Quantifying the precision of decoding models for high-dimensional stimuli

Charles Zheng and Yuval Benjamini

January 2, 2017

Abstract

The analysis of encoding and decoding models is a common theme in both cell recording studies and in neuroimaging. A basic measure of the precision of a decoder is its accuracy at distinguishing k different stimuli. However, the fixed-k accuracy becomes insensitive beyond limited range of precision: low-precision decoders saturate at the chance accuracy $1/k$, while high-precision decoders saturate near perfect accuracy. On the other hand, the entire curve of accuracies for $k = 2, 3, ...$ provides a detailed and interpretable characterization of decoder performance. However, due to limited sampling, usually only a portion of the curve can be estimated: furthermore, it is unclear how to summarize the information in the curve by a single statistic. We show that under a high-dimensional limits, the mutual information becomes a sufficient statistic for reconstructing the entire accuracy curve, therefore suggesting the adoption of the mutual information as measure of decoder precision. Based on our theory, we develop a novel estimator of mutual information suited for high-dimensional settings (such as those found in neuroimaging), and also a procedure for extrapolating the accuracy curve to arbitrarily many stimuli.

1 Introduction

Both computational and cognitive neuroscience are concerned with understanding brain function: while computational neuroscience is concerned with understanding functionality at the level of the spiking behavior of individual

neurons and small neural populations, cognitive neuroscience tends to emphasize functionality at the level of macroscale regions of the interest in the brain. While the recording technologies, motivating questions, and analytical methodologies differ between the two subdisciplines, the conceptualization of brain functionality in terms of *encoding* and *decoding* models has been widely applied in both areas [33][22]. In computational neuroscience, cell recording experiments are conducted to determine whether spike trains have a temporal and/or correlational code [24][12], to examine how the neural code adapts to changes in stimulus distribution [7] and whether downstream neurons make use of higher-order correlations for decoding [27]. Meanwhile, in neuroimaging studies, functional MRI experiments are employed to model the receptive fields of early visual areas in the human brain [17], to examine the semantic encoding of words [21] or objects [14].

The dual perspectives of encoding and decoding originate naturally from the fact that in examining the link between brain activity and function, one can either start with brain activity on one end, or with external stimulation or behavioral observation on the other end. Starting by exposing the subject to sensory stimuli or prompting the subject to engage in particular motor tasks, one can search for areas in the brain which respond to the task: in other words, one can test to see which areas of the brain *encode* the given stimulus. In the other direction, one seeks to understand the functionality of a given brain region: in other words, how to *decode* brain activity in that region.

Formulation of encoding models is relatively straightforward, since one needs only to characterize the observed brain response to a given stimulus. One can further ask how to distinguish between signal and noise in the encoding mechanism [24], or in complex stimuli, seek a linearizing feature set which reveals the nature of the brain representation [22]. However, the establishment of complete decoding models is much less amenable to experiemental manipulation, since to exhaustively characterize the functionality of a neuron, one would have to know in advance the type of information it encodes. Early advances in decoding often depended on strokes of luck: Hubel [13] originally discovered the existence of neurons with orientation-sensitive receptive fields due to the vigorous response of a cell to the perfectly angled shadow of a glass slide that they were inserting into the ophthalmoscope. Yet, even now, the goal of completely characterizing the function of a given brain region remains a difficult task, with the most promising approach being a *reverse inference* procedure [32] which aggregates information from the

literature about activity-functionality relationships.

A more feasible goal is to establish the *precision* with which a neuron can decode a particular type of feature. This can be accomplished by first training an encoding model, and then inverting the encoding model using Bayes' rule to obtain a decoding model [28][33][22].

By decoding *precision*, we mean the specificity which we can identify or reconstruct the stimulus based on the neural response. As such, in our view, the term decoder *precision* is more or less synonymous with terms such as decoder *performance* or decoder *accuracy* as they are used in the literature. However, we choose the word *precision* in particular, because it communicates the idea that the essential quality of a good decoder is that it allows one to confidently and precisely infer the stimulus.

Measures of decoding precision can be used to support several different kinds of scientific inferences. When there exist multiple plausible encoding models–for instance, a model where stimulus information is encoded solely by average firing rate versus a model where inter-spike timings also carry information–the precision of the decoder can be used as a basis for deciding the best encoding model. For two encoding models with equal complexity, such as comparing two different types of receptive field models, the model with better decoding precision could be considered the more plausible model. In the case where a more complex encoding model is compared to a strictly simpler model–such as comparing a model with a temporal code versus a model only incorporating average firing rate, a substantial improvement in decoding precision for the more complex model is needed to demonstrate its validity, since in the null hypothesis where the simpler model is correct, the more complex model should still have approximately equal decoding performance.

