Using confusion matrices to estimate mutual information between two categorical measurements

Dirk B. Walther (bernhardt-walther.1@osu.edu)
Department of Psychology, The Ohio State University

***Abstract*—Many data analysis problems in neuroimaging are set up as classification problems, frequently involving multiple classes. Typically, only the fraction of correct classifications is reported as aggregate accuracy. However, the structure of the classification errors contains valuable information as well. By reinterpreting confusion matrices as conditional probabilities we not only demonstrate a procedure for computing the mutual information between a categorical measurement and ground truth, but we also derive a mechanism for computing mutual information between two separate measurements of the same shared ground truth. We demonstrate this approach with fMRI and behavioral data for categorization of natural scenes.**

I. Introduction

How much can we learn about the stimulus presented to an observer from a measurement, e.g., of her brain activity? The typical approach in multi-voxel pattern analysis (MVPA) is to report prediction accuracy as the fraction of correctly guessed labels [1]. However, the specific pattern of confusions that is captured in the confusion matrix contains a richer representation of the relationship between ground truth labels and predicted labels. Here we use information theory to explore this relationship. For the purpose of this analysis let us consider the ground truth label as a signal that is transmitted through a noisy communication channel. In the case of an fMRI experiment the communication channel consists of the display apparatus, the eye and brain of the observer, the MRI scanner, fMRI image reconstruction, preprocessing and, finally, the MVPA classifier. The observed signal is the classifier's prediction of the label. In the case of a multiple-choice behavioral experiment, the channel consists again of the display apparatus, the eye, brain and hand of the observer, and, finally, the apparatus used to record the response. The observed signal is the key press recorded by the computer. In this paper we aim to measure the capacity of these noisy communication channels, that is, to measure the mutual information between the response and ground truth.

Going one step further, let us consider two separate measurements of the same ground truth label. How much information do the two measurements have in common? These two measurements may arise from separate experiments, even on separate groups of observers, as long as they share the same distribution of ground truth labels. For instance, one measurement could be decoding of stimulus category from fMRI activity, the other could be a behavioral experiment. We quantify how closely a brain region is associated with human behavior by measuring mutual information.

In this paper both of these information measures are derived from confusion matrices, which are routinely recorded in measurements of categorical variables. The technique is demonstrated using the example of natural scene categorization.

Thanks to Per Sederberg, Woojae Kim, and the BWlab for helpful discussions.

II. Methods

A. Mutual information with ground truth

As has been described, for instance, in [2] and [3], a confusion matrix CM can be interpreted as a table of conditional probabilities $p(R|L)$ of obtaining response R given ground truth label L. We know the marginal probability $p(L)$ from the design of the experiment. In many cases $p(L)$ will be uniform. The joint probability of R and L as well as the marginal probability of R are given by:

$$p(R,L) = p(R|L) \cdot p(L), \tag{1}$$

$$p(R) = \sum_L p(R,L). \tag{2}$$

From these probabilities we compute entropies and from those the mutual information of R and L as:

$$H(R) = -\sum_R p(R) \cdot \log(p(R)) \tag{3}$$

$$H(L) = -\sum_L p(L) \cdot \log(p(L)) \tag{4}$$

$$H(R,L) = -\sum_R \sum_L p(R,L) \cdot \log(p(R,L)) \tag{5}$$

$$I(R;L) = H(R) + H(L) - H(R,L) \tag{6}$$

B. Mutual information between two measurements

We use a similar mechanism to compute mutual information between two separate measurements R and Q, as long as both refer to the same set of ground truth labels L with the same distribution $p(L)$. Assuming conditional independence of R and Q given L, we can expand the joint conditional probability as:

$$p(R,Q|L) = p(R|L) \cdot p(Q|L). \tag{7}$$

Then the joint probability of R, Q, and L is:

$$p(R,Q,L) = p(R|L) \cdot p(Q|L) \cdot p(L). \tag{8}$$

Marginalizing over L yields:

$$p(R,Q) = \sum_L p(R|L) \cdot p(L) \cdot p(Q|L). \tag{9}$$

978-0-7695-5061-9/13 $26.00 © 2013 IEEE
DOI 10.1109/PRNI.2013.63

This can be rewritten as matrix multiplication, using the confusion matrices CM_R and CM_Q for responses R and Q:

