n = Kr$ observations. When r is large, the data can be nearly considered as i.i.d. observations from the joint distribution of $(\tilde{X}, \tilde{Y})$, where $\tilde{X}$ has a

distribution $\tilde{p}$ consisting of a mixture of point masses at $x^{(i)}$:

$$\tilde{p} = \frac{1}{K} \sum_{i=1}^{K} \delta_{x^{(i)}},$$

and $\tilde{Y}|\tilde{X} = x^{(i)}$ has the same distribution as $Y|X = x^{(i)}$.

Yet, although our data was collected from the distribution $(\tilde{X}, \tilde{Y})$, our goal is to estimate $I(X;Y)$ rather than $I(\tilde{X};\tilde{Y})$. In order for the two quantities to have any connection, the selected stimuli $x^{(i)}$ must be 'representative' of the continuous distribution X. When the stimulus X is very high-dimensional, it becomes quite reasonable to draw $x^{(i)}$ i.i.d. from the marginal distribution $p(x)$. This ensures that $I(\tilde{X};\tilde{Y})$ converges to $I(X;Y)$ as $K \to \infty$. Though, as noted by Gastpar et. al., for finite K, $I(\tilde{X};\tilde{Y})$ tends to result in an underestimate of $I(X;Y)$. This motivates their antropic correction method for estimating $I(X;Y)$, which can be applied directly in this setting supposing that one has a method for estimating the conditional entropies $h(Y|X = x^{(i)})$.

In contrast, we will consider the misclassification errror as a means to estimate the mutual information. Letting $p(x,y)$ denote the density of (X,Y), the Bayes rule for predicting $\tilde{X}$ from $\tilde{Y} = y^*$ is given by

$$\hat{X}_{Bayes} == \text{argmax}_{x=x^{(1)},\ldots,x^{(k)}} \log p(y|x)$$

where $p(y|x) = p(x,y)/p(x)$. The Bayes error is

$$\Pr[\tilde{X} \neq \hat{X}_{Bayes}],$$

where the probability is taken over the joint distribution of $(\tilde{X}, \tilde{Y})$. Since the Bayes error depends on the sample of representative stimuli $\{x^{(i)}\}$, we find it more useful to consider the average Bayes error:

$$\text{MC} = \mathbf{E}[\Pr[\tilde{X} \neq \hat{X}_{Bayes}]],$$

where the outer expectation is over the distribution of $x^{(i)} \sim p(x)$. The following sections explore the relationship between MC and $I(X;Y)$.

As a means to estimate the average Bayes error MC, we fit a predictive model for $\tilde{X}$ given $\tilde{Y}$. This results in a K-class classification problem. While in practice, a variety of multi-class classification methods can be employed, our theory depends on having a known, semiparametric generative model for

the conditional distribution of Y: we study the misclassification rate obtained by using the maximum-likelihood plugin estimate of the Bayes rule.

Hence, when deriving sample complexity results, we make the further assumptions that

$$p(x, y) = p(x)q(y|\mu(x))$$

where μ is an unknown bijection from $\mathbb{R}^p \to \mathbb{R}^p$, and $q(y|\mu)$ is a known parametric family of density functions which are jointly differentiable in y and μ. The model is semiparametric since we do not make any constraints on the function μ, other than invertibility. In fact, X can be removed from the picture since $I(X;Y) = I(\mu;Y)$, where $\mu = \mu(X)$. This reflects practice in many neuroimaging studies where the actual pixel values of the stimuli are not incorporated in the model at all; rather, one simply models the joint distribution of the class of the stimulus and the response. On the other hand, it is worth noting that the model-based approach demonstrated in Kay et al., and others, do model the mapping μ.