Yet another application of decoding precision is to track the adaptivity of the neural code. Fairhall [7] recorded the output of a motion-sensitive neuron in a fly in response to a visual stimulus with changing angular velocity. Changing the variance of the stimulus results in rapid adaptation: the neural code starts adapting to the change in stimulus distribution within tens of milliseconds, which is reflected by an increased or decreased precision (as measured by mutual information) in resolving angular velocity to match the variance of the stimulus. More generally, comparisons of decoding precisions between different conditions can show how the encoded information increases or decreases across experimental conditions. Kayser [18] demonstrated how the mutual information between a sound stimulus and neurons in the auditory

cortex increased when the subjects were also presented a matching visual stimulus (e.g. showing a picture of a lion roaring while playing the sound of a lion's roar.)

Differing types and parameterizations of stimuli naturally lead to differing measures of decoding precision. For stimuli which can be parameterized by a scalar x, the precision can be measured by the squared correlation coefficient R^2 [1]. However, the resulting measure of precision is not invariant to scaling of the parameterization: for instance, the choice of whether to parameterize volume on an absolute scale or a logarithmic scale. The mutual information [34] between the stimulus and the predicted stimulus is invariant to the parameterization of the stimulus. Due to its invariance and a number of other properties, the mutual information is widely used to measure the precision of the neural code in cell recording studies, both for single-neuron decoding models [2] and for population coding models [33][16].

However, the difficulties of estimating mutual information in small samples has been widely recognized, with a large literature on bias correction methods [30][29]. Methods for bias correction have been developed for three different sample size regimes: the moderate-sample regime, where the number of observations is larger than the number of stimulus-response pairs[20][35][36], the undersampled regime, where the number of observations is less than the number of stimulus-response pairs [25], and a *stimulus-undersampled* regime, where only a small fraction of possible stimuli are sampled, but with a large number of observations for each of the sampled stimuli [9]. Nevertheless, even the bias-corrected estimates may be unusably inaccurate in problems of moderate dimensionality, since the cardinality of response space grows exponentially with the dimensionality. In such cases, alternative approaches for estimating the mutual information include the assumption of a parametric model [3][9][37], or usage of the maximum entropy principle to obtain bounds on the mutual information subject to the empirical moments of a certain order [15][10].

Perhaps due to the technical difficulties of estimating mutual information in high dimensions, mutual information has never, to our knowledge, been used as a measure of decoding precision in neuroimaging studies, although it has been proposed for the purpose of bypassing the modelling of the hemodynamic response function for single-voxel analyses [8]. Instead, a variety of methods are employed to characterize the precision of decoding models, depending on the nature of the stimulus and the experimental setup.

In task fMRI experiments where stimuli are drawn from a number of

disjoint semantic categories– for instance, 'birds', 'insects', and 'mammals' as in [4], it is natural to construct a decoder which outputs the predicted category of a stimulus as a function of the response. Such a decoder is known as a *classifier* in the machine learning literature [11], and a natural measure of classifier precision is the probability that the decoder outputs the correct category on a new, randomly drawn test example, which is the *classification accuracy*.

In experiments where the subject is presented a number of parameterized stimuli are drawn from a continuous distribution (such as natural images or sounds), there are two types of decoders which can be constructed. In the first case, one constructs a decoder which estimates the parameters of the stimulus which we call a *reconstructor*: the precision of such a decoder is measured by the correlation between the estimated and true parameter vector [31] [26][23]. In the second case, one constructs a decoder which picks the most likely stimulus from a finite library of examples *which includes the true stimulus* [17][21]. Since the true stimulus is included in the library, the task is to 'identify' the correct stimulus from the library. A natural measure of decoder performance is therefore the probability of correct identification. However, note that this probability is dependent on the arbitrary choice of the size of the exemplar library: a different choice for library size therefore results in a different measure of precision. We refer to the probability of correct classification for a library of k exemplars as the *k-example identification accuracy*.

In their respective domains, these different measures of precision suffice to make inferences on many interesting scientific questions: to list a few examples, showing the superiority of a Gabor filters versus center-surround filters for modeling the receptive fields of V1 and V2 neurons [17], or demonstrating that brain activity in response to viewing an Engish noun can be predicted from word association frequencies [21].