$$[p(R,Q)] = CM_R^T \cdot \mathrm{diag}(p(L)) \cdot CM_Q, \tag{10}$$

where $\mathrm{diag}(p(L))$ refers to a square matrix with $p(L)$ along the diagonal and zero everywhere else. In the case of uniform $p(L)$ this matrix can be replaced with a scalar. Marginalizing eq. 9 gives us $p(R)$ and $p(Q)$, so that we can compute the entropies $H(R)$, $H(Q)$ and $H(R,Q)$ analogous to eqs. 3-5. The mutual information of R and Q is then:

$$I(R;Q) = H(R) + H(Q) - H(R,Q). \tag{11}$$

C. fMRI experiment

We evaluate these formalisms with previously published fMRI data on natural scene categories [4]. In the experiment, which was approved by the Internal Review Board (IRB) of the University of Illinois at Urbana-Champaign, ten participants passively viewed images of natural scenes from six categories (beaches, forests, mountains, city streets, highways, offices). Images were blocked by scene category, and each run of the experiment contained one block from each category. Participants viewed six runs with color photographs and six runs with line drawings while in the MRI scanner. Retinotopic mapping and a standard functional localization experiment were used to identify regions of interest (ROI): visual areas V1, V2, and V4, as well as the parahippocampal place area (PPA), the retrosplenial cortex (RSC), the lateral occipital cortex (LOC), and, as a control area, the fusiform face area (FFA). ROI-based MVP analysis was performed in individual subject space. After minimal pre-processing (motion correction, normalization to percent signal change) BOLD activity corresponding to individual image blocks was used as input to a linear support vector machine (SVM) to generate predictions for scene categories in a leave-one-run-out cross validation. Confusion matrices were recorded for each ROI and averaged over all ten participants. Decoding accuracy (mean of the diagonal elements of a confusion matrix) is shown in Table I for all ROIs for color photographs and line drawings.

In a whole-brain analysis, a sphere with a radius of 5 voxels and a volume of 81 voxels was centered on each voxel in turn. Decoding of scene category was performed in the same way as in the ROI-based analysis, resulting in a confusion matrix for each voxel in the brain. Confusion matrices for all participants were averaged in MNI space.

D. Behavioral experiment

We compare the fMRI results with data from a six-alternative forced-choice (6AFC) behavioral experiment with 18 participants [5], which was approved by the IRB of The Ohio State University. The same color photographs and line drawings of natural scenes that were used in the fMRI experiment were displayed briefly and followed by a perceptual mask. Participants were asked to press one of six keys on a keyboard to indicate the category of each scene. Assignment of keys to categories was randomized for each participant, and presentation times were adjusted for each participant individually such that they achieved 65% accuracy during training. Each participant saw 360 images, half as color photographs and half as line drawings. Responses were recorded in confusion matrices separately for color photographs and line drawings, and confusion matrices were averaged over participants.

III. Results

A. Mutual information with ground truth

We computed mutual information of decoding from fMRI activity in various ROIs with the ground truth labels using eqs 1-6. This computation distills the rich information that is contained in the full confusion matrix down to a single value, the mutual information. This value bears some relation to decoding accuracy, but it is determined by the entire confusion matrix, not just its diagonal.

Results for our fMRI experiment of natural scene categorization are shown in the second set of rows in Table I. For color photographs (CP) we observe the highest mutual information for the PPA, which also shows the highest decoding accuracy. Areas V1 and V2 also show relatively high mutual information. The added value of the mutual information approach can be seen when comparing V4 with RSC. Both ROIs have fairly similar decoding accuracy. However, RSC shows considerably higher mutual information than V4, indicating that its confusion structure is more informative about ground truth than that of area V4. Object-sensitive LOC and face-sensitive FFA show low decoding accuracy and low mutual information.

Comparing mutual information values between CP and line drawings (LD) is quite instructive. Although decoding accuracy for the PPA is almost the same for CP and LD, mutual information with ground truth is markedly higher for LD than for CP. Notably, both decoding accuracy and mutual information for V1 are higher for LD than CP, possibly due to the clearly defined contours in the line drawings.

For the behavioral experiment accuracy was at 77.3% for CP and at 66.2% for LD. Mutual information of behavior with ground truth was 1.387 bits for CP and 0.963 bits for LD, an order of magnitude larger than for the much noisier fMRI data.

B. Mutual information with behavior

How are the predictions from the neuroimaging data related to the behavioral results? We use eqs. 9-11 to compute the mutual information of the predictions from the fMRI data with the behavioral results. In the third set of rows in Table I we observe that the PPA shows by far the best match with behavioral data, for both CP and LD. This finding is in agreement with our previous work using the correlation of errors [6]. In spite of it's fairly low decoding accuracy, the RSC shows high mutual information with the behavioral results as well.

