

# Inferring exemplar discriminability in brain representations

Hamed Nili, Alexander Walther, Arjen Alink, Nikolaus Kriegeskorte

Medical Research Council Cognition and Brain Sciences Unit

## Abstract

Representational distinctions within categories are important in all perceptual modalities and also in cognitive and motor representations. Recent pattern-information studies of brain activity have used condition-rich designs to sample the stimulus space more densely. To test whether brain response patterns discriminate among a set of stimuli (e.g. exemplars within a category) with good sensitivity, we can pool statistical evidence over all pairwise comparisons. A popular test statistic reflecting exemplar information is the *exemplar discriminability index (EDI)*, which is defined as the average of the pattern dissimilarity estimates between different exemplars minus the average of the pattern dissimilarity estimates between repetitions of identical exemplars. The EDI is commonly tested with a  $t$  test ( $H_0$ : population mean EDI = 0) across subjects (subject as random effect). However, it is unclear whether this approach is either valid or optimal. Here we describe a wide range of statistical tests of exemplar discriminability and assess the validity (specificity) and power (sensitivity) of each test. The tests include previously used and novel, parametric and nonparametric tests, which treat subject as a random or fixed effect, and are based on different dissimilarity measures, different test statistics, and different inference procedures. We use simulated and real data to determine which tests are valid and which are most sensitive. The popular across-subject  $t$  test of the EDI (typically using correlation distance as the pattern dissimilarity measure) requires the assumption that the EDI is 0-mean normal under  $H_0$ , which is not strictly true. Reassuringly, our simulations suggest that the test controls the false-positives rate at the nominal level and is thus valid in practice. However, test statistics based on average Mahalanobis distances or average linear-discriminant  $t$  values (both accounting for the multivariate error covariance among responses) are substantially more powerful for both random- and fixed-effects inference. We suggest preferred procedures for safely and sensitively detecting subtle pattern differences between exemplars.

## 1. Introduction

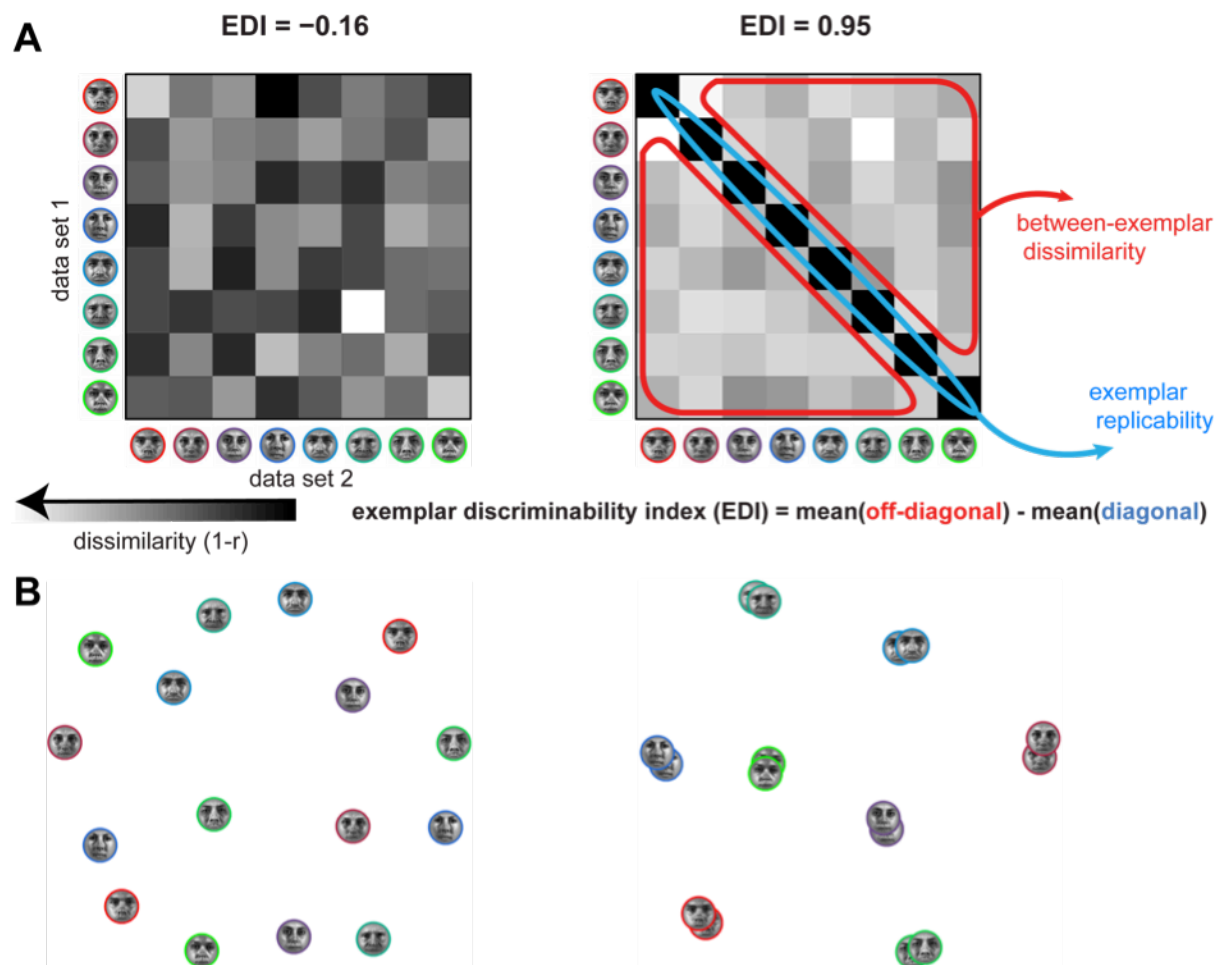
Brain representations are increasingly investigated with pattern-information analyses of brain-activity data acquired with brain imaging or neuronal recording techniques (Haxby et al., 2001; Hung et al., 2005; Kamitani and Tong, 2005; Kriegeskorte et al., 2006; Norman et al., 2006; Kiani et al., 2007; Kriegeskorte et al., 2008b; Mur et al., 2009; Kriegeskorte and Kievit, 2013; Kriegeskorte & Kreiman, 2011). These analyses seek to quantify different types of information present in the brain-activity patterns. Information carried by brain response-patterns can be explored at different levels. Two common levels are the category and the exemplar level. Category information has previously been studied using regional average activation (e.g. Kanwisher et al., 1997) and pattern decoding approaches (e.g. Haxby et al., 2001). Pattern-classifier decoding lends itself naturally to investigating category information with the category labels being decoded from the brain-activity patterns. Recent studies have estimated a unique response pattern for each individual stimulus and investigated not only category information, but also within-category exemplar information (Kay et al., 2008; Mitchell et al., 2008; Kriegeskorte et al., 2008b). Exemplar information is present to the degree that different exemplars elicit distinct representational patterns. Within-category effects are often subtle, thus powerful tests are needed to detect them. Here we use the term “exemplar discriminability” generically to denote the average discriminability across all pairs among a set of experimental conditions.

Studying representational distinctions within categories can arise at different contexts. For example, in studies of visual face representations, it is important to quantify to what extent a face region distinctly represents individual faces (Kriegeskorte et al. 2007; Nestor et al., 2011; Anzellotti et al., 2013). Classifier decoding can be used to test for exemplar information. For example, a classifier can be trained to distinguish two exemplars within a category. For power, we would then like to have many repetitions of each stimulus. This would suggest repeating a small number of exemplars many times in the experiment (e.g. Kriegeskorte et al. 2007). It is also desirable, however, to sample the category with many different exemplars in order to get a richer description of its underlying representation. There is a tradeoff between the number of stimuli and the number of repetitions of each stimulus, because the total time available for measuring brain activity in a subject is usually limited. If we have many different exemplars (i.e. a condition-rich design), we can typically repeat each stimulus only a few times. This severely limits our power to detect the discriminability of a given pair of exemplars. In the case of condition-rich designs, it is therefore desirable to pool the evidence across many pairs of exemplars. This is achieved by a summary statistic that combines the evidence of discriminability across all pairs of exemplars.

A popular summary statistic reflecting exemplar discriminability among a set of experimental conditions from the same category is the “exemplar discriminability index” (EDI; e.g. Sayres and Grill-Spector, 2008; Schwarzlose et al., 2008; Chan et al., 2010; Kravitz et al., 2010; Lee et al., 2012; Liu et al., 2013). The EDI is defined as the average between-exemplar dissimilarity estimate minus the average within-exemplar dissimilarity estimate (Fig. 1). A within-exemplar dissimilarity is a dissimilarity between independent measurements of the activity pattern elicited by repeated presentations of the same stimulus. The average within-exemplar dissimilarity estimate thus reflects the noise in the measurements. Subtracting it is essential when the dissimilarity estimate (i.e. pattern-distance estimates) is positively biased. The correlation distance, for example, which was used in the cited studies, is non-negative by definition, and therefore positively biased (although correlation coefficients are between -1 and 1, correlation distance, defined as 1 minus the correlation coefficient, is between 0 and 2). Subtracting the average within-exemplar dissimilarity removes the bias and enables inference. Typically, the EDI is computed for each subject and tested using a one-sided  $t$  test, treating subject as a random effect. It must be noted that the previous studies that use EDI do not use this nomenclature. However, we consistently use ‘EDI’ for any summary statistic computed in the way that is described above.

The cited studies using this approach rely on splitting the data into two independent halves (e.g. odd and even runs in fMRI measurements) and comparing response patterns between the two halves. The between-halves dissimilarities are assembled in a split-data representational dissimilarity matrix (*sdRDM*), which is indexed vertically by response patterns estimated from data set 1 and horizontally by response patterns estimated from data set 2. Each entry of the *sdRDM* contains one dissimilarity between two response pattern estimates spanning the two data sets. The order of the exemplars is the same horizontally and vertically, so the diagonal of the *sdRDM* contains the within-exemplar pattern dissimilarity estimates (Fig. 1A).