A commonality to all applications of decoding models in neuroimaging is the pairwise comparison of two decoding models (Gabor vs. retinotopic) or the comparison of a single decoding model to chance accuracy. Looking ahead to anticipate what kinds of analyses might be employed in the future based on neuroimaging data, it is suggestive to note that the earliest decoding studies in the cell recording literature also involved comparisons between two or three different decoders [6]. However, as neuroscientists began to consider questions of population coding, analyses of the redundancy between neurons started to make use of comparisons between large numbers of decoders: for a

population of N neurons, one might compare the precision of a decoder (mutual information) based on the entire ensemble, compared to the precisions of decoders based on each of the N individual neurons. Furthermore, one can make the same comparison for a range of different ensemble sizes N. As questions about the redundancy of the neural code are relevant on both the micro scale (the domain of cell recording studies) and the macro scale (the domain of neuroimaging), it is safe to assume that similar analyses, requiring comparisons of large numbers of decoders, will emerge in neuroimaging studies. Already in the functional MRI literature, we see similar decompositions of decoding accuracy versus ensemble size [17], but another possible type of decomposition would be to compare decoding performance as the number of stimulus features is varied, rather than the number of voxels.

The scaling properties of mutual information are highly advantageous when comparing multiple decoders, which could potentially span a wide range of decoding precision: for instance, a single neuron versus an ensemble of thousands of neurons. In contrast, classification accuracy, k-class identification accuracy and reconstruction accuracy all suffer from the issue of *limited dynamic range*: that is, they are only effective at measuring precision within a certain range.

Let us illustrate with the example of identification accuracy. A low precision decoder, such as a decoder based on a single voxel, may have an accuracy which is so close to chance accuracy, $1/k$, as to be statistically indistinguishable from chance based on the data. On the other hand, a sufficiently high-precision decoder may face the opposite problem, where it achieves perfect classification on the limited number of test examples. Any empirical estimate of identification accuracy can only be used to accurately rank decoders which have accuracies sufficiently bounded away from both $1/k$ and 1. The same issue applies to reconstruction accuracy (bounded between 0 and 1) and classification accuracy (bounded between $1/k$ and 1, where k is the number of classes): any bounded measure of precision is ineffective at comparing decoders which are too close to either the upper bound or lower bound of achievable precision.

In practice, the solution to this issue is to find a measure of precision which is well-suited for all of the decoders that needed to be compared. If there are two encoding models which both achieve perfect classification on the test set, then perhaps the more demanding measure of reconstruction accuracy can be used to distinguish them. However, this strategy begins to become impractical as the number of decoders to be compared increases.

One wishes to relate the decoding precision of an N-voxel ensemble for N spanning from 1 to 10000: however, any bounded measure of precision which is suitably stringent for distinguishing $N = 9999$ from $N = 10000$ would fail for comparing $N = 1$ to $N = 2$, and vice-versa.

We have seen that one solution to this predicament is to use an unbounded measure of precision which can remain sensitive to variations in precision across a large dynamic range: for instance, the mutual information. Yet, given the difficulty of estimating the mutual information in high-dimensional settings, one might consider another approach: to develop a systematic means for comparing decoders by using multiple (easily estimated) precision measures, each of which may only capture a limited range of precisions, but which collectively span a sufficiently large range of precisions to include all of the decoders being compared.

Our contribution in this paper is to show that both of these approaches– the estimation of mutual information, and the comparison of decoders based on a range of decoding metrics, turn out to be the very same problem in high-dimensional settings. The *identification accuracy curve*, which we define as the collection of all k-class identification accuracies for $k \geq 2$, can be used to compare a collection of decoders over a large span of precisions. Yet, a recent theoretical result[38] shows that the identification accuracy curve for the Bayes decoder (the optimal decoder) is determined by the mutual information in a certain high-dimensional regime. While it is generally not feasible to approximate the Bayes decoder in high-dimensional settings, we use this result to define the *implied information* for a non-Bayes (suboptimal) decoder. The implied information, $I_{implied}$, is not the true mutual information between the stimulus and response, but it provides a means of comparing two accuracy curves (estimate the implied information from each, and then compare the estimates), as well as providing an unbounded measure of decoding precision which, similar to mutual information, has desirable scaling properties for the purpose of comparing decoders spanning a range of precisions.

2 Methods

2.1 Experimental design

We consider experiments in which a single subject is presented with a sequence of T stimuli: each stimulus is presented during a 'task window' of a fixed duration. The stimuli are represented by real-valued feature vectors $\vec{X}$; let p be the dimensionality of the feature space. The brain activity of the subject is recorded, yielding a q-dimensional vector $\vec{Y}$: in practice, $\vec{Y}$ could consist of discretized time series data or mean firing rates for spike-sorted neurons, or BOLD response for voxels, depending on the recording modality. Let $\vec{X}^{(t)}$ denote the feature vector of the stimulus, and let $\vec{Y}^{(t)}$ denote the vector of intensities (e.g. BOLD response, mean spike) for the tth task window in the sequence.

