
Estimating mutual information in high dimensions via classification error

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Estimating the mutual information $I(X; Y)$ based on observations becomes sta-
2 tistically infeasible in high dimensions without some kind of assumption or prior.
3 One approach is to assume a parametric joint distribution on (X, Y) , but in many
4 applications, such a strong modeling assumption cannot be justified. Alternatively,
5 one can estimate the mutual information based the performance of a classifier
6 trained on the data. Existing methods include using the empirical mutual infor-
7 mation of the confusion matrix of the classifier, as well as an estimator based on
8 Fano’s inequality. However, both of these methods all produce an estimate which
9 is bounded by $\log(k)$, where k is the number of classes. This presents a substantial
10 limitation for classification-based approaches, since the number of repeats per
11 class must be large for the classifier to work well, hence limiting the number of
12 classes k that can be defined. In this paper, we construct a novel classification-
13 based estimator of mutual information which overcomes these limitations. Our
14 estimator is based on high-dimensional asymptotics: we show that in a particular
15 limiting regime, the mutual information is an invertible function of the expected
16 k -class Bayes error. While the theory is based on a large-sample, high-dimensional
17 limit, we demonstrate through simulations that our proposed estimator has superior
18 performance to the alternatives in problems of moderate dimensionality.

1 Introduction

19 **1 Introduction**
20 Mutual information $I(X; Y)$ is fundamentally a measure of dependence between random variables
21 X and Y , and is defined as

$$I(X; Y) = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy.$$

22 In its original context of information theory, the mutual information describes the rate at which a
23 noisy communications channel Y can communicate bits from a source stream X , but by now, the
24 quantity $I(X, Y)$ has found many new uses in science and engineering. Mutual information is used
25 to test for conditional independence [1], to quantifying the information between a random stimulus
26 X and the signaling behavior of an ensembles of neurons, Y [2]; for use as an objective function for
27 training neural networks [3], for feature selection in machine learning, and even as an all-purpose
28 nonlinear measure of “correlation for the 21st century” [4]. What is common to all of these new
29 applications, and what differs from the original setting of Shannon’s theory of information, is that
30 the variables X and Y have unknown distributions which must be inferred from data. In the case
31 when X and Y are both low-dimensional, for instance, when summarizing the properties of a single
32 neuron in response to a single stimulus feature, $I(X; Y)$ can be estimated nonparametrically using a
33 reasonable number of observations. There exists a huge literature on nonparametric estimation of
34 entropy and mutual information, see [5] for a review.

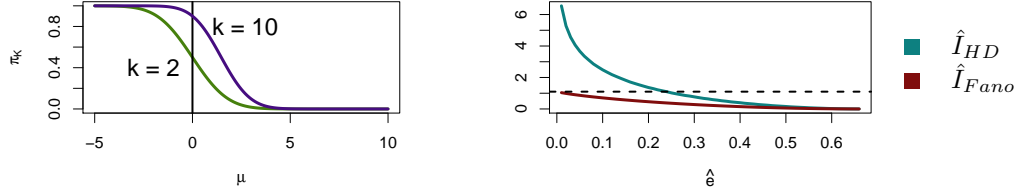


Figure 1: Left: The function $\pi_k(\mu)$ for $k = \{2, 10\}$. Right: \hat{I}_{HD} with \hat{I}_{Fano} as functions of \hat{e}_{gen} , for $k = 3$. While \hat{I}_{Fano} is bounded from above by $\log(k)$ (dotted line), \hat{I}_{HD} is unbounded.

However, the sample complexity for nonparametric estimation grows exponentially with the dimension, rendering such methods ineffective in applications with high-dimensional data [5]. In multivariate pattern analysis [6], the input X could be a natural image parameterized by $p = 10000$ image features, while the output Y is a $q = 20000$ -dimensional vector of brain activation features obtained from the fMRI scan. In problems of such dimensionality, 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.