The within-data-set pattern comparisons are not used for either within- or between-exemplar pattern comparisons. This is important because patterns measured closer in time tend to be more similar due to measurement artefacts (Hendriksson et al., 2014, Alink et al., 2015). Note that for the popular correlation distance ( $1 - \text{Pearson } r$ ), the difference of distances is equal to the negative of the difference of correlations:  $(1-r_1) - (1-r_2) = r_2 - r_1$ . For a consistent comparison with other distance measures, however, we use the correlation *distance* here.

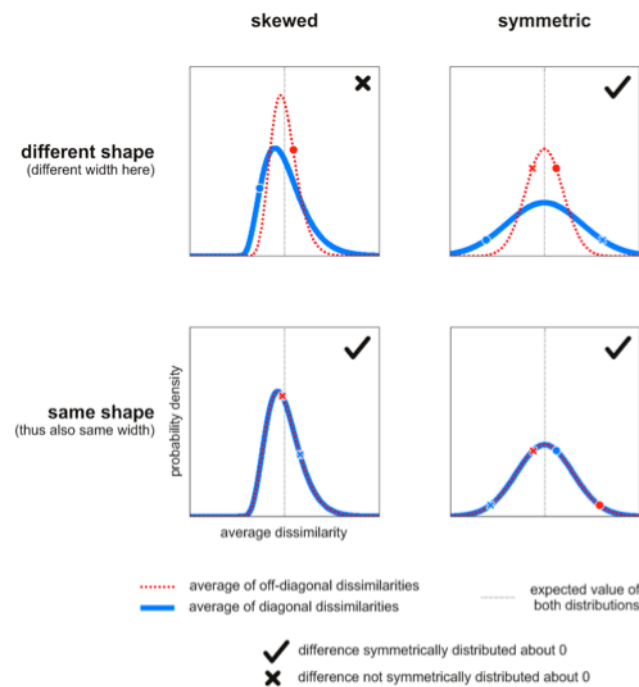


**Figure 1: Definition of the exemplar discriminability index (EDI).** (A) The exemplar discriminability index (EDI) is defined as the difference of the average between-exemplar and average within-exemplar distance. Split-data RDMs (sdRDMs) are shown for two scenarios. In the scenario on the left, the face exemplars are not discriminable and the EDI is insignificant ( $H_0$  simulation). In the scenario on the right, the EDI is large and exemplars are discriminable ( $H_1$  simulation). (B) Applying multi-dimensional scaling (MDS; Kruskal and Wish, 1978; Torgerson, 1958) allows simultaneous visualisation of the representational geometries from both data sets ( $H_0$  on the left,  $H_1$  on the right). Exemplars are colour-coded and there are two pattern estimates for each exemplar. MDS gives a low dimensional (2D here) arrangement, wherein the distances between projections approximate the original distances in the high-dimensional pattern space. Here, to obtain MDS plots, we consider a distance matrix composed of comparisons between any two patterns (aggregated from both datasets A and B). When the EDI is significantly positive (right), repetitions of the same exemplar yield patterns that are more similar to each other than presentations of two different exemplars.

The EDI  $t$  test has some caveats. First, this approach only allows testing exemplar information with subject as a random effect. Single-subject or group-level inference with subject as fixed effect cannot be accommodated in this approach. Second, it is possible that the assumptions of the test are not met. A  $t$  test requires that the data are normally distributed under  $H_0$ . The EDI is the difference of two average dissimilarities. For non-negative dissimilarities like the Euclidean distance, the distributions of the within- and the between-exemplar dissimilarities are skewed (limited by 0 on the left, unlimited on the right). Under the null hypothesis, the within- and between-exemplar dissimilarities are all samples

from the same distribution (for a given exemplar, the effect of changing the exemplar is not different from the effect of having another measurement for that exemplar). The expected value of the mean of the diagonal entries of the sdRDM (within-exemplar) is therefore equal to the expected value of the mean of the off-diagonal (between-exemplar) entries. The expected value of the difference between those means, i.e. the expectation of the EDI, is therefore also zero under  $H_0$ . However there are more between-exemplar than within-exemplar dissimilarities. For  $N$  exemplars, the sdRDM has  $N^2$  entries, so there are  $N$  diagonal entries (within-exemplars) and  $N*(N-1)$  off-diagonal entries (between-exemplars). Under  $H_0$ , the distribution of the average between-exemplar dissimilarities will therefore be narrower than that of the average within-exemplar dissimilarities. The EDI, thus, is a difference between two random variables, which have different variances and skewed distributions. The null distribution therefore need not be symmetric and can be non-Gaussian. The  $t$  test, therefore, is technically invalid as a test of the EDI.

Fig. 2 illustrates different scenarios for the distributions of the diagonal and off-diagonal means. The distributions can be symmetric or skewed; furthermore they could have the same shape (implying also the same width) or different shapes (e.g. different widths). Note that a difference between two random variables is symmetrically distributed about 0 if either (a) the variables are identically distributed (can be skewed in this case) or (b) each of the variables is symmetrically distributed about the same expected value (they can have different shapes, e.g. different variances, in this case). However, the diagonal and off-diagonal means are both skewed and non-identical (different variance), so the EDI is not symmetrically distributed under  $H_0$ . Therefore the assumptions of the  $t$  test are not strictly met.



**Figure 2: The EDI is not strictly symmetrically distributed under  $H_0$ .** EDI is the difference between the average diagonal entry (blue) and the average off-diagonal entry (red). The distributions of these averages can be skewed or non-skewed and have the same or different shapes. Note that under  $H_0$ , the two distributions have the same expected value (gray vertical line). When the two distributions are both symmetric (right panels) or when they are both of the same shape (lower panels), their difference will be symmetrically distributed about 0. To see this, consider the fact that for any two points (one from each distribution, e.g. blue X and red X), another equally probable two points exist (blue O and red O) such that their difference has the same absolute value and the opposite sign. The only case where the distribution can be asymmetric (and thus non-Gaussian) is when the two distributions are both skewed and have different shapes (upper left panel).

This paper has three aims. First, we assess the practical validity of the popular EDI  $t$  test by simulation. Second, we introduce a number of alternative tests that are valid and, unlike the EDI  $t$  test, enable single-subject inference and group-level fixed-effect inference, along with random-effects inference. Third, we compare all these tests in terms of their power to detect exemplar discriminability in real data from functional magnetic resonance imaging (fMRI). The simulations show that the theoretical violations of the assumptions of the EDI  $t$  test are minimal in practice and hence the test is generally valid. However, exemplar discriminability tests based on dissimilarity measures that account for multivariate error covariance (Mahalanobis distance and linear discriminant  $t$  values; Kriegeskorte et al., 2006; Kriegeskorte et al., 2007; Nili et al., 2014) are substantially more powerful and enable inference for single subjects.

## 2. Material and methods

### 2.1 fMRI experiment

17 participants (10 female, age range 20-38) underwent four functional runs of scanning in two separate scanning sessions. In each run, participants were presented with 24 images of real-world objects belonging to two categories with 12 objects each. Categories were changed from run to run towards ever more fine-grained categorical distinctions: animate/inanimate, faces/bodies, animal faces/human faces, and male faces/female faces. Participants were instructed to either categorize a stimulus based on the one previously shown (session one) or to complete a visual fixation task (session two). Each session also included two runs during which participants viewed retinotopic mapping stimuli (for details see Alink et al., 2013) and images depicting faces, houses, objects and scrambled objects. The analysis of brain responses to these stimuli allowed us to define regions of interest for FFA, PPA, LOC and early visual areas V1-3. Functional EPI images covering the entire brain were acquired on a 3T Siemens Trio scanner using a 32-channel head coil (32 slices, resolution = 3mm isotropic, inter slice gap = 0.75mm, TR = 2000ms). For each participant we also obtained a high-resolution (1mm isotropic) T1-weighted anatomical image using a Siemens MPRAGE sequence.

### 2.2 Tests statistics for measuring exemplar discriminability

#### 2.2.1 Exemplar Discriminability Index (EDI)

Representational similarity analysis (RSA, Kriegeskorte et al., 2008a), characterizes the representations of a brain region by a representational dissimilarity matrix (RDM). An RDM is a distance matrix composed of distances between response patterns corresponding to all pairs of experimental conditions. In cases where the responses to the same exemplars are measured in two independent data sets, a similar approach can be taken and the response patterns can be compared in a split-data RDM (sdRDM). Rows and columns of an sdRDM are indexed by the exemplars in the same order. Entries of this type of RDM are comparisons between responses to the same or different exemplars in two different data sets. Comparisons between the responses to the two measurements of the same exemplar give the diagonal entries and correspond to the within-exemplar dissimilarities. Comparisons between the responses to two different exemplars give the off-diagonal entries, which correspond to the between-exemplar dissimilarities. For N exemplars, there would be N



diagonal and  $N(N-1)$  off-diagonal entries. The exemplar discriminability index (EDI) is calculated from an sdRDM by subtracting the two averages (Fig. 1A).

The EDI is the difference of the average between-exemplar and the average within-exemplar dissimilarities. Pattern dissimilarity can be measured using various distance measures. In this paper we explore the following measures of pattern dissimilarity:

- *Euclidean distance*: For two vectors **a** and **b**, the Euclidean distance is the  $L^2$  norm of the difference vector **a-b**. It is the length of the vector that connects the two vectors (**a** and **b**) to each other.
- *Pearson correlation distance*: The Pearson correlation distance for two vectors **a** and **b** is equal to 1 minus the Pearson correlation coefficient between them. The correlation distance has a geometrical interpretation: It is one minus the cosine of the angle between the mean-centered vectors of **a** and **b** (e.g. voxel-mean centered versions of **a** and **b**).
- *Mahalanobis distance*: The Mahalanobis distance is the Euclidean distance between the two vectors after multivariate noise normalisation. Multivariate noise normalisation is a transformation that renders the noise covariance matrix between the response channels identity. In this transformation, the response patterns are normalised through scaling with the inverse square root of the error-covariance. Therefore, if we have a response matrix, **B** (Response matrix is a matrix of the distributed responses to all exemplars, with each row containing the response to one exemplar in all response channels. The size of this matrix will be number of conditions by number of response channels), we can noise-normalise it like so:

Eq. 1 
$$\mathbf{B}^* = \mathbf{B} \hat{\Sigma}^{-\frac{1}{2}}$$

where **B** is the original response matrix,  $\hat{\Sigma}$  is the estimated error variance-covariance matrix (square matrix with number of rows and columns equal to the number of response channels, e.g. voxels), and **B\*** is the response matrix formed from the response patterns after multivariate noise normalisation (number of conditions by number of response channels). If the number of response channels greatly outnumbers the number of observations per response channel,  $\hat{\Sigma}$  will be rank-deficient and hence non-invertible. To ensure invertibility, we regularize the sample covariance estimate with optimal shrinkage (Ledoit & Wolf, 2003).