2.2 Data splitting

The T stimulus-response pairs $(\vec{X}, \vec{Y})$ are randomly partitioned into a *training set* of size N and a *test set* of size $M = T - N$. Form the $N \times p$ data matrix $\boldsymbol{X}^{tr}$ by stacking the features of the N training set stimuli as row vectors, and stack the corresponding responses as row vectors to form the $N \times q$ matrix $\boldsymbol{Y}^{tr}$. Similarly, define $\boldsymbol{X}^{te}$ as the $N \times p$ matrix of test stimuli and $\boldsymbol{Y}^{te}$ as the $N \times q$ matrix of corresponding test responses.

2.3 Probabilistic encoding model

The data is used to estimate a stimulus-based encoding model [17][22][21]. The conditional mean response $\mathbf{E}[\boldsymbol{Y}|\boldsymbol{X}]$ is modelled as a linear transformation of the stimulus features,

$$\vec{Y} = \boldsymbol{B}^T \vec{X} + \boldsymbol{\epsilon}$$

where $\boldsymbol{B}$ is a $p \times q$ coefficient matrix and $\boldsymbol{\epsilon}$ is a noise variable with an assumed multivariate normal distribution, $\boldsymbol{\epsilon} \sim N(0, \Sigma)$. Hence, the conditional density of $\vec{Y}|\vec{X}$ is given by the multivariate normal density

$$p(\vec{y}|\vec{x}) = -\frac{1}{(2\pi|\Sigma|)^{-q/2}} \exp\left[-\frac{1}{2}(\vec{y} - \boldsymbol{B}^T\vec{x})^T\Sigma^{-1}(\vec{y} - \boldsymbol{B}^T\vec{x})\right].$$

The coefficient B can be estimated from the training set data $(\boldsymbol{X}^{tr}, \boldsymbol{Y}^{tr})$ using a variety of methods for regularized regression, for instance, the elastic net [39], where each column of $\boldsymbol{B} = (\beta_1, \ldots, \beta_q)$ is estimated via

$$\hat{\beta}_i = \text{argmin}_\beta ||\boldsymbol{Y}_i^{tr} - \boldsymbol{X}^{tr}\beta||^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2,$$

where λ_1 and λ_2 are regularization parameters which can be chosen via cross-validation [11] separately for each column i.

After forming the estimated coefficient matrix $\hat{\boldsymbol{B}} = (\hat{\beta}_1, \ldots, \hat{\beta}_q)$, we estimate the noise covariance Σ via a shrunken covariance estimate[19][5] from the residuals,

$$\hat{\Sigma} = \frac{1}{N}((1-\lambda)S + \lambda \text{Diag}(S))$$

where

$$S = (\boldsymbol{Y}^{tr} - \boldsymbol{X}^{tr}\boldsymbol{B})^T(\boldsymbol{Y}^{tr} - \boldsymbol{X}^{tr}\boldsymbol{B}).$$

2.4 Converting the encoding model to a decoding model

Bayes' rule can be used to convert a probabilistic encoding model into a decoding model [22]. The Bayesian decoding model gives the posterior probability of the stimulus given the response,

$$p(\vec{x}|\vec{y}) = p(\vec{y}|\vec{x})\frac{p(\vec{x})}{p(\vec{y})}.$$

In an *identification task*[17], a response $\boldsymbol{y}$ is generated by presenting the subject to a stimulus which is randomly chosen from a subset of k stimuli, $S = (\vec{x}^{(1)}, \ldots, \vec{x}^{(k)})$. The decoder is used to select the stimulus in S which is most likely to have generated the response $\boldsymbol{y}$: the performance of the the decoder is measured by the probability of correct identification. In the identification task, the prior probability $p(\vec{x})$ is uniform over the candidate set S. Therefore, the estimated log posterior probability of each candidate stimulus $\vec{x}^{(i)}$ is given by

$$\log \hat{p}(\vec{x}|\vec{y}) = \log \hat{p}(\vec{y}|\vec{x}) + \text{const.} = -\frac{1}{2}(\vec{y} - \hat{\boldsymbol{B}}^T\vec{x})^T\hat{\Sigma}^{-1}(\vec{y} - \hat{\boldsymbol{B}}^T\vec{x}) + \text{const.}$$

where we have elided the inconsequential constant terms. Therefore, the chosen stimulus $\hat{\vec{x}}$ is the stimulus which minimizes the empirical Mahalanobis distance