Rather than assuming a full parametric generative model, one can empirically select a good *discriminative* model by using machine learning. Treves [7] first proposed using the empirical mutual information of the classification matrix in order to obtain a lower bound of the mutual information $I(X; Y)$; this confusion-matrix-based lower bound has subsequently enjoyed widespread use in the MVPA literature [8]. Even earlier than this, the idea of linking classification performance to mutual information can be found in the beginnings of information theory. Fano's inequality provides a lower bound on mutual information in relation to the optimal prediction error, or Bayes error.

We derive a new classification-based estimator of mutual information by exploiting an assumption on the random sampling of the classes (described in section 1.1) and also a universality property that arises in high-dimensions. This universality phenomenon allows us to establish a relationship between the mutual information $I(X; Y)$ and the average k -class Bayes error, $e_{ABE,k}$. In short, we will identify a function π_k (which depends on k),

$$e_{ABE,k} \approx \pi_k(\sqrt{2I(X; Y)}) \quad (1)$$

and that this approximation becomes accurate under a limit where $I(X; Y)$ is small relative to the dimensionality of X , and under the condition that the components of X are approximately independent. The function π_k is given by

$$\pi_k(c) = 1 - \int_{\mathbb{R}} \phi(z - c) \Phi(z)^{k-1} dz.$$

This formula is not new to the information theory literature: it appears as the error rate of an orthogonal constellation [9]. What is surprising is that the same formula can be used to approximate the error rate in much more general class of classification problems¹—this is precisely the universality result which provides the basis for our proposed estimator.

Figure 1 displays the plot of π_k for several values of k . For all values of k , $\pi_k(\mu)$ is monotonically decreasing in μ , and tends to zero as $\mu \rightarrow \infty$, which is what we expect since if $I(X; Y)$ is large, then the average Bayes error should be small. Another intuitive fact is that $\pi_k(0) = 1 - \frac{1}{k}$, since after all, an uninformative response cannot lead to above-chance classification accuracy.

The estimator we propose is

$$\hat{I}_{HD} = \frac{1}{2} (\pi_k^{-1}(\hat{e}_{gen,\alpha}))^2,$$

obtained by inverting the relation (1), then substituting an estimate of generalization error $\hat{e}_{gen,\alpha}$ for the $e_{ABE,k}$. As such, our estimator can be directly compared to the \hat{I}_{Fano} , since both are functions

¹An intuitive explanation for this fact is that points from any high-dimensional distribution lie in an orthogonal configuration with high probability.

68 of $\hat{e}_{gen,\alpha}$ (Figure 1.) As the estimate of generalization error goes to zero, \hat{I}_{Fano} approaches $\log(k)$
69 while \hat{I}_{HD} goes to infinity. This difference in behavior is due to the fact that in contrast to Fano’s
70 inequality, the asymptotic relationship (1) is independent of the number of classes k .

71 In this paper, we argue for the advantages of our method in comparison to alternative discriminative
72 estimators under the assumption that the discriminative model approximates the Bayes rule. While
73 this is an unrealistic assumption, it simplifies the theoretical discussion, and allows us to clearly
74 discuss the principles behind our method.

75 The organization of the paper is as follows. We outline our framework in Section 2.1. In section 2.2
76 we present our key result, which links the asymptotic average Bayes error to the mutual information,
77 under an asymptotic setting intended to capture the notion of high dimensionality². In section 2.3 we
78 apply this result to derive our proposed estimator, \hat{I}_{HD} (where HD stands for “high-dimensional.”)
79 Section 3 presents simulation results, and section 4 concludes. All proofs are given in the supplement.

80 2 Theory

81 2.1 Setting

82 Let us assume that the variables X, Y have a joint distribution F , and that one can define a conditional
83 distribution of Y given X , $Y|X \sim F_X$, and let G denote the marginal distribution of X . We
84 assume that data is collected using *stratified sampling*. For $j = 1, \dots, k$, sample i.i.d. *exemplars*
85 $X^{(1)}, \dots, X^{(k)} \sim G$. For $i = 1, \dots, n$, draw Z^i iid from the uniform distribution on $1, \dots, k$, then
86 draw Y^i from the conditional distribution $F_{X^{(Z^i)}}$.