## 2.2.2 Average LD- $t$

Linear discriminant analysis (LDA) could also be used to compute the discriminability values between pairs of exemplars. Similar to EDI, this method also relies on having two independent datasets for the same set of exemplars. For each exemplar pair, the Fisher linear discriminant is fitted based on the data from one of the splits. The data from the other split is then projected onto the discriminant line and the linear discriminant  $t$  value (LD- $t$ , Kriegeskorte et al., 2007, Nili et al., 2014, Walther et al., 2015) would be obtained by computing the  $t$  value for data from that pair after projection onto the discriminant line. Under the null hypothesis, the LD- $t$  is symmetrically  $t$ -distributed around zero. Therefore it is possible to make inference on mean discriminabilities across many pairs of stimuli by performing either fixed effects analysis or across-subjects random-effects tests. The LD- $t$  can be interpreted as a cross-validated and noise-normalized Mahalanobis distance (Nili et al., 2014, Walther et al., 2015). The average  $t$  value of all exemplar pairs would then be a measure of the discriminability of all exemplars.

The discriminabilities for any two pairs of exemplars are not independent (e.g. discriminabilities of the same exemplar with two different ones). The standard error of the average  $t$  value is therefore smaller than 1 (the standard error of a conventional  $t$  value), but larger than  $1/\sqrt{n}$  where  $n$  is the number of all averaged  $t$  values (one for each pair). Therefore we understand that treating the average LD- $t$  as a proper  $t$  value would be conservative (In fact the LD- $t$  values are already averaged across the folds of crossvalidation, making such a test even more conservative.).

## 2.2.3 Removing the effect of univariate activation from the dissimilarity measures

Response-pattern dissimilarities could be influenced by differences in the average activation of a brain region. Removing the contribution of the activation differences might be interesting in scenarios where only pattern differences are investigated.

The correlation distance for any pair of conditions is 1 minus the inner product of the standardized activity patterns after subtracting the regional average activation from each. To remove the effect of univariate activation from the Euclidean and Mahalanobis distance, we computed the distance measure after subtracting the response channel mean from the pattern of each condition (results from these are denoted as “average activation removed” in Figures 8 and 9):

$$\tilde{\mathbf{b}}_i = \mathbf{b}_i - \bar{b}_i \mathbf{1}$$

where  $\bar{b}_i$  is the average value of all response channels and  $\mathbf{1}$  is a 1 by number of response channels row vector containing only ones. The LD- $t$  is computed as explained in Nili et al., (2014). However, after estimating the discriminant weights from the training data, we subtract the component along  $\mathbf{1}$  from it. The reasoning is that the component along  $\mathbf{1}$  corresponds to the overall differences in activation for the two conditions, hence removing it from the weights will result in a LD- $t$  value with no contribution from activation differences. More specifically, we estimate the weight vector,  $\mathbf{w}$ , from the training data and replace it with  $\mathbf{w} - \langle \mathbf{w}, \mathbf{1} \rangle \mathbf{1}$ , where  $\langle \cdot \rangle$  denotes the inner product.

## 2.3 Tests for inference

### 2.3.1 Subject as random effect

Treating subjects as a random effect allows inference about the population from which the subjects were randomly drawn. In a random effects analysis, the summary statistic is first estimated in individual subjects. Estimates of all subjects are then jointly tested using either a one-tailed  $t$  test or a one-tailed Wilcoxon signed-rank test.

#### 2.3.1.1 One-sided $t$ test

The standard method for testing EDIs is the  $t$  test. Since there is a hypothesis about the direction of the EDI (i.e. positive EDI for exemplar information), a one-tailed test is appropriate. The  $t$  test assumes that the data come from a population that is normally distributed under the null hypothesis.

#### 2.3.1.2 One-sided Wilcoxon signed-rank test

In cases where the assumption of normality of the data seems unreasonable, one might consider using non-parametric alternatives to the  $t$  test. In this paper, we consider the Wilcoxon signed-rank test (Wilcoxon, 1945) as the non-parametric alternative. This non-parametric statistical test compares the median of its input to zero. In this test the EDIs are first ranked according to their absolute values. The difference of the ranks for the positive and negative EDIs is then computed. The p-value corresponding to the difference of the ranks is the output of the test.

Another valid nonparametric test is the sign-test (Dixon & Mood, 1946). Significance of the sign-test is merely based on the number of positive and negative values in a given sample. (given the number of positive and negative samples, the p-value would be directly computed from the Bernoulli distribution). Therefore the sign-test discards any information about the magnitude of the samples and is not considered an appropriate non-parametric alternative for *t* test in testing EDIs.

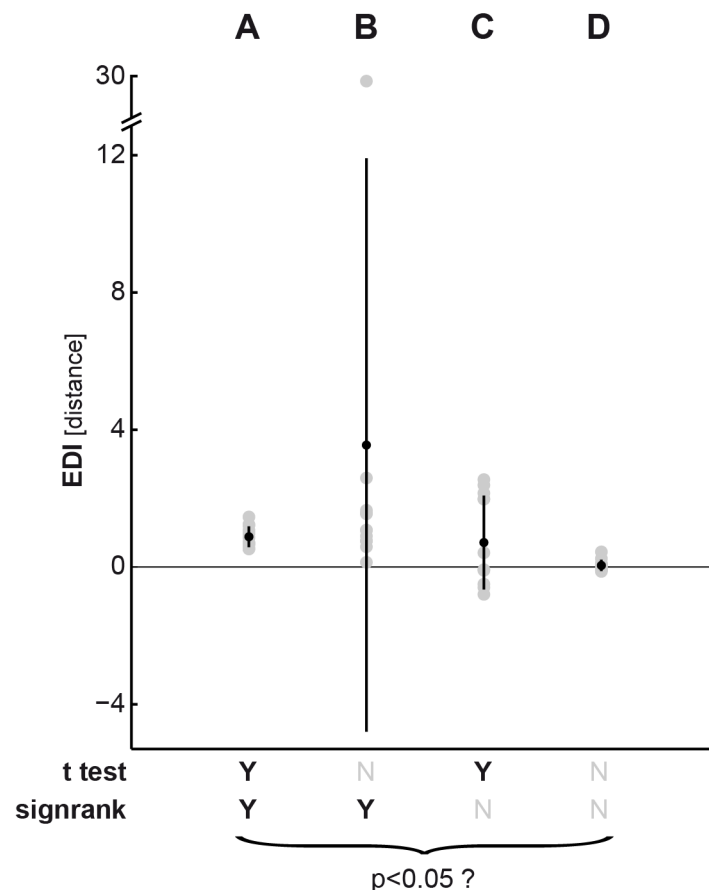
### 2.3.1.3 The *t* test and Wilcoxon signed-rank test are affected by different properties of the EDI distribution

*t* test and Wilcoxon signed-rank test base their inference on different statistics of the data distribution: while the former computes the mean and infers the standard error around it based on the sample variance, the latter computes a tail probability that relates to the sum of positive ranks of the data. These characteristics are independent from one another, therefore the tests may yield considerably different p-values depending on the degree to which each feature is pronounced in the data. To illustrate this, we simulated four sets of 12 EDIs, each by drawing random data points from four Gaussian distributions with different mean and variance (Fig. 3). We then submitted each set to a *t* test and a Wilcoxon signed-ranked test (both right-tailed and testing the null hypothesis that the data come from a zero-mean distribution) and thresholded the resulting p-values by the conventional  $p < 0.05$  criterion.

In set A, the mean of the sample is above zero and the sample variance is small; therefore, both tests yield a significant p-value. In set B, the sample mean is more positive than in set A. However, although all EDIs are positive, an outlier in the sample (value close to 30) drastically inflates the sample variance. This leads to a non-significant p-value when applying the *t* test, because the outlier increases the standard error and therefore diminishes the *t* value. By contrast, the Wilcoxon signed-rank test is not drastically affected by the outlying EDI and yields a significant p-value. The reason for this is that while the outlier has the highest rank, the rank does not contain any information on how far away this value is from the sample mean. The signed-rank test is therefore more robust against extreme outliers than the *t* test.

In set C, the variance is larger than in set A, but deviations from the sample mean are still relatively small. Therefore, the *t* test returns a significant p-value, indicating the mean EDI of the sample is different from zero, suggesting exemplar encoding. On the contrary, submitting the same set of data points to the Wilcoxon signed-rank test yields a p-value that is above the significance threshold. The reason for this is that a substantial proportion of EDIs in the

sample are negative, making the tail probability too small to result in a significant p-value. In this case the  $t$  test is the more sensitive test because it accounts for the overall sample variance rather than ranks. Finally, in set D the mean is near to zero, hence both tests do not reject the null hypothesis.



**Figure 3:  $t$  test and Wilcoxon signed-rank test may return a different p-value depending on the data distribution.** The graph depicts four different sets of simulated EDIs (A to D) sampled from Gaussian distributions with different means and variances. Gray points indicate the sample data. The superimposed black point depicts the mean of each sample. The range of the sample variance around the sample mean is depicted by black lines emanating from the mean. Each set consists of 12 values, representing 12 subjects, which are submitted to both a  $t$  test and a Wilcoxon signed-rank test. Both tests are right-tailed (meaning the tail extending into the positive direction) and test the null hypothesis that the data come from a distribution with zero mean (meaning no exemplar encoding across participants). P-values were pronounced significant (Y) if they passed the conventional threshold of  $p < 0.05$ . Otherwise, they were insignificant (N). (A) The sample mean is positive and the sample variance is small, hence both tests pass the significance threshold. (B) The sample mean is positive. However, an extreme value in the sample inflates the variance. Therefore the  $t$  test does not return a significant p-value anymore. (C) The mean is positive, but the sample contains a considerable number of negative EDIs. Accounting for the weight of the ranks of the data points, the Wilcoxon-sigend rank test penalizes this presence of negative data points, therefore its p-value does not pass the significance threshold. (D) The sample mean is close to zero, hence neither test indicates significant exemplar encoding.

