Estimating Mutual Information from Average Classification Error

Charles Zheng and Yuval Benjamini

March 17, 2016

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. In high-dimensional settings, non-parametric estimators of mutual information perform poorly, and the only remaining options for estimating mutual information are to either make a parametric assumption (such as multivariate Gaussianity) or obtain a lower bound on the mutual information $I(X;Y)$ from the estimated mutual information between X and an estimator or classification rule, $I(X;\hat{X}(Y))$. But, assuming multivariate Gaussianity reduces mutual information to a linear correlation-based statistic; meanwhile, lower bounds based on $I(X;\hat{X}(Y))$ tend to be overconservative. We propose a new estimator of mutual information based on the concept of "averaged Bayes error," which we prove to be asymptotically a function of the mutual information. The average Bayes error can be estimated using the average misclassification rate from random classification tasks; our estimation procedure consists of plugging in such an estimate of average Bayes error into the inverse of the asymptotic formula. We demonstrate the utility of our method in obtaining accurate estimates of mutual information in simulated data, as well as an fMRI dataset, without having to make parametric assumptions.

1 Introduction

Many of the key questions motivating neuroimaging studies involve some concept of information: which brain region processes sensory information? how is information encoded in the form of memories? To answer whether or not a particular brain region receives information about a given stimulus, it suffices to demonstrate a systematic dependence between brain activity in that region and stimulus features; this was the approach taken for studying single-cell recordings. Moving beyond single-cell recordings, it becomes desirable to use methods which can quantify the information encoded by populations of cells.

Following the seminal work by Haxby (2001), the dominant approach in this area, known as "multivariate pattern analysis" (MVPA), is to quantify information in the form of classification error. The classification approach contrasts with more traditional measures of information, such as mutual information and Fisher information. For example, to demonstrate that a particular brain region responds to a certain type of sensory information, one employs supervised learning to build a classifier which classifies the stimulus class from the brain activation in that region, and tests the statistical hypothesis that the classifier has above-chance classification accuracy. In principle, one could just as well test the statistical hypothesis that the Fisher information or mutual information between the stimulus and the activation patterns is nonzero. But in practice, the machine learning approach enjoys several advantages over approaches based on estimating Fisher information or mutual information. First, one does not need a parameterization of the stimulus space: this is highly relevant for studying complex stimuli such as faces. Secondly, the machine learning approach does not require parametric model assumptions, unlike Fisher information. Thirdly, the machine learning approach scales well with dimensionality of both the stimulus space and the response space. In contrast, nonparametric methods for estimating mutual information scale extremely poorly with dimensionality of the response space. Meanwhile, a major detractor for using classification error as a measure of information is that the classification accuracy depends on the particular choice of stimuli exemplars employed in the study and the number of partitions used to define the classes for the classification task. The difficulty of the classification task depends on the number of classes defined: high classification accuracy can be achieved relatively easily by using a coarse partition of stimuli exemplars into classes. In a meta-analysis on visual decoding, Coutanche

et al (2016) quantified the strength of a classification study using the formula

$$\text{decoding strength} = \frac{\text{accuracy} - \text{chance}}{\text{chance}}.$$

Such an approach may compensate for the differences in accuracy due purely to choice of number of classes defined; however, no theory is provided to justify the formula.

In contrast, mutual information has ideal properties for the quantitatively comparing information between different studies, or between different brain regions, subjects, featurization models, or modalities. Not only is the mutual information defined independently of the arbitrary definition of stimulus classes (albeit still dependent on an implied distribution over stimuli), it is even meaningful to discuss the difference between the mutual information measured for one system and the mutual information for a second system. Due to these appealing properties, as well as the success of information theoretic approaches in more traditional neuroscience studies, mutual information has been frequently proposed for application in neuroimaging studies (Quiroga 2009). However, since nonparametric approaches for estimating mutual information are unusably inaccurate for typical high-dimensional data sets, alternative approaches must be employed. One can tractably estimate mutual information by assuming a multivariate Gaussian model: however, this approach essentially assumes a linear relationship between the input and output, and hence fails to quantify nonlinear dependencies. Alternatively, one can obtain a lower bound on the mutual information from the confusion matrix of the classifier. This is the most popular approach for estimating mutual information in neuroimaging studies, but suffers from known shortcomings (Gastpar 2010, Quiroga 2009).