87 Stratified sampling is commonly seen in controlled experiments, where an experimenter chooses an
88 input X to feed into a black box, which outputs Y . An example from fMRI studies is an experimental
89 design where the subject is presented a stimulus X , and the experimenter measures the subject’s
90 response via the brain activation Y .³

91 When stratified sampling is employed, one can define an *exemplar-based* classification task. One
92 defines the *class function* Z by

$$93 \quad Z : \{X^{(1)}, \dots, X^{(k)}\} \rightarrow \{1, \dots, k\},$$

$$94 \quad Z(X^{(i)}) = i \text{ for } i = 1, \dots, k.$$

94 One defines the generalization error by

$$e_{gen}(f) = \frac{1}{k} \sum_{i=1}^k \Pr[f(Y) \neq Z|X = X^{(i)}]. \quad (2)$$

95 In an exemplar-based classification, there is no need to specify an arbitrary partition on the input
96 space (as is the case in category-based classification), but note that the k classes are *randomly* defined.
97 One consequence is that the Bayes error e_{Bayes} is a random variable: when the sampling produces k
98 similar exemplars, e_{Bayes} will be higher, and when the sampling produces well-separated exemplars
99 e_{Bayes} may be lower. Therefore, it is useful to consider the *average Bayes error*,

$$e_{ABE,k} = \mathbf{E}_{X^{(1)}, \dots, X^{(k)}}[e_{Bayes}], \quad (3)$$

100 where the expectation is taken over the joint distribution of $X^{(1)}, \dots, X^{(k)} \stackrel{iid}{\sim} G$.

101 We use the terminology *classifier* to refer to any algorithm which takes data as input, and produces
102 a classification rule f as output. Mathematically speaking, the classifier is a functional which

²Namely, one where the number of classes is fixed, and where the information $I(X; Y)$ remains fixed, while the dimensionality of the input X and output Y both grow to infinity. We make a number of additional regularity conditions to rule out scenarios where (X, Y) is really less “high-dimensional” than it appears, since most of the variation is captured a low-dimensional manifold.

³Note the asymmetry in our definition of stratified sampling: our convention is to take X to be the variable preceding Y in causal order. Such causal directionality constrains the stratified sampling to have repeated X rather than repeated Y values, but has no consequence for the mutual information $I(X; Y)$, which is a symmetric function.

maps a set of observations to a classification rule, $\mathcal{F} : \{(x^1, y^1), \dots, (x^m, y^m)\} \mapsto f(\cdot)$. The data $(x^1, y^1), \dots, (x^m, y^m)$ used to obtain the classification rule is called *training data*. When the goal is to obtain *inference* about the generalization error e_{gen} of the classification rule f , it becomes necessary to split the data into two independent sets: one set to train the classifier, and one to evaluate the performance. The reason that such a splitting is necessary is because using the same data to test and train a classifier introduces significant bias into the empirical classification error [10]. The classification rule is obtained via $f = \mathcal{F}(S_{train})$, where S_{train} is the training set, and the performance of the classifier is evaluated by predicting the classes of the test set. The results of this test are summarized by a $k \times k$ *confusion matrix* M with $M_{ij} = \sum_{\ell=r_1+1}^r I(f(y^{(i),\ell}) = j)$. The i, j th entry of M counts how many times a output in the i th class was classified to the j th class. The *test error* is the proportion of off-diagonal terms of M , $e_{test} = \frac{1}{kr} \sum_{i \neq j} M_{ij}$, and is an unbiased estimator of e_{gen} . However, in small sampling regimes the quantity e_{test} may be too variable to use as an estimator of e_{gen} . We recommend the use of Bayesian smoothing, defining an α -smoothed estimate $\hat{e}_{gen,\alpha}$ by $\hat{e}_{gen,\alpha} = (1 - \alpha)e_{test} + \alpha \frac{k-1}{k}$, which takes a weighted average of the unbiased estimate e_{test} , and the natural prior of *chance classification*.