### 2.3.2 Single-subject or subject as fixed effect

In clinical cases or basic research, it is sometimes desirable to perform tests at the single-subject or group-average level. For example we may want to test if familiar faces are represented distinctly in one subject or in a particular group (e.g. patients). These tests do not require generalization to the larger population from which the current subjects are sampled from. We suggest two tests for the fixed effect analysis of exemplar information.

#### 2.3.2.1 RDM-level condition-label randomisation test

Under the null hypothesis, the within-exemplar and between-exemplar dissimilarities are exchangeable. Therefore for a given split-data RDM (single-subject sdRDM or group-average sdRDM, for single-subject or group-level fixed-effect analysis, respectively), one can permute the rows or the columns of the sdRDM many times and compute the EDI at each iteration. The p-value for the non-parametric test is then the proportion of more extreme (greater) values in the null distribution compared to the EDI for the single subject or group average. So if  $N$  denotes the null distribution of EDIs and  $EDI_m$  the single-subject or group-average EDI, the p-value is obtained according to the following formula:

$$\text{Eq. 3} \quad p = \frac{n(N > EDI_m)}{n(N)}$$

where  $n()$  is an operator that counts the number of elements in a set (i.e. its cardinality).

#### 2.3.2.2 Pattern-level condition-label randomization test

In contrast to sdRDMs, the within-exemplar distances are not estimated in LD- $t$  RDMs and the exchangeability could not be applied in the same way. However, one can still estimate the null distribution of the EDI by computing the LD- $t$  RDMs under the null hypothesis for the subject or group of subjects. For this single-subject analysis, we fit the discriminant line for the training dataset (e.g. dataset 1) and test it on the remaining data (e.g. dataset 2) after condition-label randomisation (randomly shuffled test-data). We then obtain the null distribution of the average LD- $t$  values by aggregating the results from the null LD- $t$  RDMs (i.e. LD- $t$  RDMs obtained under the null hypothesis for different iterations). Fixed-effect

group-level analysis is carried out in the same way as single-subject analysis. The null LD- $t$  values are estimated for each subject and the null distribution for group-level analysis will be the distribution of subject-averaged values. Note that this test requires more computational resources than the RDM-level condition label randomisation test.

## 2.4 Scenarios for testing the statistical tests

A statistical test can lead to errors in two cases: 1) When the data comes from the null distribution and the test gives significant results (*type I error*). 2) when the alternative hypothesis holds but the test does not detect it (*type II error*). If the type I error is large, the test lacks *specificity* and is not valid. If the type II error is large, the test lacks *sensitivity* and is not powerful. Ideally we seek a test that is both sensitive and specific. In this section, we first test the validity of the underlying assumptions for the  $t$  test and then estimate the specificity and sensitivity of different EDI tests.

### 2.4.1 H0: simulation

Simulations allow us to simulate multivariate ensemble vectors with known properties. We use a number of parameters to simulate multi-normal activity patterns under H0 for two datasets in each simulated subject. For any point in the parameter space (*i.e.* any combination of parameters), we simulated patterns for a large number of subjects (10,000 subjects). The distribution of the EDIs is then a good estimate of the EDI null distribution for that particular set of parameter values. The main purpose of this rich estimation of the EDI null distribution is to assess the validity of the required assumptions for the  $t$  test. For the conventional  $t$  test approach to give interpretable results, the null distribution of the population needs to be zero-centered and reasonably Gaussian.

Figure 4 illustrates our simulation setup. For any combination of parameters, we apply three different tests to the null EDIs:

- *One-sided Wilcoxon signed-rank test*: Tests if the median of the EDI null distribution is different from zero
- *Lilliefors test of Gaussianity*: Tests if the EDI null distribution is Gaussian. The Lilliefors test (Lilliefors, 1967) is a normality test with its H0 being that the data is normal (unknown mean and standard deviation). This test is based on the maximum discrepancy between the empirical distribution function and the

cumulative distribution function of the normal distribution with the estimated mean and estimated variance.

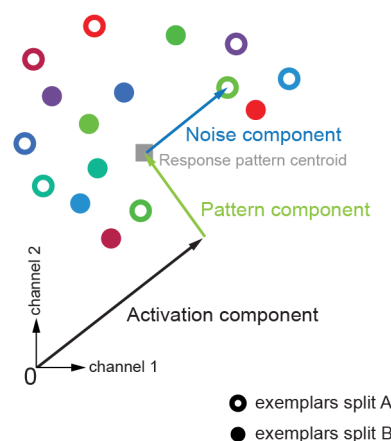
- *One-sided  $t$  test*: Tests if the mean of the EDI null distribution is different from zero. The motivation is to estimate the false-positive rate of the standard  $t$  test approach for every point of the parameter space.

As illustrated in Figure 4, the simulated multivariate responses were simulated with five parameters. All simulated patterns can be imagined as vectors in a space with as many dimensions as response channels. *the activation component* is a univariate variance component (emanating from the origin) that affects all response channels (e.g., voxels in fMRI) of all exemplars equally in both datasets. The *pattern component* is then the variance of a response channel that is added to the activation component. Here, zero pattern component variance means that the centroid is on the allones vector, i.e. the response is equal in all response channels and equal to the grand mean, hence there is no spatial variability across the simulated response channels. Conversely, a high variance means that the average pattern across splits and exemplars has great spatial variability. Adding activation and pattern component gave the response pattern centroid. Response patterns for two data splits of individual exemplars were then simulated by sampling random numbers from a Gaussian distribution with a certain *noise variance* for each split, which were added to the centroid. The noise variance parameter therefore determined the variability of the exemplars within and between repetitions. This is sensible because under the null hypothesis, all exemplars are indiscriminable.

The remaining two parameters are the *number of response channels* and the *number of exemplars*. Consider an fMRI experiment investigating whether the representations of a brain region distinguish between different face images. In that case, the number of exemplars would correspond to the number of face images that were displayed to each participant and number of response channels would be the number of voxels whose activities were recorded during the scan. Note that in practice the values of these two parameters are chosen by the experimenter/analyst. One can intuitively think that the number of exemplars may play an important role. While having more exemplars reduces the sampling error by having richer samples of the stimulus space, it also increases the gap between the number of diagonal and off-diagonal estimates of an sdRDM. Therefore one can speculate that more exemplars may not be as advantageous for the  $t$  test, because this also makes it more likely that the underlying assumptions of the test are violated. Exploring the effect of the number of response channels is also important. It would be essential to know how exemplar information in a distributed pattern depends on the number of response channels. For example in fMRI analysis, researchers often replicate the same effect for a



range of ROI sizes (e.g., Kriegeskorte et al., 2008b). This analysis helps reveal potential *special* effects for some number (or range of numbers) of response channels that are due to sensitivity of the tests or validity of the required assumptions and not due to the properties of the distributed representations.



**Figure 4:  $H_0$  Simulation settings.** We use simulations to test the distribution of EDIs under the null hypothesis,  $H_0$  (here illustrated for only two response channels). For both datasets 1 and 2, patterns of exemplars were simulated in a multi-dimensional response space. The number of dimensions of the response space is the same as the number of response channels (e.g. voxels in a region of interest in fMRI data). Response patterns were simulated by first moving along the all-ones vector to reach the average activation, whose strength was determined by the *activation component* variance. A Gaussian pattern with a *pattern component* variance was added to the activation component, resulting in the response pattern centroid. Individual exemplars of each data split were then generated by adding random Gaussian noise with a *noise component* variance to the centroid. At each iteration of the simulation, the same centroid underlay all exemplars in both data splits (as exemplars are indiscriminable under  $H_0$ ). Exemplar labels were randomly assigned to experimental conditions and their repetitions. For details of the simulations see section 2.3.1.

Exploring the parameter space at multiple simulation settings can lead to false positives due to the testing of multiple hypotheses at once (each test is repeatedly applied for all possible combinations of the parameters). To control the type-I error rate we keep the false discovery rate (FDR) at 5% (Benjamini and Hochberg, 1995). Moreover, we also report results from applying the more conservative Bonferroni correction for multiple comparisons (i.e., controlling the family-wise error rate).

#### 2.4.2 $H_0$ : simulation by shuffling fMRI data

The null hypothesis,  $H_0$ , can also be simulated from real fMRI data. We simulate  $H_0$  from data at the single-subject level and obtain group data under  $H_0$  by aggregating the simulated data from all subjects. For tests that rely on statistics obtained from an sdRDM (e.g. EDI based on Pearson correlation distance), null-sdRDMs are obtained in each subject by

permuting the rows and columns independently (note that exchangeability holds for  $H_0$ ). For tests that rely on the LD- $t$  values, the exchangeability is applied by randomly permuting the order of exemplar predictors in the design matrices of both data splits. The group aggregate is estimated for a large number of  $H_0$  iterations and each test is applied to the null data. Once the p-values are obtained, we compute the proportion of significant scenarios and compare it to what is expected under  $H_0$  (i.e. number of iterations times the number of tested scenarios times the threshold). We choose the conventional threshold of 5 %.

This procedure estimates the false positive rate (type-I error rate i.e. proportion of an incorrect rejection of the null hypothesis amongst all tested hypotheses) and allows validating the tests without thorough exploration of the parameter space (i.e., the approach explained in 2.4.1).