These results demonstrate that the mutual information computed from the full confusion matrices of two separate measurements (here: fMRI decoding and a 6AFC behavioral experiment) can uncover relationships that are not obvious from mere accuracy measures. This is possible, because the

TABLE I: ROI-based fMRI results: Accuracy of decoding scene categories, mutual information with ground truth, and mutual information with behavior. Results are shown separately for color photographs (CP) and line drawings (LD). Last row: Mutual information between fMRI decoding from CP and LD.

		V1	V2	V4	PPA	RSC	LOC	FFA
Mean decoding accuracy in percent (chance: 16.7)	CP	24.2	26.7	23.7	31.9	22.6	21.0	17.9
	LD	28.8	20.3	25.6	29.2	22.5	21.2	18.3
Mutual information with ground truth ($\times 0.01$ bits)	CP	10.36	11.39	6.58	17.35	11.75	6.07	4.28
	LD	18.57	9.44	11.24	25.75	11.33	3.70	3.72
Mutual information with behavior ($\times 0.01$ bits)	CP	5.77	6.47	4.25	11.53	6.91	3.57	2.45
	LD	8.36	3.78	5.26	12.48	5.02	1.54	1.49
Mutual information between CP and LD ($\times 0.01$ bits)		0.833	0.562	0.255	3.251	0.705	0.108	0.050

two measurements, although following different experimental procedures with separate sets of subjects, used the same stimulus set, and the probability for the occurence of each scene category was the same in both cases, namely uniform.

C. Mutual information between ROIs

Eq. 11 can also be used to measure the information that two brain areas have in common. In this case the confusion matrices obtained from predicting scene categories in two separate ROIs are used to compute mutual information. The results for all pairings of our ROIs (except for the FFA, which only served as a control area) are shown separately for CP and LD in Tables II and III.

For color photographs we observe a strong relationship only between PPA and RSC, which are both brain areas known to specialize in scene processing, and between visual areas V1 and V2 (Table II). For line drawings, however, we see a high level of mutual information for V1 with V2 and V4, and also with the PPA, as well as for V4 with PPA (Table III). Together with the high mutual information of V1 with ground truth for LD (Table I, second set of rows) this suggests that for line drawings processing in V1 gains in importance for determining scene category, and that this information is shared readily with the PPA as well as with V4. PPA, in turn, is closely related to human behavior (Table I, third set of rows). The change in information sharing between these brain areas from CP to LD could reflect a reorganization of the flow of visual information in response to the different feature composition of line drawings compared to color photographs of scenes. Please note, however, that no directionality of information flow is implied by the MI analysis.

While these values can be useful for comparing the two stimulus conditions, the absolute values of the shared information between brain areas need to be interpreted with caution. The conditional independence of the two measurements R and Q that is required for eq. 7 cannot be guaranteed when considering concurrent fMRI activity in different brain regions. In fact, since both measurements follow the exact same time course, more potent measures of functional or effective connectivity can be employed, such as mutual information of the time series [7] or transfer entropy [8].

TABLE II: Mutual information ($\times 0.01$ bits) for pairs of ROIs for decoding scene category from color photographs.

	V2	V4	PPA	RSC	LOC
V1	**0.898**	0.357	0.444	0.146	0.255
V2		0.401	0.460	0.155	0.210
V4			0.513	0.285	0.138
PPA				**1.113**	0.495
RSC					0.228

TABLE III: Mutual information ($\times 0.01$ bits) for pairs of ROIs for decoding scene category from line drawings.

	V2	V4	PPA	RSC	LOC
V1	**1.212**	**1.630**	**2.717**	0.693	0.362
V2		0.727	0.783	0.169	0.142
V4			**1.619**	0.425	0.214
PPA				**1.903**	0.449
RSC					0.142

D. Mutual information between photographs and line drawings

Methods based on time series are not available, however, when comparing two separate, non-concurrent experimental conditions within the same experiment. In our fMRI experiment, color photographs and line drawings were presented in separate runs. Our method allows us to assess the amount of shared information between these two conditions in specific regions of interest. We used equations 7-11 with decoding from CP as measure R and decoding from LD as measure Q to explore this relationship. Results for all ROIs are shown in the last row of Table I. In agreement with the results obtained when using error correlations in [4], we find relatively high mutual information between CP and LD in the PPA, but considerably lower values in other high-level visual areas as well as early visual cortex. Specifically, the FFA, which is not involved in processing scene content, shows very low values.