We define a discriminative estimator to be a function which maps the misclassification matrix to a positive number, $\hat{I} : \mathbb{N}^{k \times k} \rightarrow \mathbb{R}$. We are aware of the following examples of discriminative estimators: (1) estimators \hat{I}_{Fano} derived from using Fano's inequality, and (2) the empirical information of the confusion matrix, \hat{I}_{CM} , as introduced by Treves [7]. We discuss these estimators in section 3.

2.2 Universality result

We obtain the universality result in two steps. First, we link the average Bayes error to the moments of some statistics Z_i . Secondly, we use Taylor approximation in order to express $I(X; Y)$ in terms of the moments of Z_i . Connecting these two pieces yields the formula (1).

Let us start by rewriting the average Bayes error:

$$e_{ABE,k} = \Pr[p(Y|X_1) \leq \max_{j \neq 1} p(Y|X_j) | X = X_1].$$

Defining the statistic $Z_i = \log p(Y|X_i) - \log p(Y|X_1)$, where $Y \sim p(y|X_1)$, we obtain $e_{ABE} = \Pr[\max_{j \geq 1} Z_i > 0]$. The key assumption we need is that Z_2, \dots, Z_k are asymptotically multivariate normal. If so, the following lemma allows us to obtain a formula for the misclassification rate.

Lemma 1. *Suppose (Z_1, Z_2, \dots, Z_k) are jointly multivariate normal, with $\mathbf{E}[Z_1 - Z_i] = \alpha$, $\text{Var}(Z_1) = \beta \geq 0$, $\text{Cov}(Z_1, Z_i) = \gamma$, $\text{Var}(Z_i) = \delta$, and $\text{Cov}(Z_i, Z_j) = \epsilon$ for all $i, j = 2, \dots, k$, such that $\beta + \epsilon - 2\gamma > 0$. Then, letting*

$$\mu = \frac{\mathbf{E}[Z_1 - Z_i]}{\sqrt{\frac{1}{2}\text{Var}(Z_i - Z_j)}} = \frac{\alpha}{\sqrt{\delta - \epsilon}},$$

$$\nu^2 = \frac{\text{Cov}(Z_1 - Z_i, Z_1 - Z_j)}{\frac{1}{2}\text{Var}(Z_i - Z_j)} = \frac{\beta + \epsilon - 2\gamma}{\delta - \epsilon},$$

we have

$$\begin{aligned} \Pr[Z_1 < \max_{i=2}^k Z_i] &= \Pr[W < M_{k-1}] \\ &= 1 - \int \frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{(w-\mu)^2}{2\nu^2}} \Phi(w)^{k-1} dw, \end{aligned}$$

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 .

To see why the assumption that Z_2, \dots, Z_k are multivariate normal might be justified, suppose that X and Y have the same dimensionality d , and that joint density factorizes as

$$p(x^{(j)}, y) = \prod_{i=1}^d p_i(x_i^{(j)}, y_i)$$

139 where $x_i^{(j)}$, y_i are the i th scalar components of the vectors $x^{(j)}$ and y . Then,

$$Z_i = \sum_{m=1}^d \log p_m(y_m | x_m^{(i)}) - \log p_m(y_m | x_1^{(m)})$$

140 where $x_{i,j}$ is the i th component of x_j . The d terms $\log p_m(y_m | x_{m,i}) - \log p_m(y_m | x_{m,1})$ are
 141 independent across the indices m , but dependent between the $i = 1, \dots, k$. Therefore, the multivariate
 142 central limit theorem can be applied to conclude that the vector (Z_2, \dots, Z_k) can be scaled to converge
 143 to a multivariate normal distribution. While the componentwise independence condition is not a
 144 realistic assumption, the key property of multivariate normality of (Z_2, \dots, Z_k) holds under more
 145 general conditions, and appears reasonable in practice.

146 It remains to link the moments of Z_i to $I(X; Y)$. This is accomplished by approximating the
 147 logarithmic term by the Taylor expansion

$$\log \frac{p(x, y)}{p(x)p(y)} \approx \frac{p(x, y) - p(x)p(y)}{p(x)p(y)} - \left(\frac{p(x, y) - p(x)p(y)}{p(x)p(y)} \right)^2 + \dots$$

148 A number of assumptions are needed to ensure that needed approximations are sufficiently accurate;
 149 and additionally, in order to apply the central limit theorem, we need to consider a *limiting sequence*
 150 of problems with increasing dimensionality. We now state the theorem.