---

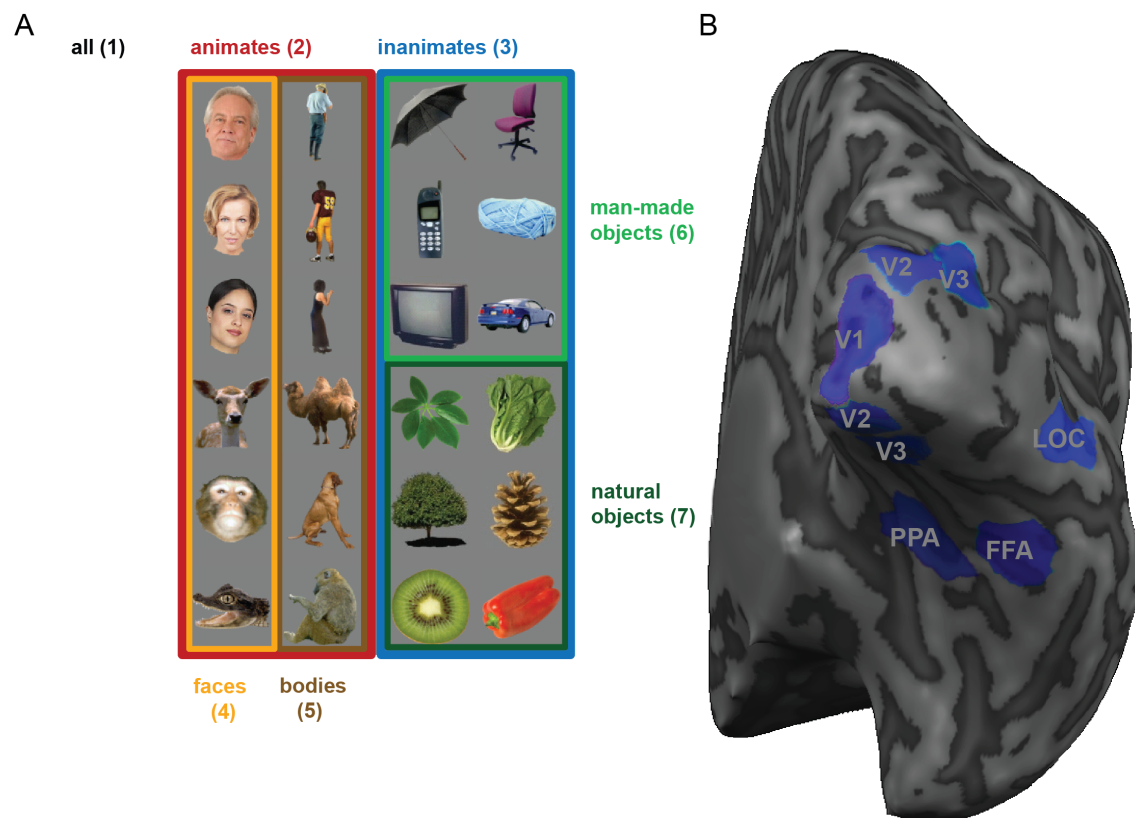
#### 2.4.3 H1: fMRI data

In order to assess the sensitivity of different tests, we apply all tests to the same data and a wide range of exemplar-discriminability test scenarios. To do this, fMRI data from six regions of interest (i.e., V1, V2, V3, LOC, FFA, and PPA) of 17 subjects were considered. In addition, we consider discriminability of different subsets of experimental conditions. Figure 5 shows the regions of interest and the different stimulus sets. Pattern dissimilarities were computed based on the beta coefficients from the GLM. The design matrix consisted of one regressor per exemplar and six motion parameters.

Having six ROIs and seven discrimination sets (42 scenarios in total) includes a range of effects. For a given threshold, the number of significant scenarios is an indicator of the power. A test that is more powerful will give more significant cases compared to a less powerful test. If an effect is very strong, all tests would most likely detect it (resulting in significantly small p-value) but a weak effect will only be detected by a sensitive test. Thus, the comparison is fair since different tests are applied to the same data and also reasonably general since a range of effects are considered.

Applying all tests to the same dataset would not allow statistical comparison of the power of different tests. For that, we need to estimate the “variability” in the power of tests as well. To do that, we employ subject bootstrapping repeating the same procedure for a large number of subject replacements. At each bootstrap iteration, tests are applied to the bootstrapped group-data and the number of significant scenarios are counted. The standard deviation of the bootstrapping distribution would be an estimate of the standard error of the mean for the actual data (Efron and Tibshirani, 1994). We obtain p-values for each pair of tests by computing the proportion of cases where the number of significant cases is different in the

two tests (e.g. for two tests  $\text{test}_1$  and  $\text{test}_2$ , the p-value for the null hypothesis that the power of  $\text{test}_1$  is greater than the power of  $\text{test}_2$  is the proportion of bootstrapping iterations in which the number of significant cases reported by  $\text{test}_1$  is less than or equal to the number of significant cases reported by  $\text{test}_2$ ).



**Figure 5: Different exemplar sets (A) and ROIs (B) considered for the analysis of real fMRI data.** In order to evaluate the different tests, we apply the tests to real fMRI data. Having data from different regions of interest (V1, V2, V3, LOC, FFA and PPA, displayed for an example subject on the right) and testing the discriminability for different exemplar sets (*i.e.* different sets of experimental conditions, displayed on the left) confronts the tests with a wide range of situations.. To obtain an estimate of the type I error, different tests were applied to randomised fMRI data (simulating the null hypothesis with real data). Seven subsets of conditions are considered: 1) all 24 images 2) 12 images belonging to the animate category 3) 12 images belonging to the inanimate category 4) 6 face exemplars 5) 6 images of bodies 6) 6 man-made objects 7) 6 natural objects.

### 3. Results

#### 3.1 Simulation results: all tests are empirically valid

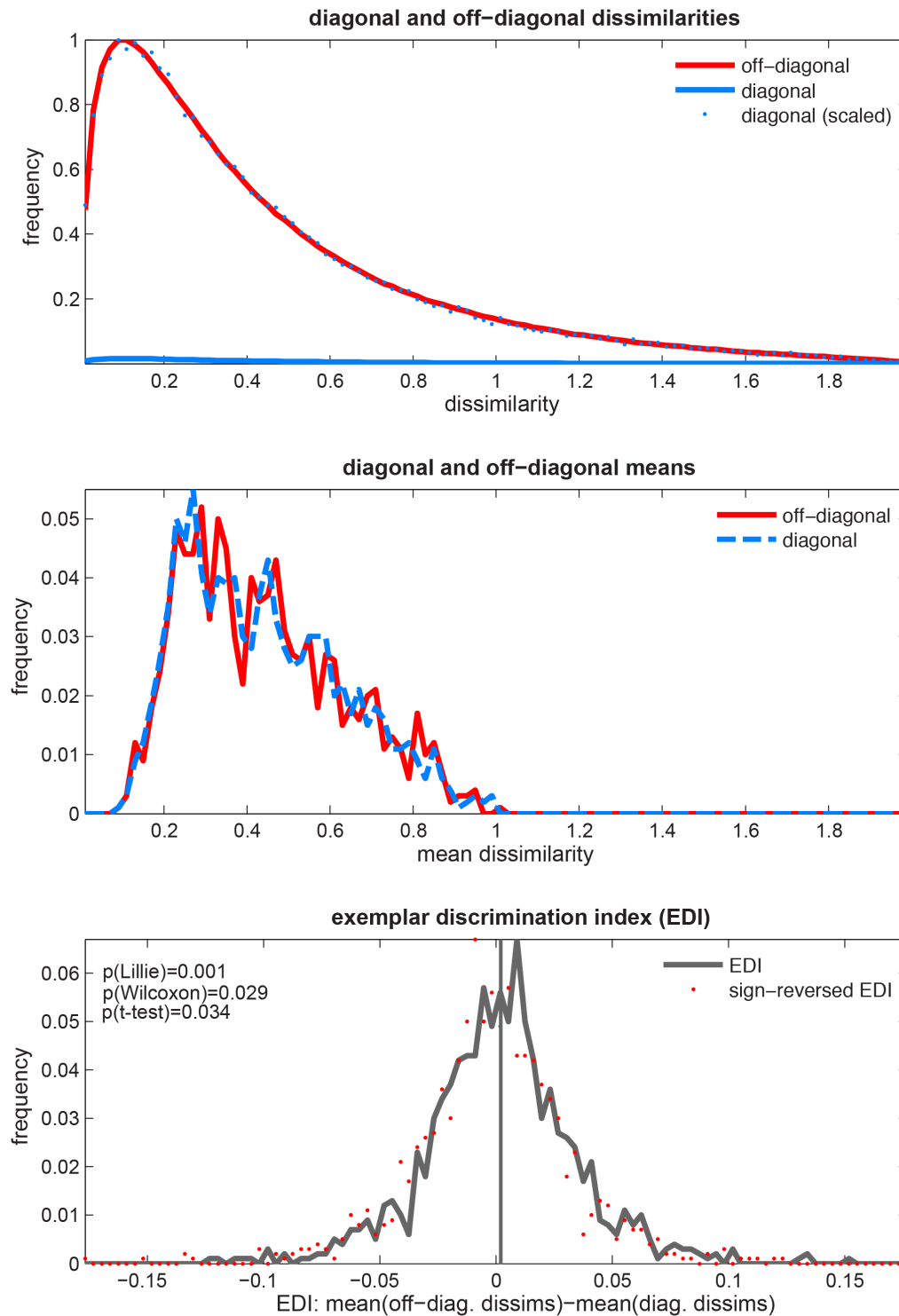
The simulations enabled us to test hypotheses about the EDI distribution under  $H_0$ . In particular, we explored a wide range of parameter settings (e.g. number of exemplars, etc.).

For each setting, we assessed the validity of the  $t$  testing approach by testing whether the null distribution conforms to the  $t$  test distributional assumptions and if the  $t$  test protects against false positives at a reasonable rate. Furthermore, we simulated  $H_0$  using real fMRI data and obtained estimates of the false-positive rate for all exemplar discriminability tests. In both cases (simulating  $H_0$  using real or simulated data), if the tests are valid, the false-positive rates will not be different from what is expected to be significant by chance (the number of false-positives depends on the threshold and the number of tests e.g. when we apply a threshold of 5 % to 100 tests of the data under  $H_0$ , we expect 5 tests to be significant).

### 3.1.1 The $t$ test assumptions are not met: the EDI is zero-mean, but not Gaussian under $H_0$

Although the EDI is not exactly symmetrically distributed about 0 under  $H_0$ , in practice it comes very close to a symmetric distribution about 0, for three reasons. (1) Though the representational distances are positively biased and have an asymmetric distribution, their distribution becomes more and more symmetrical as the dimensionality of the patterns (i.e. the number of response channels, e.g. voxels) increases. Even for very small numbers of response channels (e.g. five) the distances approximate a symmetrical distribution under  $H_0$ . (2) When dissimilarities are averaged to obtain diagonal and off-diagonal means, the distribution of each of these means even more closely approximates symmetry as it becomes more Gaussian (according to the central limit theorem). (3) The variances of the diagonal and off-diagonal means are clearly different, but this difference is smaller than one might intuitively expect. The diagonal elements are independent, so their average has a variance reduced by factor  $\sqrt{N}$ . The off-diagonal elements are dependent and although  $N^2 - N$  of them are averaged, the reduction in the variance is much smaller than factor  $\sqrt{N^2 - N}$ .