$$d_{\hat{\Sigma}}(\vec{y}, \hat{\boldsymbol{B}}^T\vec{x}) = (\vec{y} - \hat{\boldsymbol{B}}^T\vec{x})^T\hat{\Sigma}^{-1}(\vec{y} - \hat{\boldsymbol{B}}^T\vec{x})$$

among the stimuli in S, and supposing that the correcty stimulus has index i, the probability of correct identification is

$$\Pr[\text{correct}] = \Pr[d_{\hat{\Sigma}}(\vec{y}, \hat{\boldsymbol{B}}^T \vec{x}^{(i)}) \leq \min_{j \neq i} d_{\hat{\Sigma}}(\vec{y}, \hat{\boldsymbol{B}}^T \vec{x}^{(j)})].$$

2.5 Computation of identification accuracy curve

The probability of correct identification varies depending on the choice of stimulus set S. Therefore, to obtain a well-defined measure of decoder precision, we define the k-class *identification risk* as the expected acccuracy when the set S is constructed by drawing $x^{(1)}, \ldots, x^{(k)}$ independently from the prior distribution $p(\vec{x})$.

An unbiased estimate of the k-class identification risk for any $k \leq M$ can be obtained, where M is the number of test observations. The idea is to evaluate the empirical accuracy (the proportion of correct identifications) over all combinations of $\binom{M}{k}$ stimulus subsets S times all k choices for the correct stimulus within S. Yet, this empirical accuracy can be computed without explicitly looping over all $\binom{kM}{k}$ combinations via a computational trick.

Suppose without loss of generality that the indices of the test observations are $i = 1, \ldots, M$. Define

$$M_{i,j} = \log \hat{p}(\vec{x}^{(j)} | \vec{y}^{(i)})$$

Furthermore, define

$$R_{i,j} = \sum_{\ell \neq j} I\{M_{i,\ell} \geq M_{i,j}\}.$$

The computational trick is to look at each combination of test response $\vec{y}^{(i)}$ and stimulus $\vec{x}^{(\ell)}$, and to count count the number of subsets $N_{i,\ell}$ where (i) both i and ℓ are included in S, and (ii) $\hat{x}^{(i)} = \vec{x}^{(\ell)}$. One can then verify that the empirical accuracy over all subsets is equal to

$$\text{EmpAcc}_k = 1 - \frac{1}{\binom{M}{k}} \frac{1}{k} \sum_{i=1}^{k} \sum_{\ell \neq i} C_{i\ell} N_{i,\ell}. \tag{1}$$

Now it is just a matter of simple combinatorics to compute $N_{i,\ell}$. We require both $\vec{x}^{(i)}$ and $\vec{x}^{(\ell)}$ to be included in S. This implies that if $M_{i,i} > M_{i,\ell}$, then $\vec{x}^{(\ell)}$ will never have the highest margin in any of those subsets, so $N_{i,\ell} = 0$.

Otherwise, there are $R_{i,\ell}-1$ elements with a lower margin than $\vec{x}^{(\ell)}$. Since $i \neq \ell$, then there are $k-2$ elements in $S \setminus \{i, \ell\}$, so therefore $N_{i,j,\ell} = \binom{R_{i,j,\ell}-2}{k-2}$. Therefore, we can write

$$N_{i,\ell} = I\{R_{i,\ell} > R_{i,i}\}\binom{R_{i,\ell}-2}{k-2} \tag{2}$$

The *identification accuracy curve* is defined as the function which maps $k \in 2, 3, \ldots$ to the k-class identification risk. Therefore, an estimate of a portion of the curve can be obtained by estimating the k-class identification risk for $k = 2, \ldots, M$.

2.6 Implied information

Define the Bayes risk as the identification risk of the optimal decoder. The result of ZB 2016 says that under certain regularity conditions, for for sufficiently high-dimensional $p(\vec{x}, \vec{y})$, we have

$$\text{BayesAcc}_k \approx \bar{\pi}_k(\sqrt{2I(\vec{x}; \vec{y})})$$

where $\bar{\pi}_k$ is the function

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

and where $I(\vec{x}; \vec{y})$ is the Shannon information

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

This is an important result because it implies that the entire identification accuracy curve can be summarized by a single parameter–the mutual information. This means that in the asymptotic regime specified by ZB 2016, (i) any portion of the curve can be used to estimate the mutual information and therefore reconstruct the entire curve, and (ii) that there exists a strict ordering over identification accuracy curves: for any two curves A_k and A'_k, one dominates the other for all k: either $A_k \geq A'_k$ for all $k \geq 2$, or $A'_k \geq A_k$ for all $k \geq 2$.