In order to get an estimate of the misclassification rate, one has to *hold out* a number r_{test} of the repeats from each class. The classification rule is based on estimates of $\mu^{(i)} = \mu(x^{(i)})$, given by the MLE estimator on the training set,

$$\hat{\mu}^{(i)} = \text{argmax}_\mu \sum_{j=1}^{r_{train}} \log q(y^{(i,j)}|\mu).$$

The MLE classification rule is therefore defined as

$$\hat{X}_{MLE} = x^{(i)} \text{ where } i = \text{argmax}_i \log q(y^*|\mu).$$

The sample test error is therefore

$$\frac{1}{Kr_{test}} \sum_{i=1}^{K} \sum_{j=r_{train}+1}^{r} I(\hat{x}_{MLE}^{(i,j)} \neq x^{(i)}).$$

As an estimate of MC, the sample test error has variability both from the randomness in $\tilde{Y}$ conditional on the sampled stimuli $x^{(i)}$, and from the randomness in the sampled stimuli drawn from $p(x)$. Therefore, it makes sense to repeat the procedure for m independent *samples* of $(x^{(1)}, \ldots, x^{(K)})$, and then averaging the resulting test errors. Let the resulting average misclassification rate be denoted $\hat{\text{MC}}$. In later sections we will study the discrepancy between $\hat{\text{MC}}$ and MC, and how to optimally choose the experimental parameters K and r given a total budget of $N = Kmr$ observations.

2 Theory

2.1 Application of classical results

- Using Fano's inequality
- Limitations
- Define $\tilde{X}$ to be the discretization of X
- Define $I(F)$ to be the mutual information $I(X;Y)$ when $(X,Y) \sim F$.

2.2 Low-SNR model

We have seen in the previous section that the lower bound implied by Fano's inequality is quite inaccurate when (...). Certainly, an exact relationship between $I(X;Y)$ and the Bayes error cannot hold since given two different joint distibutions F, G with $I(F) = I(G)$, the K-class misclassification rate MC may be quite different between F and G. Yet, we observe that under the two conditions that (i) the dimensionality of (X,Y) is high, and (ii) the signal-to-noise ratio is low, in the sense that $H(X,Y) >> I(X;Y)$, the relationship between information and misclassification rate begins to cohere.

- Given a counterexample (gaussian)
- Plot of MC depending on K.
- Examples of low SNR regime. Varying $I(X;Y)$ and also dimensionality
- In all plots, compare with Fano inequality
- As we can see, low SNR regime gets more accurate than Fano

Assume that X and Y have joint density $p(x,y)$ with respect to Lesbegue measure on $\mathbb{R}^{2d}$. Draw i.i.d. $(X^{(i)}, Y^{(i)})$ from the joint distribution, for $i = 0, \ldots, K-1$, and let (X^*, Y^*) denote $(X^{(0)}, Y^{(0)})$. Define

$$Z_i = \log p(Y^*|X_i) = \log p(Y^*, X_i) - \log p(X_i).$$

The Bayes rule is therefore

$$\hat{X} = x^{(i)} \text{ where } i = \text{argmax}_i Z_i$$

It turns out the reason why the dimensionality and signal-to-noise ratio play a role is because those conditions ensure that the vector $Z = (Z_*, Z_1, \ldots, Z_{K-1})$ has an approximately normal distribution. However, to formally prove this fact, we require an asymptotic framework.

2.2.1 Asymptotics

In order to establish asympotic normality, we consider a limiting sequence of problems of increasing dimensionality d. Let $(X^{[d]}, Y^{[d]})$ denote the joint distributions in the sequence, for $d \in \{1, 2, \ldots\}$. As d increases, the ratio of the information $I(X^{[d]}; Y^{[d]})$ and the joint entropy $H(X^{[d]}, Y^{[d]})$ decreases.

Before giving a general result, we illustrate this aymptotic regime by the following gaussian example. Let

$$\begin{bmatrix} X^{[d]} \\ Y^{[d]} \end{bmatrix} \sim N\left(0, \begin{bmatrix} I & \frac{1}{\sqrt{1+d\sigma^2}} I \\ \frac{1}{\sqrt{1+d\sigma^2}} I & I \end{bmatrix}.\right)$$

For fixed d, we have $(X_i^{[d]}, Y_i^{[d]})$ drawn i.i.d. from a bivariate normal $N(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$ where $\rho = (1 + d\sigma^2)^{-1}$. Recalling that the mutual information of the components of such a bivariate normal is $-\log(1-\rho^2)/2$, we easily calculate:

$$I(X^{[d]}, Y[d]) = \sum_{i=1}^{d} I(X_i^{[d]}, Y_i^{[d]}) = -\frac{d}{2}\log(1 - \frac{1}{1+d\sigma^2}).$$