For these reasons, it is very difficult indeed to find a scenario by simulation where the EDI is not approximately symmetrically distributed about 0 under  $H_0$ . Figure 6 gives an example of a simulated extreme scenario, which we selected to illustrate the theoretical violations of the normality and symmetry assumptions. The upper two panels illustrate the approximate symmetry of the dissimilarities and their diagonal and off-diagonal means. The lower panel shows the EDI distribution, which is also close to, but not exactly, zero-mean symmetric. The EDI in this selected simulated null scenario is slightly but significantly non-Gaussian (Lilliefors test) and the  $t$  test and Wilcoxon signed-rank test are also significant.

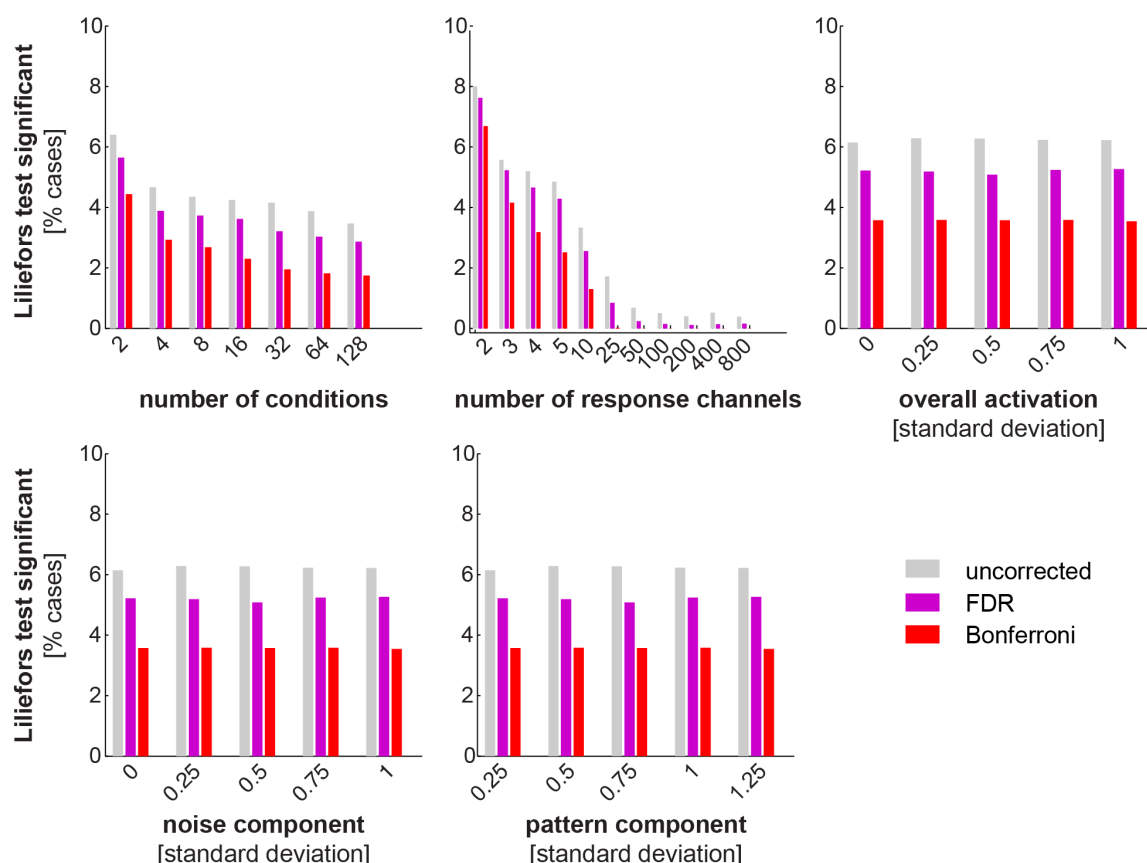


**Figure 6: The EDI can be non-Gaussian under the null hypothesis ( $H_0$ ).** *Top panel:* The distribution of the diagonal (blue) and off-diagonal (red) dissimilarities. Distributions are obtained from the dissimilarities aggregated across pairs and simulated subjects. There are fewer diagonal entries but they are sampled from the same distribution as off-diagonal dissimilarities under the null hypothesis. Therefore the two distributions differ only by a scaling factor. *Middle panel:* The distribution of the diagonal and off-diagonal average dissimilarities. Data were simulated for each subject (1,000 in total) and the distributions were obtained by pooling the averages across all subjects. *Bottom panel:* The EDI distribution in this case is significantly non-Gaussian. Furthermore, a one-sided  $t$  test or Wilcoxon signed rank test both yield significant  $p$ -values. This means that EDI is

not strictly zero-mean and Gaussian under  $H_0$ . This could potentially inflate both types of error and weaken inference based on the  $t$  test. For each of the 1,000 subjects, data were simulated for 64 exemplars and five response channels.

### 3.1.2 The $t$ test is empirically valid because of assumption robustness

In the previous section we showed that the EDI is not strictly Gaussian under  $H_0$ . Here, we used simulations to see how often these violations occur. To this end, sdRDMs were computed under the null hypothesis for a large number of simulated subjects. Simulations were carried out independently for every point of the 5-dimensional parameter space. The parameters were the number of exemplars, number of response channels (e.g. voxels) and three others characterizing the multivariate response space (see Fig. 4). At each point of the parameter space, we tested if the EDI null-distribution was Gaussian and if it was zero-centered (the two main assumptions of the  $t$  test of EDIs). Additionally, for every point, we estimated the false positive rate of the  $t$  test. Estimating the false positive rate at each point in the parameter space is carried out by independently repeating the whole simulation 1,000 times and obtaining the false positive rates from the proportion of cases where the  $t$  test is significant for a specified threshold.



**Figure 7: Marginal histograms of deviations from Gaussianity of the simulated EDI null distribution.** In order to richly characterise the properties of the EDI distribution under  $H_0$ , we simulate data for a large number of subjects for a range of values of any parameter. At every point of the cross-product parameter space, sdRDMs (based on Pearson correlation distance) were obtained from simulated patterns for 10,000 subjects (patterns were simulated independently for each subject). Lilliefors tests (testing whether the distribution of EDIs is significantly non-Gaussian) were applied to the EDI distributions. These bars correspond to results from the Lilliefors Gaussianity test. A significance threshold of 5 % was applied to the tests. Each column also corresponds to one dimension of the parameter space. The first two (number of exemplars and number of response channels) mainly depend on the experiment and analysis and the other three characterise the multivariate response space. For example, the red bar on the leftmost bar-graph would correspond to the proportion of all Lilliefors tests that were significant ( $p < 0.05$ , Bonferroni correction) when the number of conditions were fixed at 2, 4, 8, etc. and the other parameters had any possible value. Two other tests were performed in a similar way to test EDIs for the simulated data under  $H_0$ . One was a one-sided Wilcoxon signed rank test and the other a one-sided  $t$  test. The signed rank test was applied to assess whether the EDIs were zero centered and the  $t$  test was applied to obtain an estimate of the false-positive rates. Interestingly, for both tests less than 1% violations occurred ( $p < 0.05$ ) and none of those survived correction for multiple comparisons (FDR and Bonferroni,  $p < 0.05$ ).

Figure 7 shows the results for the Lilliefors test. Each panel corresponds to one of the parameters of the parameter space (5 in total, see section 2.3.1). Each bar graph gives the marginal histograms for the frequency of violations of the Lilliefors test for different levels of a parameter. Frequency of violations for the other two tests, i.e.  $t$  test and signed-rank test were close to zero for the different levels of all parameters (after controlling FDR at 5%) and

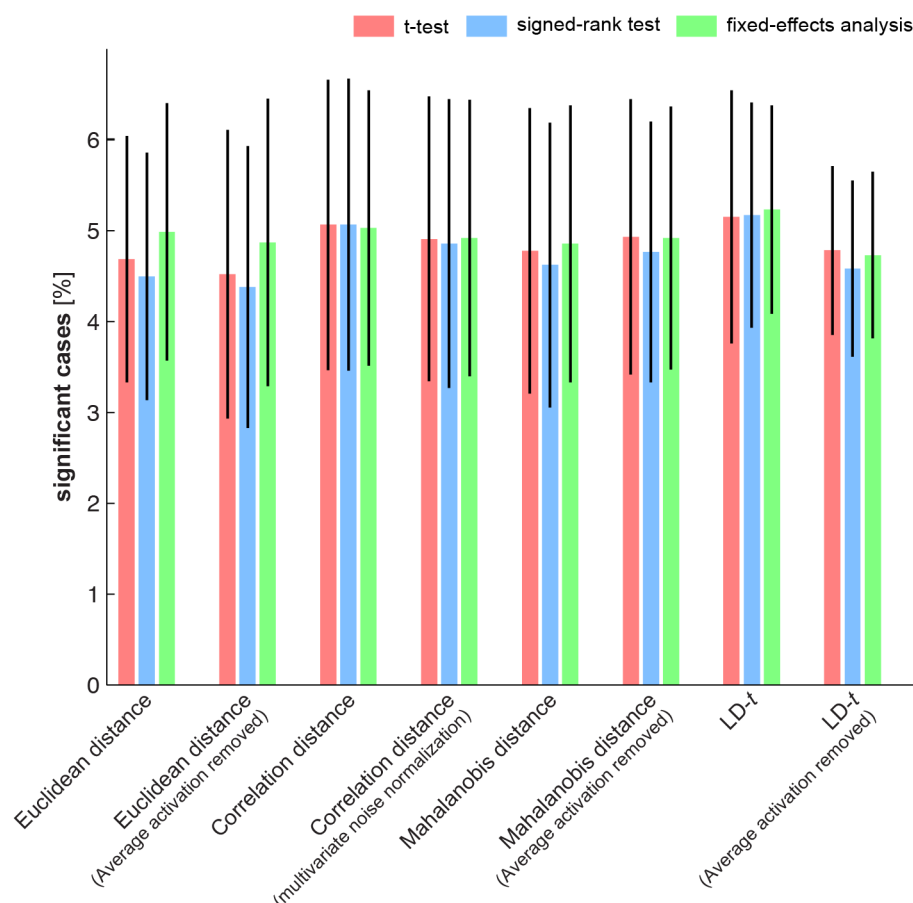


not displayed. As expected, the null distribution is very rarely centered at a point different from zero.