However, the result in ZB 2016 only applies to the optimal decoder, or *Bayes decoder*. Yet, it is impossible to obtain the Bayes decoder in practice,

since constructing the Bayes decoder requires knowing $p(\vec{x}, \vec{y})$. Therefore, we propose that under similar conditions to those stipulated in ZB 2016, for a certain class of classifiers[1], we have

$$\text{IdAcc}_k \approx \bar{\pi}_k(\sqrt{2I_{implied}})$$

where IdRisk_k is the k-class identification risk for a given classifier trained from the training set, and where $I_{implied}$ is a real-valued attribute of the classifier called the *implied information*. Furthermore, since $\text{IdAcc}_k \leq \text{BayesAcc}_k$ by definition (as BayesAcc_k is the best achievable accuracy), we have

$$I_{implied} \leq I(\vec{X}; \vec{Y}).$$

In order to estimate the implied information, we can rely on the fact that the empirical identification accuracy curve EmpAcc_k is an unbiased estimate of the true identification accuracy curve IdAcc_k. Therefore, we can estimate $I_{implied}$ by finding the theoretical curve which gives the best fit to the empirical accuracies in terms of mean-squared error. Thus, define $\hat{I}_{implied}$ as the nonlinear least-squares estimator

$$\hat{I}_{implied} = \text{argmin}_{\iota \geq 0} \sum_{k=2}^{M} (\text{EmpAcc}_k - \bar{\pi}_k(\sqrt{2\iota}))^2.$$

3 Theory

The main condition needed for the theory of ZB 2016 is that the log-likelihoods $\log p(\vec{Y}^{(i)}|\vec{X}^{(j)})$ have a jointly multivariate normal distribution for $(\vec{X}^{(i)}, \vec{Y}^{(i)}) \overset{iid}{\sim} p(\vec{x}, \vec{y})$ for $i = 1, \ldots, k$. The multivariate normal distributional assumption, in turn, approximately holds in high dimensions due to central limit theorem, due to the decomposition

$$\log p(\vec{Y}|\vec{X}) = \sum_{j=1}^{\dim(\vec{y})} \log p(Y_j|\vec{X}, Y_1, \ldots, Y_{j-1})$$

under the condition that the covariances between summands,

$$\text{Cov}[\log p(Y_j|\vec{X}, Y_1, \ldots, Y_{j-1}), \log p(Y_\ell|\vec{X}, Y_1, \ldots, Y_{\ell-1})] \tag{3}$$

[1]We leave it to future work to specify the conditions on the joint density and classifiers needed to formally establish the desired property.

are not too large. Therefore, two conditions are needed for the approximation to be effective

1. The dimensionality of the response $\vec{Y}$ must be large, so that the central limit theorem be applied.

2. The components of the response, Y_i, should be "close to independent" in the sense that the covariances of of log-likelihoods (3) are small–again, so that the assumptions of the central limit theorem can be satisfied.

However, due to the symmetry of mutual information, we can exchange the places of $\vec{X}$ and $\vec{Y}$ to get another set of necessary conditions: namely, that the dimensionality of $\vec{X}$ be large and that the components of $\vec{X}$ must be approximately indepndent. The reason why the $\vec{Y}$ conditions are equivalent to the $\vec{X}$ conditions is that low dimensionality in $\vec{Y}$ implies correlation among components in $\vec{X}$, and vice-versa.

The most effective way to illustrate the practical implications of these conditions is to list examples of experiments in neuroscience where the conditions are either well-satisfied, or violated.

- When $\vec{Y}$ is the mean spike count of a single neuron, and $\vec{X}$ is the wind direction, as in [**?**], the dimensionality of both the stimulus and the response is too low for the theory to apply.

- When the stimulus $\vec{X}$ consists of a images drawn from a mixture of a k different disjoint classes, as in [4] (mammal, insect, bird), even if $\vec{X}$ is high-dimensional, the mixture structure creates correlations between the components of $\vec{X}$. Whether or not the theory is still effective depends on the degree of separation between mixture components–we discuss this issue in more detail in Section 4.

- The ideal case is when $\vec{X}$ is high-dimensional, and drawn from a population with either no clear cluster structure or with very many clusters, and where $\vec{Y}$ is high-dimensional (e.g. a large region of interest in an fMRI scan), as in [17][21].

4 Simulations

To demonstrate the effectiveness of our procedure for recovering the identification accuracy curve in various settings, we apply our method to three different simulated examples: (a) a low dimensional case, (b) a high-dimensional case, (c) a high-dimensional mixture, as seen in Figure 4. Since only case (b) is close to satisfying the high-dimensionality and component dependence assumptions, we only expect the method to be effective in case (b).