Hence

$$\begin{aligned} \lim_{d\to\infty} I(X^{[d]}, Y^{[d]}) &= \lim_{d\to\infty} -\frac{d}{2}\log(1 - \frac{1}{1+d\sigma^2}) \\ &= \lim_{d\to\infty} \frac{d}{2}\frac{1}{1+d\sigma^2} = \frac{1}{2\sigma^2}. \end{aligned}$$

Meanwhile, $H(X^{[d]}) = H(Y^{[d]}) = \frac{d}{2}\log(2\pi)$, so it is clear that $H(X^{[d]}, Y^{[d]}) >> I(X^{[d]}; Y^{[d]})$.

A simple calculation shows that

$$Z_i = \log p(Y^*|X^{(i)}) = -\frac{1}{2(1-\rho^2)}||Y^* - \rho X^{(i)}||^2 + \text{ const.}$$

where the first term is a scaled chi-squared distribution with d degrees of freedom: the scale is -1/2 for $i = 1, \ldots, K-1$ and $-(1-\rho^2)/2$ for $i = 0$. Since we can separate Z_i into independent, componentwise sums,

$$Z_i = \text{const.} \;\; - \frac{1}{2(1-\rho^2)} \sum_{j=1}^{d} (Y_j^* - \rho X_j^{(i)})^2,$$

it follows from the multivariate central limit theorem that Z_i are asymptotically jointly normal.

A straightforward computation using multivariate normal moments (c.f. Muirhead) yields the limiting moments:

$$\mathbf{E}[Z_*] = -\frac{d}{2} + \text{const.}, \;\; \mathbf{E}[Z_i] = -\frac{d}{2}\frac{1+\rho^2}{1-\rho^2} + \text{const.}$$

$$\text{Var}[Z_*] = \frac{d(1-\rho^2)^2}{2}, \;\; \text{Var}[Z_*, Z_i] =$$

Taking limits, the moments simplify to yield

$$\begin{bmatrix} Z_* \\ Z_1 \\ \vdots \\ Z_{K-1} \end{bmatrix} \sim N \left(\begin{bmatrix} -\frac{d}{2} \\ -\frac{d}{2} + \frac{1}{\sigma^2} \\ \vdots \\ -\frac{d}{2} + \frac{1}{\sigma^2} \end{bmatrix}, \begin{bmatrix} d/2 & d/2 & \cdots & d(1-\rho^2)^2/2 \\ d/2 & & \cdots & \end{bmatrix} \right).$$

The misclassification probability is

$$\text{MC} = \Pr[Z_* > \min_{i=1}^{K-1} Z_i] = \Pr[N(\mu, \sigma^2) < M_{K-1}]$$

where $\mu =?$ and $\sigma^2 =?$, and M_{K-1} is the maximum of $K-1$ independent standard normal variates. Note that our setup automatically gives the averaged Bayes error rate! Taking the limit, we get $\mu =? = \sqrt{I(X;Y)}$ and $\sigma^2 = 1$. Therefore, by inverting the function

$$f_K(j) = \Pr[N(\sqrt{j}, 1) < M_{K-1}]$$

we get

$$I(X;Y) = f_K^{-1}(\text{MC}).$$

This very result holds under much more general conditions. Our following proof only relaxes the condition that X and Y are jointly normal, still

retaining the coordinatewise independence, and assuming some additional conditions for notational convenience.