E. Whole-brain analysis

In addition to testing specific hypotheses about pre-defined ROIs we can also use mutual information analysis as an exploratory tool. Here we explore the mutual information between the prediction of scene categories at each location in the brain and ground truth (see methods and reference [4] for details of the searchlight analysis). We obtain this mutual information for every voxel in the brain. In Figure 1 we show the results separately for CP and LD.

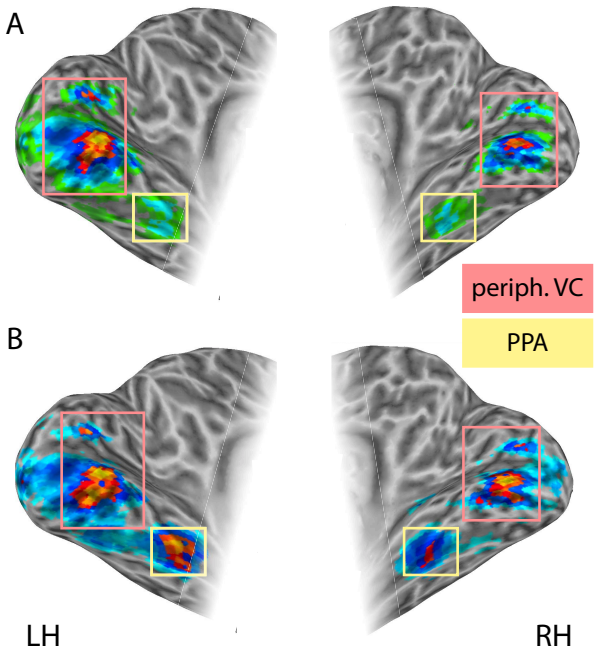

Fig. 1: Whole-brain group analysis (10 subjects) of mutual information of predicted scene category labels with respect to ground truth for (A) color photographs and (B) line drawings. Visual cortex regions corresponding to the periphery of the visual field and the parahippocampal place area (PPA) show high levels of mutual information, validating our initial choice of regions of interest. The PPA has a higher level of mutual information for line drawings than for photographs, mirroring the results shown in Table I.

The maps were thresholded such that only the top 2% of voxels are displayed. The maps are dominated by high values of mutual information in the parts of visual cortex that encode the periphery of the visual field on both banks of the calcarine fissure (and, therefore, in both the upper and lower visual field) and in the PPA. The higher values on the lower than the upper bank of the calcarine fissure presumably reflect the contributions of area V4, which is only located on the lower bank. No further brain regions show a similarly high level of mutual information with ground truth. This finding re-confirms our a-priori choice of regions of interest. Inspection of the distribution of mutual information between visual cortex and PPA shows higher values in the PPA for LD than for CP, mirroring the ROI-based results in Table I.

IV. Discussion

We have derived procedures for computing the mutual information between a categorical measurement and ground truth as well as between two separate categorical measurements of the same ground truth. We have demonstrated the use of these procedures with the example of an fMRI experiment and a behavioral experiment of natural scene categorization. The method allows for some new insights into the processing of categorical stimuli. However, several open questions need to be addressed to make it fully viable.

A. Open questions

Multi-voxel pattern analysis (MVPA) of fMRI data is fairly noisy, and accuracies, though significantly above chance, are relatively low. This leads to fairly small values for mutual information with ground truth. Note that the values in all three tables are in units of 0.01 bits, where maximally attainable mutual information for the six-way classification would be $\log_2 6 \approx 2.585$ bits. What level of mutual information should we accept as reliably above zero? Can we develop something like a significance test for mutual information? The most robust method for establishing significance for MVPA is a non-parametric permutation analysis. Permuting the rows and columns of one of the confusion matrices, however, would yield the same mutual information as before, because summation to compute the entropies is commutative. What would happen if we permuted the ground truth labels for individual experiment trials or blocks and repeated the MVP analysis? The expected confusion matrix for such an analysis would be a uniform matrix with chance level in every cell, which would result in mutual information being equal to zero. This would not help in finding a significance test. What, then, is a criterion for accepting certain values of mutual information as high enough? The choice of the 2% cut-off used in Figure 1 was completely arbitrary. Clearly, a more rigorous framework for establishing significance for mutual information will need to be developed. One possible avenue could be using maximum likelihood estimates as described in the supplementary material of [9].