The idea of linking classification performance to mutual information is not new: after all, Shannon's original motivation was to characterize the minimum achievable error probability of a noisy communication channel. More explicitly, Fano's inequality provides a lower bound on mutual information in relation to the optimal prediction error, or Bayes error. Fano's inequality can be further refined to obtain a tighter lower bound on mutual information (Tebbe and Dwyer 1968.)

In practice, the idea of deriving estimates of mutual information from classification performance occupies an advantageous middle ground between the two extremes of nonparametric and parametric approaches for estimating mutual information. In neuroimaging data, we lack prior knowledge for spec-

ifying parametric models, and the data is too high-dimensional for nonparametric approaches, but we have a sufficient idea of the general "structure" in the data to achieve above-chance classification rates.

However, there are several practical issues for estimating mutual information from classification accuracy. One, the achievable classification accuracy depends on the amount of data available for training. The test error is a unbiased estimate of the true error rate of the classifier on new data (called generalization error); but in turn, the generalization error can only be meaningfully interpreted as a lower bound on the error rate of the ideal classfier, or Bayes error. Secondly, even if one could obtain the Bayes error, the Bayes error itself depends on the choices made by the experimenter in regards to the stimuli exemplar chosen in the experiment, and the decision of how to partition those exemplars in the classification task. For example, Nishimoto et al classified segments of a movie clips based on activation patterns, but the definition of the classification task, and the achievable classification performance, depends not only on the particular move clips used in the experiment, but also the choice of time interval used to define discrete classes: defining each class to be a 1sec segment of movie results in more distinct classes and lower classification accuracy than defining each class to be a 4sec segment of movie. The Bayes error, and any estimate of mutual information based on the Bayes error, would therefore be necessarily dependent on the experimental parameters.

In principle, the first issue can be overcome by having sufficiently many observations: but a practical issue is that an astronomical amount of data might be needed to learn the optimal classification rule. A more efficient approach would be to estimate the Bayes error from the classification performance. Since the Bayes error is the large-sample limit of the achieved classification error, a promising approach is to perform classification using differently sized subsamples of the training data, producing a plot of classification error versus sample size–a "learning curve." One can then extrapolate the learning curve to estimate the Bayes error (Cortes et al. 1994.) However, much work remains to develop rigorous methodology for estimating Bayes error, and so we leave this first issue for future work.

The starting point for our methodology is a proposal for overcoming the second issue: the non-uniqueness of the Bayes error. We define a notion of K-class *average Bayes error* which is uniquely defined for any given stimulus distribution and stochastic mapping from stimulus to response. The K-class average Bayes error is simply the expectation of the Bayes error when

K stimuli exemplars are drawn i.i.d. from the stimulus distribution, and treated as distinct classes. Hence the average Bayes error can in principle be estimated if the appropriate randomization is employed for designing the experiment.

While the K-class average Bayes error is defined independently of the particular choice of stimuli, the quantity still depends on the choice of number of classes, K. In comparison, the mutual information provides a quantification of information which does not depend on K, allowing more flexible comparisons and easier interpretation. Hence our main theoretical contribution is the derivation of an asymptotic relationship between the K-class average Bayes error and the mutual information, which in practice means that any estimate of the K-class Bayes error can be converted into an estimate of mutual information; and the resulting estimator of mutual information should be asymptotically independent of the choice of number of classes K.

2 Average Bayes Error

The following simplified model captures the essence of many neuroimaging studies. Let $\mathcal{X}$ be a space of stimuli, represented by q-dimensional vectors. In the design stage of the experiment, a set of stimuli exemplars $\{x_1, \ldots, x_k\}$ are selected, and then one specifies a T-length sequence of stimuli $(x_{i_1}, \ldots, x_{i_n})$ to be presented to the subject. In the execution of the experiment, an activation pattern, or *response* y_t is obtained for each of the stimuli presentations x_{i_t}. Generally y_t is a p-dimensional vector, representing activity levels of p disjoint brain regions. To simplify, assume that each of the k stimuli is presented a total of r times in the sequence; further assume that the responses to each stimulus presentation are conditionally independent, hence the ordering of the sequence does not matter. Henceforth we can let y_i^j denote the response to the jth presentation of the stimulus x_i.