151 **Theorem 1.** Let $p^{[d]}(x, y)$ be a sequence of joint densities for $d = 1, 2, \dots$. Further assume that

152 A1. $\lim_{d \rightarrow \infty} I(X^{[d]}; Y^{[d]}) = \iota < \infty$.

153 A2. There exists a sequence of scaling constants $a_{ij}^{[d]}$ and $b_{ij}^{[d]}$ such that the random vector
 154 $(a_{ij} \ell_{ij}^{[d]} + b_{ij}^{[d]})_{i,j=1,\dots,k}$ converges in distribution to a multivariate normal distribution,
 155 where $\ell_{ij} = \log p(y^{(i)} | x^{(i)})$ for independent $y^{(i)} \sim p(y | x^{(i)})$.

156 A3. Define

$$u^{[d]}(x, y) = \log p^{[d]}(x, y) - \log p^{[d]}(x) - \log p^{[d]}(y).$$

157 There exists a sequence of scaling constants $a^{[d]}$, $b^{[d]}$ such that

$$a^{[d]} u^{[d]}(X^{(1)}, Y^{(2)}) + b^{[d]}$$

158 converges in distribution to a univariate normal distribution.

159 A4. For all $i \neq k$,

$$\lim_{d \rightarrow \infty} \text{Cov}[u^{[d]}(X^{(i)}, Y^{(j)}), u^{[d]}(X^{(k)}, Y^{(j)})] = 0.$$

160 Then for $e_{ABE,k}$ as defined above, we have

$$\lim_{d \rightarrow \infty} e_{ABE,k} = \pi_k(\sqrt{2\iota})$$

161 where

$$\pi_k(c) = 1 - \int_{\mathbb{R}} \phi(z - c) \Phi(z)^{k-1} dz$$

162 where ϕ and Φ are the standard normal density function and cumulative distribution function,
 163 respectively.

164 Assumptions A1-A4 are satisfied in a variety of natural models. One example is a multivariate
 165 Gaussian sequence model where $X \sim N(0, \Sigma_d)$ and $Y = X + E$ with $E \sim N(0, \Sigma_e)$, where Σ_d and
 166 Σ_e are $d \times d$ covariance matrices, and where X and E are independent. Then, if $d\Sigma_d$ and Σ_e have
 167 limiting spectra H and G respectively, the joint densities $p(x, y)$ for $d = 1, \dots$, satisfy assumptions
 168 A1 - A4. Another example is the multivariate logistic model, which we describe in section 3. We
 169 further discuss the rationale behind A1-A4 in the supplement, along with the detailed proof.

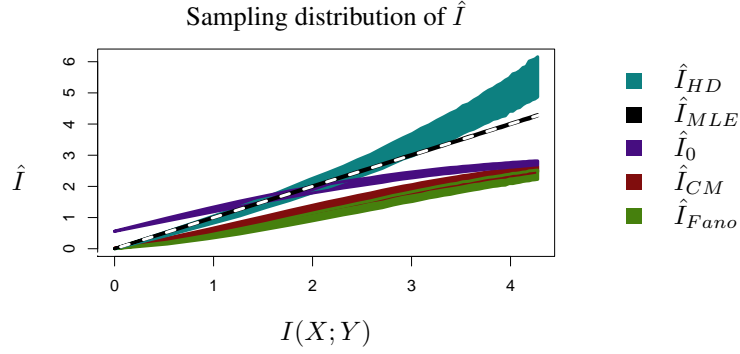


Figure 2: Sampling distributions of \hat{I} for data generated from the multiple-response logistic model. $p = q = 10$; $k = 20$; $B = sI_{10}$, where $s \in [0, \sqrt{200}]$; and $r = 1000$.