Theorem. *Let $X^{[d]}, Y^{[d]}$ be a sequence of distributions satisfying:*

- *For each $d = 1, 2, \ldots$ has an associated bivariate distribution $b_d(x, y)$ and the components $(X_i^{[d]}, Y_i^{[d]})$ are drawn iid from $b_d(x, y)$ for $i = 1, \ldots, d$, where the distributions b_d have uniformly bounded third moments.*
- *$I(X^{[d]}, Y^{[d]}) = c$ is constant*
- *$H(X_i^{[d]})$ and $H(Y_i^{[d]})$ are constant.*

Then, as $d \to \infty$, the misclassification probability

$$MC = \Pr[Z_* > ...]$$

satisfies

$$\lim_{d\to\infty} MC = f_K(I) = \Pr[N(\sqrt{I}, 1) > M_{K-1}].$$

Remark. It follows from the assumptions that $H(X^{[d]}, Y^{[d]}) = O(d)$ while $I(X^{[d]}; Y^{[d]})$ is fixed. These assumptions are sufficient for asymptotic normality of Z_i, but asymptotic normality also follows from much more general conditions, as we will discuss.

Proof. Let $b(x)$ and $b(y)$ denote the marginal distributions of the bivariate density $b(x, y)$. Note that due to the componentwise i.i.d. assumptions,

$$I(X^{[d]}, Y^{[d]}) = \sum_{i=1}^{d} I(X_i, Y_i) = dI(X_1, Y_1)$$

hence

$$I(X_1, Y_1) = c/d.$$

Now define $u(x, y)$ by

$$b(x, y) = b(x)b(y)(1 + u(x, y)).$$

It follows that

$$0 = \mathbf{E}[u(X, Y)|X] = \mathbf{E}[u(X, Y)|Y].$$

Meanwhile, observe that

$$\begin{aligned}
-H(X_1, Y_1) &= \int \log(b(x,y))b(x,y)dxdy \\
&= \int \log(b(x)b(y)(1+u(x,y))b(x)b(y)(1+u(x,y))dxdy \\
&= \int \log(b(x))b(x)\left[\int b(y)(1+u(x,y))dy\right]dx \\
&\quad + \int \log(b(y))b(y)\left[\int b(x)(1+u(x,y))dx\right]dx \\
&\quad + \int \log(1+u(x,y))(1+u(x,y))b(x)b(y)dxdy \\
&= \int \log(b(x))b(x)\mathbf{E}[1+u(X,Y)|X=x]dx \\
&\quad + \int \log(b(y))b(y)\mathbf{E}[1+u(X,Y)|X=y]dx \\
&\quad + \mathbf{E}[\log(1+u(X_1,Y_*))(1+u(X_1,Y_1^*))] \\
&\quad = -H(X_1) - H(Y_1) + \mathbf{E}[\log(1+u(X_1,Y_1^*))(1+u(X_1,Y_1^*))]
\end{aligned}$$

where here X_1, Y_1^* are drawn from the product marginal $b(x)b(y)$. Hence

$$I(X_1;Y_1) = \mathbf{E}[\log(1+u(X_1,Y_1^*))(1+u(X_1,Y_1^*))].$$

But, since $I(X_1;Y_1) = O(1/d)$, it follows that $u(X_1,Y_1)$ is also of order $O(1/d)$ in probability. This allows us to use the approximatation

$$\log(1+u(x,y)) = u(x,y) - \frac{1}{2}u(x,y)^2 + O(d^{-3}).$$

Hence,

$$I(X_1;Y_1) = \mathbf{E}[u(X_1,Y_1^*) + u(X_1,Y_1^*)^2/2] + O(d^{-3}) = \frac{1}{2}\mathrm{Var}[u(X_1,Y_1^*)].$$

Due to componentwise independence, the scores $Z_i = \log p(Y^*|X^{(i)}$ converge in distribution to a multivariate normal. Now let us compute the moments of Z_i:

$$\begin{aligned}
\mathbf{E}[Z_1] &= d\mathbf{E}[\log b(X_1,Y_1^*) - \log b(X_1)] \\
&= d\mathbf{E}[\log b(Y_1^*) + u(X_1,Y_1^*) - u(X_1,Y_1^*)^2/2] \\
&= -H(Y) - I(X;Y)
\end{aligned}$$

Meanwhile, we know that

$$\mathbf{E}[Z_*] = \mathbf{E}[\log p(Y^*|X^*)] = H(X) - H(X,Y),$$

hence

$$\mathbf{E}[Z_* - Z_i] = 2I(X;Y).$$