We now restate the model in application-independent terms. Let $\mathcal{X} \subset \mathbb{R}^q$ and $\mathcal{Y} \subset \mathbb{R}^p$; let $p(x)$ be a probability density on $\mathcal{X}$. For every $x \in \mathcal{X}$, let $p_x(y)$ be a probability density on $\mathcal{Y}$. Define the joint density

$$p(x, y) = p(x)p_x(y)$$

so that $p_x(y)$ can also be written $p(y|x)$. Take a subset $\{x_1, \ldots, x_k\} \subset \mathcal{X}$. Let Y_i^j be a random vector distributed according to density $p_{x_i}(y)$; and let Y_i^j be conditionally independent given $\{x_1, \ldots, x_k\}$ for $i = 1, \ldots, k$ and $j = 1, \ldots, r$.

In MVPA, one carries out a classification task to assess whether y contains information about x. Formally, a classification rule is any (possibly stochastic) mapping $f : \mathcal{Y} \to \{1, \ldots, k\}$. The *generalization error* of the classification rule is

$$e_{gen}(f) = \frac{1}{k} \sum_{i=1}^{k} \Pr[f(Y) \neq i | X = x_i].$$

A trivial classification rule which outputs the result of a k-sided die roll for all inputs y would achieve a generalization error of $e_{gen} = \frac{1}{k}$. Conversely, even a single counterexample with $e_{gen} < \frac{1}{k}$ is indicative that y contains nonzero information about x. Hence, in order to demonstrate that y is informative of x, one tests the null hypothesis

$$H_0 : e_{gen}(f) = \frac{1}{k}.$$

Rejecting the null hypothesis for a given classification rule f can be taken as evidence that y is informative of x. In the discussion thus far, "information" refers to an informal, intuitive notion, but it is also possible to formally establish that violation of the null hypothesis implies nonzero mutual information $I(X; Y)$.

The classifier is a functional which maps a set of observations to a classification rule,

$$\mathcal{F} : \{(x_1, y_1), \ldots, (x_m, y_m)\} \mapsto f(\cdot);$$

informally, a classifier is an algorithm which "learns" a classification rule from data. For a valid test of H_0, either the data-splitting approach or the cross-validation approach can be used. In the data-splitting approach, one creates a *training set* consisting of r_1 repeats per class,

$$\{(x_1, y_1^1), \ldots, (x_1, y_1^{r_1}), \ldots, (x_m, y_m^1), \ldots, (x_m, y_m^{r_1})\}$$

and a *test set* consisting of the remaining $r_2 = r - r_1$ repeats.

$$\{(x_1, y_1^{r_1+1}), \ldots, (x_1, y_1^r), \ldots, (x_m, y_m^{r_1+1}), \ldots, (x_m, y_m^{r_1})\}.$$

In the cross-validation approach, one applies the data-splitting approach multiple times, with different training and test partitions, and aggregates the results. We further discuss the cross-validation approach in section X.

In the data-splitting approach, one obtains the classification rule f by applying the classifier to the training data,

$$f = \mathcal{F}(\{(x_1, y_1^1), \ldots, (x_1, y_1^{r_1}), \ldots, (x_m, y_m^1), \ldots, (x_m, y_m^{r_1})\})$$

The test statistic of interest is the test error, defined as

$$e_{test} = \frac{1}{kr_2} \sum_{i=1}^{k} \sum_{j=r_1+1}^{r} \mathrm{I}(f(y_i^j) \neq i).$$

Due to the conditional independence of the training set and test set, e_{test} is an unbiased estimate of e_{gen}. Hence various approaches can be used to obtain a threshold c_α such that $\Pr[e_{test} < c_\alpha] \leq \alpha$ holds (approximately) under the null hypothesis. To name a few, one can employ permutation tests, tests based on a universal variance bound, or the generalized likelihood ratio test. Regardless of the procedure used to derive the threshold c_α, The hypothesis H_0 is rejected at level α if $e_{test} < c_\alpha$.