2.3 High-dimensional estimator

As stated in the introduction, we propose the estimator

$$\hat{I}_{HD}(M) = \frac{1}{2}(\pi_k^{-1}(\hat{e}_{gen,\alpha}))^2.$$

For sufficiently high-dimensional problems, \hat{I}_{HD} can accurately recover $I(X; Y) > \log k$, supposing also that the classifier \mathcal{F} consistently estimates the Bayes rule. The number of observations needed depends on the convergence rate of \mathcal{F} and also the complexity of estimating $e_{gen,\alpha}$. Therefore, without making assumptions on \mathcal{F} , the sample complexity is at least exponential in $I(X; Y)$. This is because when $I(X; Y)$ is large relative to $\log(k)$, the Bayes error $e_{ABE,k}$ is exponentially small. Hence $O(1/e_{ABE,k})$ observations in the test set are needed to recover $e_{ABE,k}$ to sufficient precision. While the sample complexity exponential in $I(X; Y)$ is by no means ideal, by comparison, the nonparametric estimation approaches have a complexity exponential in the dimensionality. Hence, \hat{I}_{HD} is favored over nonparametric approaches in settings with high dimensionality and low signal-to-noise ratio.

3 Simulation

We compare the discriminative estimators \hat{I}_{CM} , \hat{I}_{Fano} , \hat{I}_{HD} with a nonparametric estimator \hat{I}_0 in the following simulation, and the correctly specified parametric estimator \hat{I}_{MLE} . We generate data according to a multiple-response logistic regression model, where $X \sim N(0, I_p)$, and Y is a binary vector with conditional distribution

$$Y_i | X = x \sim \text{Bernoulli}(x^T B_i)$$

where B is a $p \times q$ matrix. One application of this model might be modeling neural spike count data Y arising in response to environmental stimuli X [12]. We choose the naive Bayes for the classifier \mathcal{F} : it is consistent for estimating the Bayes rule.

The estimator \hat{I}_{Fano} is based on Fano's inequality, which reads

$$H(Z|Y) \leq H(e_{Bayes}) + e_{Bayes} \log ||Z| - 1|$$

where $H(e)$ is the entropy of a Bernoulli random variable with probability e . Replacing $H(Z|Y)$ with $H(X|Y)$ and replacing e_{Bayes} with $\hat{e}_{gen,\alpha}$, we get the estimator

$$\hat{I}_{Fano}(M) = \log(K) - \hat{e}_{gen,\alpha} \log(K - 1) + \hat{e}_{gen,\alpha} \log(p) + (1 - \hat{e}_{gen,\alpha}) \log(1 - \hat{e}_{gen,\alpha}).$$

Meanwhile, the confusion matrix estimator computes

$$\hat{I}_{CM}(M) = \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k \log \frac{M_{ij}}{r/k},$$

194 which is the empirical mutual information of the discrete joint distribution $(Z, f(Y))$.

195 It is known that \hat{I}_{CM} , \hat{I}_0 tend to underestimate the mutual information. Quiroga et al. [8] discussed
 196 two sources of ‘information loss’ which lead to \hat{I}_{CM} underestimating the mutual information: the
 197 discretization of the classes, and the error in approximating the Bayes rule. Meanwhile, Gastpar et al.
 198 [11] showed that \hat{I}_0 is biased downwards due to undersampling of the exemplars: to counteract this
 199 bias, they introduce the anthropic correction estimator \hat{I}_α ⁴.

200 In addition to the sources of information loss discussed by Quiroga et al., an additional reason why
 201 \hat{I}_{CM} and \hat{I}_{Fano} underestimate the mutual information is that they are upper bounded by $\log(k)$, where
 202 k is the number of classes. As $I(X; Y)$ exceeds $\log(k)$, the estimate \hat{I} can no longer approximate
 203 $I(X; Y)$, even up to a constant factor. In contrast, \hat{I}_{HD} is unbounded and may either underestimate
 204 or overestimate the mutual information in general, but performs well when the high-dimensionality
 205 assumption is met.