B. Relation to other work

Information theory has been used to formulate techniques for univariate [10, 11] and multivariate [7] functional connectivity as well as effective connectivity using transfer entropy [8]. These methods rely on the time course of the activity signal and should be used for this purpose when the time course is available and shared between the two measures being compared as,for instance, for functional connectivity analysis [12, 7]. The method that we have put forward in this paper serves a different purpose. Our method can be used with any experimental technique that generates a confusion matrix: neuroimaging, behavioral, computational models etc., even if they do not share the same experimental time course.

We have previously related confusion matrices from fMRI decoding to those from a behavioral experiment by correlating the errors (off-diagonal elements) between the two measurements [6, 4, 13]. Our new method is similar in spirit, but by considering the entire confusion matrix, including the diagonal elements, we integrate the previously separate considerations of accuracy (diagonal) and error terms (off-diagonal). Furthermore, the new method allows for the condensation of an entire confusion matrix into a single number.

Representational similarity analysis (RSA) is also similar in spirit [14]. However, the correlations between representations that are at the heart of RSA are not easily interpreted as

probabilities, whereas confusion matrices are already properly normalized conditional probabilities. Many problems that can be addressed using RSA can be reformulated into a classification problem yielding a confusion matrix. The reverse is not necessarily true. For instance, the 6AFC behavioral experiment described in this paper, which naturally results in a confusion matrix, cannot be reformulated in the RSA framework in a straight-forward manner. In this sense, using confusion matrices to compute mutual information can be more general than RSA in its potential applications and more readily interpreted in a probabilistic framework.

C. Conclusions

In summary, analyzing confusion matrices using the mutual information framework presented in this paper could open new avenues for the analysis of categorical data. Once several open questions have been addressed satisfactorily, this method is likely to be most useful for relating prediction results across different modalities, or for comparing measurements to model predictions or to human behavior. The method is powerful, because it utilizes the rich error representations contained in confusion matrices.

References

[1] D. D. Cox and R. L. Savoy. Functional magnetic resonance imaging (fMRI) brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, 19:261–270, 2003.

[2] I. Kononenko and I. Bratko. Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6:67–80, 1991.

[3] R. Q. Quiroga and S. Panzeri. Extracting information from neuronal populations: information theory and decoding approaches. *Nat Rev Neurosc*, 10:173–185, 2009.

[4] D. B. Walther, B. Chai, E. Caddigan, D. M. Beck, and L. Fei-Fei. Simple line drawings suffice for functional mri decoding of natural scene categories. *Proc Natl Acad Sci U S A*, 108(23):9661–6, 2011.

[5] D. Shen and D. B. Walther. Categorization of line drawings of natural scenes using non-accidental properties matches human behavior. *Journal of Vision*, 12(9):a593, 2012.

[6] D. B. Walther, E. Caddigan, L. Fei-Fei, and D. M. Beck. Natural scene categories revealed in distributed patterns of activity in the human brain. *J Neurosci*, 29(34):10573–81, 2009.

[7] B. Chai, D. B. Walther, D. M. Beck, and L. Fei-Fei. Exploring functional connectivities of the human brain using multivariate information analysis. In *Neural Information Processing Systems*, 2009.

[8] T. Schreiber. Measuring information transfer. *Phys Rev Lett*, 85(2):461–4, 2000.

[9] E. Y. Kimchi and M. Laubach. Dynamic encoding of action selection by the medial striatum. *J. Neurosc.*, 29:3148–3159, 2009.

[10] A. Tsai, J. W. Fisher III, C. Wible, W. M. Wells III, J. Kim, and A. S. Willsky. Analysis of functional MRI data using mutual information. In C. Taylor and A. Colchester, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI99*, pages 473–480, 1999.

[11] B. D. Ward and Y. Mazaheri. Information transfer rate in fMRI experiments measured using mutual information theory. *J Neurosci Methods*, 167(1):22–30, 2008.

[12] K. J. Friston. Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2:56–78, 1994.

[13] D. B. Walther, D. M. Beck, and L. Fei-Fei. To err is human: correlating fMRI decoding and behavioral errors to probe the neural representation of natural scene categories. In N. Kriegeskorte and G. Kreiman, editors, *Visual population codes – Toward a common multivariate framework for cell recording and functional imaging*, pages 391–415. MIT Press, Cambridge, MA, 2012.

[14] N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.*, 2:4, 2008.