Details of the simulation [to be moved to appendix]

- $\vec{X}$ is standard multivariate normal.
- $\vec{Y} = \vec{X}^T \boldsymbol{B} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is standard multivariate normal, and $\boldsymbol{B}$ is a randomly generated $p \times q$ coefficient matrix with independent $N(0, 1)$ entries.
- Training size n is varied; test set size is 100.
- Our method is used to estimate the accuracy curve from $K = 2$ to $K = 100$ from the empirical accuracies from $K = 2$ to $K = 50$.

5 Real-data examples

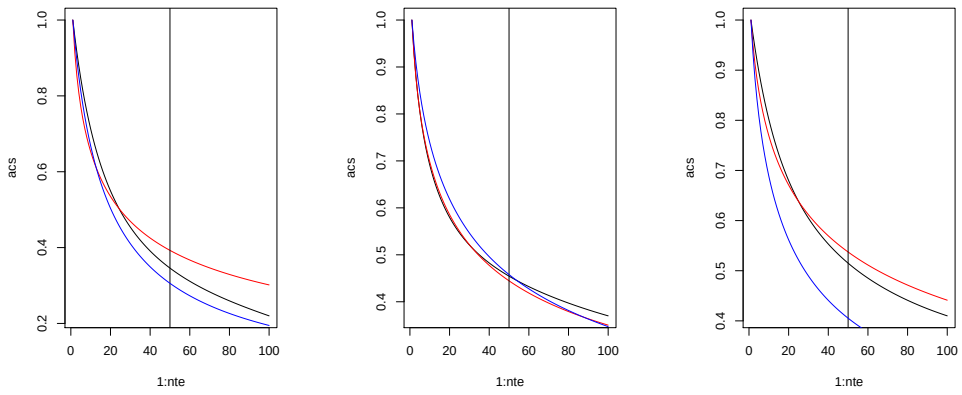
Examples using Kay et al data:

- Comparison of different feature models (subsets of pyramid levels)
- Comparison of differing voxel numbers
- Effect of voxel resolution (by averaging neighboring voxels into bigger voxels)

References

[1] Larry F. Abbott. Decoding neuronal firing and modelling neural networks. *Quarterly reviews of biophysics*, 27(3):291–331, aug 1994.

[2] Alexander Borst and Frédéric E. Theunissen. Information theory and neural coding. *Nature Neuroscience*, 2(11):947–957, nov 1999.



(a) (b) (c)

Figure 1: Success at recovering identification accuracy curves for various settings. Black = empirical accuracy curve, Blue = ground truth accuracy curve, Red = estimated accuracy curve. (a) Low-dimensional setting ($n = 20, p = 3, q = 5$). The estimated accuracy curve is not close to either the empirical curve nor the true accuracy curve. (b) High-dimensional setting ($n = 2000, p = 30, q = 50$). The estimated accuracy curve is close to both the empirical and true accuracy curves, and is a better estimate of the true accuracy than the empirical accuracy. (c) High-dimensional mixture setting. ($n = 2000, p = 30, q = 50$) Although the estimated curve is similar to the empirical curve, the true accuracy is much diffferent from either.

[3] Nicolas Brunel and J P Nadal. Mutual information, Fisher information, and population coding. *Neural computation*, 10(7):1731–57, 1998.

[4] Andrew C. Connolly, J. Swaroop Guntupalli, Jason Gors, Michael Hanke, Yaroslav O. Halchenko, Yu-Chien Wu, Herv?e Abdi, and James V. Haxby. The Representation of Biological Classes in the Human Brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(8):2608–2618, 2012.

[5] Michael J. Daniels and Robert E. Kass. Shrinkage Estimators for Covariance Matrices. *Biometrics*, 57(4):1173–1184, dec 2001.

[6] R Eckhorn, O.-J Grfisser, J Kr611er, K Pellnitz, and B P6pel. Efficiency of Different Neuronal Codes: Information Transfer Calculations for Three Different Neuronal Systems. *Biol. Cybernetics*, 22:49–60, 1976.

[7] Adrienne L. Fairhall, Geoffrey D. Lewen, William Bialek, and Robert R. de Ruyter van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(23):787–792, aug 2001.

[8] Galit Fuhrmann Alpert, Fellice T. Sun, Daniel Handwerker, Mark D'Esposito, and Robert T. Knight. Spatio-temporal information analysis of event-related BOLD responses. *NeuroImage*, 34(4):1545–1561, 2007.