206 In Figure 2 we show the sampling distributions of the five estimators as $I(X; Y)$ is varied in the
 207 interval $[0, 4]$. The estimator \hat{I}_{MLE} is a plug-in estimator using \hat{B} , the coefficient matrix estimated
 208 via multinomial regression of Y on X ; it recovers the true mutual information within $\pm 2\%$ with a
 209 probability of 90%. We see that \hat{I}_{CM} , \hat{I}_{Fano} , and \hat{I}_0 indeed begin to asymptote as they approach
 210 $\log(k) = 2.995$. In contrast, \hat{I}_{HD} remains a good approximation of $I(X; Y)$ within the range,
 211 although it begins to overestimate at the right endpoint. The reason why \hat{I}_{HD} loses accuracy as
 212 the true information $I(X; Y)$ increases is that the multivariate normality approximation used to
 213 derive the estimator becomes less accurate when the conditional distribution $p(y|x)$ becomes highly
 214 concentrated.

215 4 Discussion

216 Discriminative estimators of mutual information have the potential to estimate mutual information
 217 in high-dimensional data without resorting to fully parametric assumptions. However, a number of
 218 practical considerations also limit their usage. First, one has to find a good classifier \mathcal{F} for the data:
 219 techniques for model selection can be used to choose \mathcal{F} from a large library of methods. However,
 220 there is no way to guarantee how well the chosen classifier approximates the optimal classification
 221 rule. Secondly, one has to estimate the generalization error from test data: the complexity of
 222 estimating e_{gen} could become the bottleneck when e_{gen} is close to 0. Thirdly, for previous estimators
 223 \hat{I}_{Fano} and \hat{I}_{CM} , the ability of the estimator to distinguish high values of $I(X; Y)$ is limited by
 224 the number of classes k . Our estimator \hat{I}_{HD} is subject to the first two limitations, along with any
 225 conceivable discriminative estimator, but overcomes the third limitation under the assumption of
 226 stratified sampling and high dimensionality.

227 It can be seen that additional assumptions are indeed needed to overcome the third limitation, the
 228 $\log(k)$ upper bound. Consider the following worst-case example: let X and Y have joint density
 229 $p(x, y) = \frac{1}{k} I([kx] = [ky])$ on the unit square. Under partition-based classification, if we set
 230 $Z(x) = [kx] + 1$, then no errors are made under the Bayes rule. We therefore have a joint
 231 distribution which maximizes any reasonable discriminative estimator but has *finite* information
 232 $I(X; Y) = \log(k)$. The consequence of this is that under partition-based classification, we cannot
 233 hope to distinguish distributions with $I(X; Y) > \log(k)$. The situation is more promising if we
 234 specialize to stratified sampling: in the same example, a Bayes of zero is no longer likely due to the
 235 possibility of exemplars being sampled from the same bin (‘collisions’)—we obtain an approximation
 236 to the average Bayes error through a Poisson sampling model: $e_{ABE, k} \approx \frac{1}{e} \sum_{j=1}^{\infty} \frac{1}{j(j!)} = 0.484$. By
 237 specializing further to the high-dimensional regime, we obtain even tighter control on the relation
 238 between Bayes error and mutual information. Our estimator therefore provides more accurate
 239 estimation at the cost of more additional assumptions, but just how restrictive are these assumptions?

240 The assumption of stratified sampling is usually not met in the most common applications of
 241 classification where the classes are defined *a priori*. For instance, if the classes consist of three
 242 different species of iris, it does not seem appropriate to model the three species as i.i.d. draws from

⁴However, without a principled approach to choose the parameter $\alpha \in (0, 1]$, \hat{I}_α could still vastly underesti-
 mate or overestimate the mutual information.

some distribution on a space of infinitely many potential iris species. Yet, when the classes have been pre-defined in an arbitrary manner, the mutual information between a latent class-defining variable X and Y may be only weakly related to the classification accuracy. We rely on the stratified sampling assumption to obtain the necessary control on how the classes in the classification task are defined. Fortunately, in many applications where one is interested in estimating $I(X; Y)$, a stratified sampling design can be practically implemented.