While tests of the generalization error suffice to establish the presence of information, the generalization error is less satisfactory as a measure of the information between X and Y, since the e_{gen} depends on the classification rule f obtained, and hence, varies randomly depending on the sampling of the data. For most reasonable classifiers, the generalization error decreases as the number of training observations increases. Therefore, the sample size is an additional confounding factor for interpreting the generalization error.

A more ideal measure of information, still related to the classification, is the Bayes error, which is simply the *optimal* generalization error

$$e_{Bayes} = \min_f e_{gen}(f).$$

Due to Bayes' theorem, the optimal classification rule f^* which achieves the Bayes error can be given explicitly: it is the maximum a posteriori (MAP) rule

$$f^*(y) = \mathrm{argmax}_{i=1}^{k}\ p(y|x_i).$$

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

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

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:

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

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,

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

The MLE classification rule is therefore defined as

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

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.

3 Theory

3.1 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 $(X,Y) \sim F$.

3.2 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.

- Give 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.

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.

3.2.1 Gaussian Example

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 $\rho = (1 + d\sigma^2)^{-\frac{1}{2}}$. Recalling that the mutual information

of the components of such a bivariate normal is $-\log(1-\rho^2)/2$, we easily calculate:

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

Let ι denote the limit of the mutual information as $d \to \infty$: we have

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

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

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

where the first term is a scaled chi-squared distribution with d degrees of freedom: the scale is $-\frac{1}{2}\frac{1+\rho^2}{1-\rho^2}$ for $i = 1, \ldots, K-1$ and $-1/2$ for $i = 0$. The omitted constant is

$$C_\rho = -\frac{1}{2}\log(2\pi(1-\rho)^2)$$

Since we can separate Z_i into independent, componentwise sums,

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

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:

$$\begin{aligned}
\mathbf{E}[Z_*] &= -\frac{d}{2} + C_\rho \\
\mathbf{E}[Z_i] &= -\frac{d}{2}\frac{1+\rho^2}{1-\rho^2} + C_\rho \\
\mathrm{Var}[Z_*] &= \mathrm{Cov}(Z_*, Z_i) = \frac{d}{2} \\
\mathrm{Var}[Z_i] &= \frac{d}{2}\frac{(1+\rho^2)^2}{(1-\rho^2)^2} \\
\mathrm{Cov}[Z_i, Z_j] &= \frac{d}{2}\frac{1}{(1-\rho^2)^2}
\end{aligned}$$

for $i \neq j \neq 0$. Taking limits, the moments simplify to yield

$$\begin{bmatrix} Z_* \\ Z_1 \\ \vdots \\ Z_{K-1} \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} C_0 - \frac{d}{2} \\ C_0 - \frac{d}{2} - \frac{1}{\sigma^2} \\ \vdots \\ C_0 - \frac{d}{2} - \frac{1}{\sigma^2} \end{bmatrix}, \begin{bmatrix} \frac{d}{2} & \frac{d}{2} & \cdots & \frac{d}{2} \\ \frac{d}{2} & \frac{d}{2} + \frac{2}{\sigma^2} & \cdots & \frac{d}{2} + \frac{1}{\sigma^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d}{2} & \frac{d}{2} + \frac{1}{\sigma^2} & \cdots & \frac{d}{2} + \frac{2}{\sigma^2} \end{bmatrix} \right),$$

where $C_0 = -\log(2\pi)/2$. By the central limit theorem, the misclassification probability is

$$\mathrm{MC} = \Pr[Z_* < \max_{i=1}^{K-1} Z_i]$$

for a random multivariate normal vector $(Z_*, Z_1, \ldots, Z_{K-1})$ with the given mean and covariance matrix. It is worth noting that the probability $\Pr[Z_* < \max_{i=1}^{K-1} Z_i]$ directly gives the *averaged* Bayes error: indeed, in high dimensions it is not trivial to compute the Bayes error for fixed configuration. To obtain a simplified expression of this multivariate normal probability, we employ the following lemma.