[9] Michael C. Gastpar, Patrick R. Gill, and Frédéric E. Theunissen. Anthropic correction of information estimates. *Proceedings - 2009 IEEE Information Theory Workshop on Networking and Information Theory, ITW 2009*, 56(2):152–155, 2009.

[10] Amir Globerson, Eran Stark, Eilon Vaadia, and Naftali Tishby. The minimum information principle and its application to neural code analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(9):3490–5, mar 2009.

[11] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 1. Springer, 2 edition, 2009.

[12] N G Hatsopoulos, C L Ojakangas, L Paninski, and J P Donoghue. Information about movement direction obtained from synchronous activity of

motor cortical neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 95(26):15706–11, dec 1998.

[13] David H. Hubel. Evolution of ideas on the primary visual cortex, 19551978: A biased historical account. *Bioscience Reports*, 2(7):435–469, 1982.

[14] Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, 76(6):1210–1224, 2012.

[15] Robin A A Ince, Rasmus S Petersen, Daniel C Swan, and Stefano Panzeri. Python for information theoretic analysis of neural data. *Frontiers in neuroinformatics*, 3:4, 2009.

[16] Robin A.A. Ince, Riccardo Senatore, Ehsan Arabzadeh, Fernando Montani, Mathew E. Diamond, and Stefano Panzeri. Information-theoretic methods for studying population codes. *Neural Networks*, 23(6):713–727, 2010.

[17] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(March):352–355, 2008.

[18] Christoph Kayser, Nikos K. Logothetis, and Stefano Panzeri. Visual Enhancement of the Information Representation in Auditory Cortex. *Current Biology*, 20(1):19–24, 2010.

[19] Olivier Ledoit and Michael Wolf. Honey, I Shrunk the Sample Covariance Matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.

[20] Miller. Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods*, 1955.

[21] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320(5880), 2008.

[22] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011.

[23] Thomas Naselaris, Ryan J. Prenger, Kendrick N. Kay, Michael Oliver, and Jack L. Gallant. Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, 63(6):902–915, 2009.

[24] Israel Nelken, Gal Chechik, Thomas D Mrsic-Flogel, Andrew J King, and Jan W H Schnupp. Encoding Stimulus Information by Spike Numbers and Mean Response Time in Primary Auditory Cortex. *Journal of Computational Neuroscience*, 19:199–221, 2005.

[25] Ilya Nemenman, William Bialek, and Rob de Ruyter van Steveninck. Entropy and information in neural spike trains: progress on the sampling problem. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69(5 Pt 2):056111, may 2004.

[26] Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21:1641–1646, 2011.

[27] Masafumi Oizumi, Toshiyuki Ishii, Kazuya Ishibashi, Toshihiko Hosoya, and Masato Okada. Mismatched Decoding in the Brain. *Journal of Neuroscience*, 30(13):4815–1826, 2010.

[28] Mike W. Oram, Peter Földiák, David I. Perrett, Mike W. Oram, and Frank Sengpiel. The 'Ideal Homunculus': decoding neural population signals. *Trends in Neurosciences*, 21(6):259–265, 1998.

[29] Liam Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, 2003.

[30] Stefano Panzeri, Riccardo Senatore, Marcelo A. Montemurro, and Rasmus S. Petersen. Correcting for the Sampling Bias Problem in Spike Train Information Measures. *Journal of Neurophysiology*, 98(3), 2007.

[31] Brian N. Pasley, Stephen V. David, Nima Mesgarani, Adeen Flinker, Shihab a. Shamma, Nathan E. Crone, Robert T. Knight, and Edward F. Chang. Reconstructing speech from human auditory cortex. *PLoS Biology*, 10(1), 2012.

[32] Russell A. Poldrack. Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10(2):59–63, 2006.

[33] Rodrigo Quian Quiroga and Stefano Panzeri. Extracting information from neuronal populations: information theory and decoding approaches. *Nature reviews. Neuroscience*, 10(3):173–185, 2009.

[34] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, jul 1948.

[35] S. P. Strong, Roland Koberle, Rob R. de Ruyter van Steveninck, and William Bialek. Entropy and Information in Neural Spike Trains. *Physical Review Letters*, 80(1):197–200, jan 1998.

[36] Alessandro Treves and Stefano Panzeri. The Upward Bias in Measures of Information Derived from Limited Data Samples. *Neural Computation*, 7(2):399–407, 1995.

[37] Stuart Yarrow, Edward Challis, and Peggy Seriès. Fisher and Shannon Information in Finite Neural Populations. *Neural Computation*, 24(7):1740–1780, 2012.

[38] Charles Y. Zheng and Yuval Benjamini. Estimating mutual information in high dimensions via classification error. jun 2016.

[39] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, apr 2005.