The assumption of high dimensionality is not easy to check: having a high-dimension response Y does not suffice, since even then Y could still lie close to a low-dimensional manifold. In such cases, \hat{I}_{HD} could either overestimate or underestimate the mutual information. In situations where (X, Y) lie on a manifold, one could effectively estimate mutual information by would be to combining dimensionality reduction with nonparametric information estimation [13]. We suggest the following diagnostic to determine if our method is appropriate: subsample within the classes collected and check that \hat{I}_{HD} does not systematically increase or decrease with the number of classes k .

The assumption of approximating the Bayes rule is impractical to check, as any nonparametric estimate of the Bayes error requires exponentially many observations. Hence, while the present paper studies the ‘best-case’ scenario where the model is well-specified, it is even more important to understand the robustness of our method in the more realistic case where the model is misspecified. We leave this question to future work.

Even given a classifier which consistently estimates the Bayes error, the estimator \hat{I}_{HD} can still be improved. One can employ more sophisticated methods to estimate $e_{ABE,k}$: for example, extrapolating from learning curves [14]. Furthermore, depending on the risk function, one may debias or shrink the estimate \hat{I}_{HD} to achieve a more favorable bias-variance tradeoff.

All of the necessary assumptions are met in our simulation experiment, hence our proposed estimator is seen to dramatically outperform existing estimators. It remains to assess the utility of our estimation procedure in a real-world example, where both the high-dimensional assumption and the model specification assumption are likely to be violated. In a forthcoming work, we apply our framework to evaluate visual encoding models in human fMRI data.

References

- [1] De Campos, Luis M. (2006). “A scoring function for learning Bayesian networks based on mutual information and conditional independence tests.” *The Journal of Machine Learning Research* 7 : 2149-2187.
- [2] Borst, A. & Theunissen, F. E. (1999). “Information theory and neural coding” *Nature Neurosci.*, vol. 2, pp. 947-957.
- [3] Linsker, R. (1989). “An application of the principle of maximum information preservation to linear systems.” *Advances in neural information processing systems*.
- [4] Speed, T. “A correlation for the 21st century.” (2011). *Science* 334.6062: 1502-1503.
- [5] Beirlant, J., Dudewicz, E. J., Gy orfi, L., & der Meulen, E. C. (1997). “Nonparametric Entropy Estimation: An Overview.” *International Journal of Mathematical and Statistical Sciences*, 6, 17-40.
- [6] Haxby, James V., et al. (2001). “Distributed and overlapping representations of faces and objects in ventral temporal cortex.” *Science* 293.5539: 2425-2430.
- [7] Treves, A. (1997). “On the perceptual structure of face space.” *Bio Systems*, 40(1-2), 189-96.
- [8] Quiroga, Q. R., & Panzeri, S. (2009). Extracting information from neuronal populations: information theory and decoding approaches. *Nature Reviews. Neuroscience*, 10(3), 173-185.
- [9] Tse, D., & Viswanath, P. (2005). *Fundamentals of wireless communication*. Cambridge university press,
- [10] Friedman, J., Hastie, T., & Tibshirani, R. (2008). *The elements of statistical learning*. Vol. 1. Springer, Berlin: Springer series in statistics.
- [11] Gastpar, M. Gill, P. Huth, A. & Theunissen, F. (2010). “Anthropic Correction of Information Estimates and Its Application to Neural Coding.” *IEEE Trans. Info. Theory*, Vol 56 No 2.
- [12] Banerjee, A., Dean, H. L., & Pesaran, B. (2011). “Parametric models to relate spike train and LFP dynamics with neural information processing.” *Frontiers in computational neuroscience* 6: 51-51.

- 292 [13] Theunissen, F. E. & Miller, J.P. (1991). "Representation of sensory information in the cricket cercal sensory
293 system. II. information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary
294 interneurons," *J. Neurophysiol.*, vol. 66, no. 5, pp. 1690-1703.
- 295 [14] Cortes, C., et al. "Learning curves: Asymptotic values and rate of convergence." (1994). *Advances in*
296 *Neural Information Processing Systems*.