Lemma. *Suppose $(Z_*, Z_1, \ldots, Z_{K-1})$ are jointly multivariate normal, with $\boldsymbol{E}[Z_* - Z_1] = \alpha$, $Var(Z_*) = \beta$, $Cov(Z_*, Z_i) = \gamma$, $Var(Z_i) = \delta$, and $Cov(Z_i, Z_j) = \epsilon$ for all $i, j = 1, \ldots, K-1$. Then, letting*

$$\mu = \frac{\boldsymbol{E}[Z_* - Z_i]]}{\sqrt{\frac{1}{2}\,Var(Z_i - Z_j)}} = \frac{\alpha}{\sqrt{\delta - \epsilon}},$$

$$\nu^2 = \frac{Cov(Z_* - Z_i, Z_* - Z_j)}{\frac{1}{2}\,Var(Z_i - Z_j)} = \frac{\beta + \epsilon - 2\gamma}{\delta - \epsilon},$$

we have

$$\begin{aligned}\Pr[Z_* < \max_{i=1}^{K-1}] &= \Pr[W < M_{K-1}] \\ &= 1 - \int -\frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{(w-\mu)^2}{2\nu^2}} (1-\Phi(w))^{K-1} dw,\end{aligned}$$

where $W \sim N(\mu, \nu^2)$ *and* M_{K-1} *is the maximum of* $K-1$ *independent standard normal variates, which are independent of* W*.*

(See appendix for proof.)

Applying the lemma, we compute

$$\mu = \frac{\sigma^{-2}}{\sqrt{\sigma^{-2}}} = \frac{1}{\sigma},$$

$$\nu^2 = \frac{\sigma^{-2}}{\sigma^{-2}} = 1,$$

hence

$$\mathrm{MC} = \Pr[N(\frac{1}{\sigma}, 1) < M_{K-1}].$$

Defining the function $f_K(\mu) = \Pr[N(\mu, 1) < M_{K-1}]$, we therefore get

$$\mathrm{MC} = f_K\left(\frac{1}{\sigma}\right) = f_K(\sqrt{2\iota}),$$

recalling that ι is the limiting value of the mutual information.

Hence we obtain a fairly explicit relationship between average Bayes error MC and limiting mutual information in the gaussian case. In the following section, we see that the formula

$$\mathrm{MC} = f_K(\sqrt{2\iota})$$

applies more generally!

3.2.2 Generalization

How far can we generalize the previous example? For starters, we can allow $(X_i^{[d]}, Y_i^{[d]})$ to have a non-Gaussian bivariate distribution, with a density with respect to Lesbegue measure. However, we still require that $(X_i^{[d]}, Y_i^{[d]})$ are i.i.d. for $i = 1, \ldots, d$. We let $b_d(x, y)$ denote the bivariate joint density of

$(X_i^{[d]}, Y_i^{[d]})$, and $b(x)$ and $b(y)$ to denote the marginal distributions of $b_d(x, y)$, which are also assumed to be fixed. The independence allows us to decompose the mutual information

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

Given some additional conditions on the marginal bivariate distributions $b_d(x, y)$, we can conclude joint asymptotic normality of all of the quantities $\log p(X^{(i)}, Y^{(j)}$ for $i, j = 1, \ldots, K$.

Now consider what happens if the total mutual information stays fixed, $I(X^{[d]}.Y^{[d]}) = c$, while the dimensionality increases. Since $I(X_i^{[d]}.Y_i^{[d]}) = c/d$, and since the marginals are fixed, we conclude that the density functions $b_d(x, y)$ are converging to the product $b(x)b(y)$. Now if we define

$$u_d(x, y) = \frac{b_d(x, y)}{b(x)b(y)} - 1,$$

we can say that $u_d(x, y) \to 0$ as $d \to \infty$.

It turns out that in such a limit, the moments of $u_d(X_i^{(j)}, Y_i^{(k)})$ determines many of the information theoretic-quantities of interest. From the definition, we have

$$0 = \mathbf{E}[u(X_i, Y_i)|X_i] = \mathbf{E}[u(X_i, Y_i)|Y_i] = \mathbf{E}[u(X_i^{(j)}, Y_i^{(k)})]$$

for $j, k \in \{1, \ldots, K\}$. Then observe that

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

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

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

Since $I(X_1; Y_1)$ is 'small'–order $O(1/d)$, the function $u(x,y)$ must become 'small' in some sense as well, as d grows. Assume for the moment that we can justifiably replace $\log(1+u(x,y))$ with its second-order Taylor expansion,