Importantly, in accordance with the theoretical argument for the non-Gaussianity of the EDI, cases in which the EDI null distribution was significantly non-Gaussian were not infrequent. The validity of the assumptions also seemed to interact with the parameter level (e.g. fewer number of response channels are more likely to give rise to a non-Gaussian EDI null distribution). Interestingly, despite these violations of assumptions, the false-positive rates of the  $t$  test approach did not survive the multiple comparisons in any of the tested scenarios. These seemingly contradictory results can be explained by the robustness of the  $t$  test: Though the  $t$  test relies on distributional assumptions, it is robust to assumption violations. Overall, these results suggest that the assumptions are violated in some cases but the violations are small and tolerated by the  $t$  test.

### 3.1.3 All tests of exemplar information protect against false positives at the expected rate.

Our simulations showed that the one-sided  $t$  test is a valid approach to test EDIs. Section 2.2 introduced a variety of EDI tests from which one is the commonly used  $t$  test. In this section we estimated the false positive rate (type-I error rate) of all exemplar information tests. To this end, we applied the tests to many instantiations of group data under  $H_0$ . Each instantiation was obtained by shuffling fMRI data (as explained in 2.3.2). Fig. 8 shows the false-positives rates of the different tests.



**Figure 8: False-positives rates for different tests of exemplar discriminability.** We estimated the false positive rate (type-I error rate) by applying tests to randomised fMRI data (simulating  $H_0$ , see section 2.4.2 for details). The null data were simulated for a large number of iterations. At each iteration an ROI was randomly selected and data from the ROI for every subject were randomised by permuting the condition labels for a full set containing experimental conditions from both data splits. Randomisation was independently carried out for the two data splits. Under the null hypothesis, exchangeability implies that there should not be a difference between replications of the same exemplar and a different exemplar. At each iteration (1,000 in total), we simulated group level data under  $H_0$  and applied all tests to it. We then computed the average percentage of false positives (proportion of significant cases from the total number of 1,000 simulated cases, i.e. height of the bars) and their standard deviation (error bars). All tests had a false positive rate that was not significantly different from 5% after a threshold of 0.05 was applied to the tests ( $p > 0.5$ , two-sided Wilcoxon test). This means that all tests protect against false positives at an expected rate.

As the results show, all tests have an acceptable false positive error rate: at a significance threshold of 5%, about 5% of the tests give significant results.

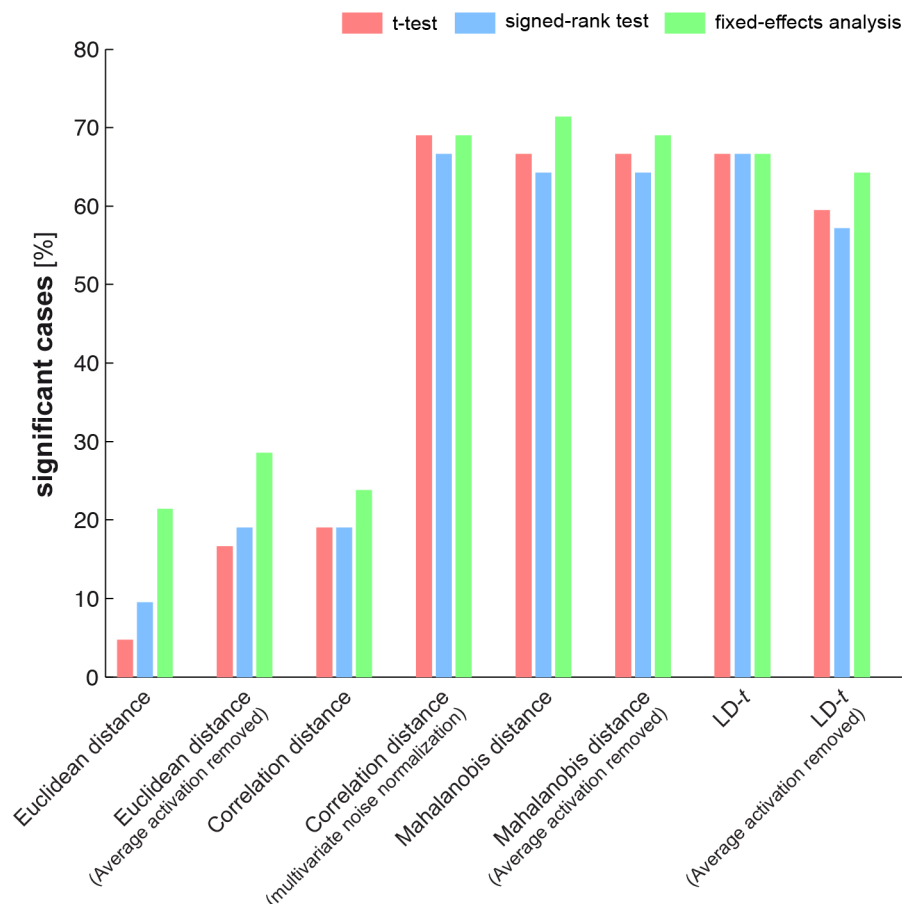
Simulating the null hypothesis using real data is more realistic than using simulations, as it incorporates all the complexities of measured data. In simulations, it can be impractical to model voxel dependencies or the extent to which response patterns change from one session to the other (different data splits). Using real data inherits all those dependencies and allows answering the same questions. However, one disadvantage of real data is that parameters of interest cannot be studied as principled as in simulations.

It must be noted that having a reasonable false positive rate is a necessary criterion for the validity of a test and if any test gives greater false positive rate than what is expected by chance, the test would not be considered further.

## 3.2 Real-fMRI-data results

### 3.2.1 Significantly greater sensitivity through multivariate noise normalization

To complete our assessment of exemplar information tests, we compared their sensitivity, i.e. the power of the tests to detect an effect when present. One way to proceed would then be to simulate data with known ground truth (i.e. exemplar discriminabilities), apply the tests to the simulated data and compare them (or their ranks) in terms of the number of significant cases. However, this approach is both computationally expensive and unrealistic. In the parameter space, each parameter can span an enormous range of values. Therefore, it would be impractical to carry out a comparison for each possible combination of parameters. Moreover, as mentioned in the last section, voxel dependencies and appropriate levels of replicability observed in real fMRI data are not known a priori. For these reasons we compared the tests by applying all of them to the same dataset. The dataset contained six brain regions and seven exemplar subsets (see Fig. 5). Therefore the total number of significant scenarios for a given test could vary from 0 to 42. Note that in order to make the test comparisons valid and fair, the same significance threshold was applied in all tests. **Figure 9** shows the results of the comparison.



**Figure 9: Comparing the sensitivity of different tests.** Each test was applied to fMRI data from six different ROIs. In each ROI, discriminabilities of seven category subsets were assessed. A test was pronounced significant if  $p < 0.001$ . The height of the bars shows the percentage of significant cases of the 42 tested scenarios (e.g. 50% corresponds to 21 significant tests). Tests on multivariately noise-normalized response patterns were significantly more powerful. Inference was done using subject bootstrap. All pairwise comparisons between the two groups of tests (i.e. the first three triple of bars and the next five triple of bars corresponding to unnormalized and multivariate noise normalized data) were significant after controlling the expected false discovery rate at 0.01.

The results show a striking effect of multivariate noise normalization in the sensitivity of EDI tests. All tests that are applied to data after multivariate noise normalization yielded a greater proportion of significant cases. Statistical comparison of the tests was carried out by bootstrapping subjects and applying the tests to the group data at each bootstrapping iteration. Obtaining p-values for comparing pairs of tests and controlling the false discovery rate at 1%, we concluded that the difference between the two groups of tests was significant. In other words, summary statistics that were obtained from data after accounting for the noise covariance between voxels had significantly greater power in detecting exemplar effects.

## 4. Discussion

This article explores different approaches to testing subtle within-category effects in brain representations.

### 4.1 Testing for exemplar information in condition-rich designs is important for understanding brain representations

Understanding brain representations asks for characterising the capabilities of the representations. One important capability would be to allow discrimination between different members of a category i.e. category exemplars. One could use a classifier and test for exemplar information by calculating the performance of a classifier that is trained to distinguish between exemplars (e.g. employing a support vector machine, Burges 1998, Misaki et al., 2010).

In contrast to the classification-based approach that requires having many repetitions of each exemplar, we aim to extract exemplar information for condition-rich designs where many exemplars are used with fewer repetitions for each. Given the limited time for measuring brain representations, there would be a trade-off between the number of tested exemplars and number of repetitions of each exemplar. Resolving this trade-off would allow higher inferential power and a stronger claim about the representational structure of a brain region. For example if we could have more exemplars *i.e.* richer sampling of the space of exemplars, inference would be based on a larger group and that would reduce the sampling error. *Condition-rich decoding* refers to quantifying and assessing exemplar information in cases where discriminability of many experimental conditions (e.g. category exemplars) are tested.

Condition-rich decoding from distributed brain representations is possible using a simple and intuitive idea: if (on average) repetitions of the same exemplar produce less pattern-change than produced by changing the exemplar, there is exemplar information in brain representations. Since this approach averages the effects across many pairs of exemplars, presence of exemplar information does not imply that any two pairs would be discriminable but that on average there is enough information to discriminate between exemplars on the basis of brain representations.