$$\log(1+u_d(x,y)) \approx u_d(x,y) - \frac{1}{2}u_d(x,y)^2.$$

Then we get

$$I(X_1; Y_1) \approx \mathbf{E}\left[u_d(X_1,Y_1^*) + \frac{1}{2}u_d(X_1,Y_1)^2 - \frac{1}{2}u_d(X_1,Y_1^*)^3\right],$$

which, since $\mathbf{E}[u_d(X_1,Y_1^*)] = 0$, and since we have been neglecting third-order terms, gives

$$I(X_1; Y_1) \approx \frac{1}{2}\mathbf{E}[u_d(X_1,Y_1^*)^2] = \frac{1}{2}\mathrm{Var}[u_d(X_1,Y_1^*)].$$

Many other similar identities occur in the following proof–all of which depend on neglecting higher-order terms of u_d. But when can we justifiably ignore terms of the form $u_d(X_i, Y_i^*)^k$? Ideally, we need $\mathbf{E}[|u_d(X_i, Y_i^*)|^k]$ uniformly bounded by $O(d^{-1+\epsilon})$ for $k \geq 3$. However, in order to conclude such a result, it is necessary to assume a few regularity conditions. For instance, it suffices to assume that

????

We mention it to give a concrete example of a sufficient condition for being able to neglect higher moments of $u_d(X_i, Y_i^*)$.

Collected together, we will assume the following:

A1. For all d, $I(X^{[d]}, Y^{[d]}) = \iota$.

A2. For all d, d' and $i \leq d$, $j \leq d'$, we have $X_i^{[d]}$ is equal to $X_j^{[d']}$ in distribution, $Y_i^{[d]}$ is equal to $Y_j^{[d']}$ in distribution. Let $b(x)$ denote the marginal distribution of X_j and $b(y)$ denote the marginal distribution of Y_j.

A3. Assume X_j and Y_j have finite third moments.

A4. For each $d = 1, 2, \ldots$, the components $(X_i^{[d]}, Y_i^{[d]})$ are drawn i.i.d. from a bivariate density $b_d(x, y)$ for $i = 1, \ldots, d$.

A6. Defining $u_d(x, y) = \frac{b_d(x,y)}{b(x)b(y)}$, we have $\mathbf{E}[|u_d(X_i, Y_i^*)^3|] = O(d^{-1-\epsilon})$, where $\epsilon > 0$.

Theorem. *Let $X^{[d]}, Y^{[d]}$ be a sequence of distributions satisfying assumptions A1-A6. Then, as $d \to \infty$, the misclassification probability*

$$MC = \Pr[Z_* < \max_{i=1}^{K-1} Z_i]$$

satisfies

$$\lim_{d\to\infty} MC = f_K(\sqrt{2\iota}),$$

where f_K is defined in Lemma.

Proof. As mentioned in the preceding discussion, assumption A6 allows us to find some $\epsilon > 0$, allowing us to write

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

Furthermore, we conclude that

$$\mathrm{Cov}(u_d(X_i^{(j)}, Y_i^*)^2, u_d(X_i^{(k)}, Y_i^*)) = O(d^{-1-\epsilon})$$

$$\mathrm{Cov}(u_d(X_i^{(j)}, Y_i^*), u_d(X_i^{(k)}, Y_i^*)) = O(d^{-1-\epsilon})$$

for all $j, k = 0, \ldots, K-1$. And for $i \neq j \neq 0$, we have

$$\begin{aligned}\mathrm{Cov}(u_d(X_1^{(i)}, Y_1^*), u_d(X_1^{(j)}, Y_1^*)) =& \mathrm{Cov}(\mathbf{E}[u_d(X_1^{(i)}, Y_1^*)|Y_1^*], \mathbf{E}[u_d(X_1^{(j)}, Y_1^*)|Y_1^*]) \\ &+ \mathbf{E}[\mathrm{Cov}(u_d(X_1^{(i)}, Y_1^*), u_d(X_1^{(j)}, Y_1^*))|Y_1^*] \\ &= \mathrm{Cov}(0,0) + \mathbf{E}[0] = 0\end{aligned}$$

due to the identity $\mathbf{E}[u_d(X_1, y)|y] = 0$, and conditional independence, respectively. By a similar argument,