## 4.2 The conventional $t$ test approach is theoretically problematic but practically acceptable

The conventional approach for condition-rich decoding uses  $t$  test on the EDIs. The  $t$  test is a parametric test and, for a given sample, tests if the parameters of the sample distribution are different from the Gaussian distribution obtained under the null hypothesis. Using parametric statistics can be more powerful when the required assumptions are satisfied. However, if the requirements are not met, test results are not interpretable. For reasons that we explain in the paper, the null distribution of the EDI is likely to violate the required assumptions of the  $t$  test (e.g. violation of the Gaussianity assumption). Therefore, we thought that  $t$  test would be problematic. Using simulations, we first appreciate this concern by observing cases where applying a  $t$  test is wrong. This was then followed by a systematic exploration of the space of possible parameters (characterizing the response space in a wide range of settings) and showing that violations of assumptions are not frequent and to the extent that in most cases could be tolerated by the test (although the  $t$  test presumes distributional properties, it is tolerant to violations of the assumptions).

## 4.3 Testing exemplar information at the single-subject level or group-level with subject as fixed effect is possible using novel randomization tests

Randomization tests, and more generally non-parametric statistics, are becoming more and more popular in the analysis of neuroimaging data (Nichols and Holmes, 2002). We can also use randomization tests and non-parametric statistics in the context of testing for information in brain response-patterns. For example, randomization tests have been proposed before to test for the similarity (i.e. correlation) of two RDMS (Kriegeskorte et al., 2008b, Nili et al, 2014). Here, randomization techniques could be employed to design non-parametric tests of exemplar information.

Hitherto, it has not been possible to test the exemplar information at the single-subject level or to do group-level analysis with subject being the fixed effect. Here we introduce tests for performing fixed effect analysis. To our knowledge this is the first attempt to fixed effect analysis of exemplar information using pattern-information analysis. Our tests are based on randomization methods that estimate the distribution of the statistic under  $H_0$ . For group-level analysis, these tests are usually more sensitive since they ignore the between-subject variability.

#### 4.4 Tests that take the covariance structure of the noise into account are significantly more powerful

In the past, only few studies have attempted to do condition-rich decoding. All those studies were based on quantifying pattern-changes by computing Pearson correlation distances and using a one-sided  $t$  test to test the net pattern-effect (average effect of changing exemplars minus average effect of repeating an exemplar) quantified in each subject, against zero. The  $t$  test is widely used in psychology and neuroscience; however, its use is less motivated for testing exemplar information in distributed patterns. The main reason against using  $t$  test is the theoretical concern about the distribution of the test statistic (i.e., EDI) under the null hypothesis.

Having established the validity of the standard approach, we consider different ways of testing exemplar information by exploring different tests and test statistics. In particular we introduce ways of testing exemplar information at the single-subject level or fixed effects analysis at the group level. In addition to the fixed effect tests, we also consider other tests including non-parametric alternatives to the  $t$  test. Those distribution-free methods can be used for testing exemplar information even in the extreme cases where  $t$  test is not valid.

The various test statistics considered in this paper are different due to their different ways of quantifying pattern-effects. In particular different measures could be used for computing pattern dissimilarities. Therefore, the situation is as follows: there are different ways of assessing exemplar information. The net effect could be *summarized* and *tested* in different ways. Interestingly, we see that all the possible combinations of tests and test statistics yield reasonable false positive rates. In other words, choosing any of the proposed ways of summarizing the results and any of the tests would result in reasonable specificities. However, when looking at the sensitivities, we did not find equal levels of power for the different combinations.

By contrast, we observed that within the explored tests (i.e.  $t$  test, Wilcoxon signed rank test and fixed effect tests) and test statistics (i.e. EDIs based on different distance measures or average LD- $t$  values), the most important factor in determining the power of the test was the way to quantify and summarize the net effect, and not the test itself. All pattern-effects that were computed after multivariate noise normalisation could be detected with a greater power than those that were not. This effect of multivariate noise normalisation afforded an almost 3-fold boost of power for the analysed dataset. Therefore, these results have clear practical implications for future studies. For example, researchers could add multivariate noise normalization to the pre-processing stages of data analysis and use the Euclidean distance -



which has a simple geometric interpretation – to estimate pattern dissimilarities on the pre-processed datasets. This would have the advantage of being both simple and powerful.

#### 4.5 Quantifying category information with the category discrimination index (CDI)

Although category information and exemplar information consider conceptually different characteristics of representational geometries, they could both be assessed using similar methods. In this paper, we focus on exemplar information. Category information could also be quantified from an RDM by subtracting the average-within-category dissimilarities from the average-between-category dissimilarities *i.e.* category discriminability index (CDI). The CDIs could then be tested in exactly the same way as EDIs. The main difference is that for CDI there is no need to split the data in two halves. One could use the whole dataset and compute an RDM from the full dataset. Using more data (*i.e.* the whole dataset as opposed to data-halves) may likely result in more stable patterns, which would be an advantage of this approach.

Another approach to testing category information is to test the rank correlation of RDMs with a categorical RDM that has equal and smaller values (*e.g.* zeros) for within-category dissimilarities and equal and larger values (*e.g.* ones) for between-category dissimilarities (Khaligh-Razavi & Kriegeskorte, 2014). We have recently proposed using Kendal's tau-a for RDM correlations in such cases where tied ranks are predicted for either between- or within-category dissimilarities (Nili et al., 2014).

An alternative approach is to fit categorical RDMs to brain RDMs using linear regression and test the regression coefficients against zero (Mur et al., 2013). LD-*t* RDMs also allow direct inference on the categorical structure. Category information could be tested by testing the average-between-category LD-*t* values (see Cichy, Pantazis and Oliva (2014) for a similar approach using SVM). A stricter test is to test category *clustering*. That is possible by testing average between-category effects after subtracting average-within-category LD-*t* values.

#### 4.6 Relation to model testing

Representational similarity analysis enables testing computational models of information processing in the brain. This is achieved by comparing the predicted representational geometries of models with observed geometries of the brain. Computational models also imply exemplar information but only test for a particular representational geometry. Therefore testing a model is a more focused test which would be more powerful if the hypothesis is true.

Note that the EDI is maximally general in that it is sensitive to any differences between conditions. It is highly powerful when differences exist for many of the pairs, but in principle it can detect a difference between just one pair of exemplars. Note also that the EDI can detect equidistant representational geometries, to which tests of RDM models using correlation coefficients are not sensitive.

Additionally, one could use sdRDMs to test computational models. The model prediction would be a symmetric matrix that predicts zeros along the diagonal entries. This would enable us to test the model predictions of the dissimilarities at the level of a ratio scale, not just an interval scale. It would increase the power when the model predicts very similar distances among the representational patterns. (Consider the extreme case where the model predicts equidistant patterns. In this scenario the sdRDM is required, and the model test would be equivalent to the EDI if the Pearson correlation was used).

## 5. Conclusions

When independent measurements of the same set of exemplars are available, condition-rich decoding can be carried in different ways. Despite the theoretical concerns against applying the  $t$  test to the EDIs, we validated the use of the  $t$  test, which is the established approach for testing exemplar information. Furthermore, we explored different ways of summarizing and testing exemplar discriminabilities and introduced a novel randomization test that is assumption-free and also allows testing EDIs at the single-subject level or group level analysis with subject as fixed effect. Comparing the different methods, we conclude that it is mostly critical to compute the EDI after applying multivariate noise normalization to the response patterns. Multivariate noise normalization considers the noise variance-covariance structure of the data and enables a more reliable estimation of RDMs. This means that computing the EDIs after multivariate noise normalization or using the average LD- $t$  will likely result in higher sensitivity to detect exemplar information in brain representations.

## References

- Anzellotti, S., Fairhall, S. L., & Caramazza, A. (2013). Decoding representations of face identity that are tolerant to rotation. *Cerebral Cortex*, bht046.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2), 121-167.

Chan, A. W., Kravitz, D. J., Truong, S., Arizpe, J., & Baker, C. I. (2010). Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nature neuroscience*, 13(4), 417-418.

Dixon, W. J., & Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association*, 41(236), 557-566.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425-2430.

Henriksson, L., Khaligh-Razavi, S. M., Kay, K., & Kriegeskorte, N. (2014). Intrinsic cortical dynamics dominate population responses to natural images across human visual cortex. *bioRxiv*, 008961.

Hung, C.P., Kreiman, G., Poggio, T., DiCarlo, J.J., 2005. Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science* 310: 863–866. doi:10.1126/science.1117593

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5), 679-685.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, 10(11), e1003915.

Kiani, R., Esteky, H., Mirpour, K., Tanaka, K., 2007. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology* 97, 4296–4309.

Kravitz, D. J., Kriegeskorte, N., & Baker, C. I. (2010). High-level visual object representations are constrained by position. *Cerebral Cortex*, 20(12), 2916-2925.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10), 3863-3868.

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, 104(51), 20600-20605.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008a). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... & Bandettini, P. A. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6), 1126-1141.

Kriegeskorte, N., & Kreiman, G. (Eds.). (2012). Visual population codes: toward a common multivariate framework for cell recording and functional imaging. MIT press.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8), 401-412.

Kruskal, J.B., Wish, M., n.d. Multidimensional Scaling. 1978. Beverly Hills, CA.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5), 603-621.

Lee, S. H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. *Neuroimage*, 59(4), 4064-4073.

Liu, N., Kriegeskorte, N., Mur, M., Hadj-Bouziane, F., Luh, W. M., Tootell, R. B., & Ungerleider, L. G. (2013). Intrinsic structure of visual exemplar and category representations in macaque brain. *The Journal of Neuroscience*, 33(28), 11346-11360.

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage*, 53(1), 103-118.

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social cognitive and affective neuroscience*, 4(1), 101-109.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in psychology*, 4.

Nestor, A., Plaut, D. C., & Behrmann, M. (2011). Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences*, 108(24), 9998-10003.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1), 1-25.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4), e1003553.

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature neuroscience*, 17(3), 455-462.

Sayres, R., & Grill-Spector, K. (2008). Relating retinotopic and object-selective responses in human lateral occipital cortex. *Journal of Neurophysiology*, 100(1), 249.

Schwarzlose, R. F., Swisher, J. D., Dang, S., & Kanwisher, N. (2008). The distribution of category and location information across object-selective regions in human visual cortex. *Proceedings of the National Academy of Sciences*, 105(11), 4447-4452.

Torgerson, W.S., 1958. Theory and methods of scaling. Wiley.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2015). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*.

Wilcoxon, F., 1945. Individual Comparisons by Ranking Methods, Biometrics Bulletin. Vol1, No 60, 80-83.