$$\mathrm{Cov}(u_d(X_1^{(i)}, Y_1^*), u_d(X_1^*, Y_1^*)) = 0.$$

Noting that $\mathbf{E}[u_d(X_1, Y_1^*)^2] = 2\iota/d$, we can compute

$$\begin{aligned}\mathbf{E}[u_d(X_1^*, Y_1^*)] &= \int u_d(x,y) b(x,y) dxdy \\ &= \int u_d(x,y)(1 + u_d(x,y)) b(x) b(y) dxdy \\ &= \mathbf{E}[u_d(X_1, Y_1^*)] + \mathbf{E}[u_d(X_1, Y_1^*)^2] \\ &= 0 + 2\iota/d = 2\iota/d\end{aligned}$$

and

$$\begin{aligned}\mathbf{E}[u_d(X_1^*, Y_1^*)^2] &= \int u_d(x,y)^2 (1 + u_d(x,y)) b(x) b(y) dxdy \\ &= \mathbf{E}[u_d(X_1, Y_1^*)^2] + \mathbf{E}[u_d(X_1, Y_1^*)^3] \\ &= 2\iota/d + O(d^{-1-\epsilon}),\end{aligned}$$

hence

$$\mathrm{Var}(u_d(X_1^*, Y_1^*)) = \frac{2\iota}{d} - \frac{4\iota^2}{d^2} + O(d^{-1-\epsilon}) = \frac{2\iota}{d} + O(d^{-1-\epsilon})$$

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:

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

Meanwhile, we know that

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

hence

$$\mathbf{E}[Z_* - Z_i] = 2I(X;Y) + O(d^{-\epsilon}) = 2\iota + O(d^{-\epsilon}).$$

We get

$$\begin{aligned}\mathrm{Var}(Z_i - Z_j) &= d\mathrm{Var}(\log(b_d(Y_1^*|X_1^{(i)})) - \log(b_d(Y_1^*|X_1^{(j)}))) \\ &= d\mathrm{Var}(\log(b_d(X_1^{(i)}, Y_1^*)) - \log(b_d(X_1^{(j)}, Y_1^*) - \log b(X_1^{(i)}) + \log b(X_1^{(j)}))) \\ &= d\mathrm{Var}(u_d(X_1^{(i)}, Y_1^*) - u_d(X_1^{(j)}, Y_1^*) - u_d(X_1^{(i)}, Y_1^*)^2/2 + u_d(X_1^{(j)}, Y_1^*)^2/2) \\ &= d\mathrm{Var}(u_d(X_1^{(i)}, Y_1^*) - u_d(X_1^{(j)}, Y_1^*)) + O(d^{-\epsilon}) \\ &= 2d\mathrm{Var}(u_d(X_1^{(i)}, Y_1^*)) + O(d^{-\epsilon}) = 4\iota + O(d^{-\epsilon}).\end{aligned}$$

Looking ahead to the application of the Lemma, we can already compute

$$\mu = \lim_{d\to\infty} \frac{\mathbf{E}[Z_* - Z_i]}{\sqrt{\frac{1}{2}\mathrm{Var}(Z_i - Z_j)}} = \sqrt{2\iota}.$$

It remains to compute

$$\begin{aligned}\mathrm{Cov}(Z_* - Z_i, Z_* - Z_j) &= d\mathrm{Cov}(u_d(X_1^*, Y_1^*) - u_d(X_1^{(i)}, Y_1^*), u_d(X_1^*, Y_1^*) - u_d(X_1^{(j)}, Y_1^*)) + O(d^{-\epsilon}) \\ &= d\mathrm{Var}(u_d(X_1^*, Y_1^*)) = 2\iota + O(d^{-1-\epsilon})\end{aligned}$$

We conclude that

$$\nu^2 = \lim_{d\to\infty} \frac{\mathrm{Cov}(Z_* - Z_i, Z_* - Z_j)}{\frac{1}{2}\mathrm{Var}(Z - i - Z_j)} = \frac{2\iota}{2\iota} = 1.$$

Hence, the desired result follows from the asymptotic normality of $(Z_*, Z_1, \ldots, Z_{K-1})$ and Lemma. □