Covariance matrices of fMRI scans yield information about a person. Attempts have been made to establish the similarity between two scans of the same person at different points in time. To find the similarity, previous works have calculated the elementwise correlation between entries of covariance matrices at two time points. If the highest correlated matrix of person x at time 1 is that of person x at time 2, then this algorithm is considered to have correctly identified person x. The *id rate* of this method is defined as the percent of subjects correctly matched between time points.

This work addresses two improvements in identifying individuals from their scans. First, the id rate is dependent on the size of the sample studied. Id rate is the probability that the most similar matrix to subject *x*’s scan at time 1, out of all scans at time 2, is subject *x*’s scan at time 2. Using subscripts to denote the time index when the correlation matrix was collected, id rate for a sample may be defined as

where is a chosen measure of similarity between two correlation matrices.

For large samples, id rates will likely degrade due to the increased density of data points. Id rate is sample dependent, and it is difficult to compare id rate measures across different samples. Even when there is no discernable relationship between correlation matrices taken at different time points, by random chance one would expect an id rate of for a sample of size .

For samples where a relationship exists between obtained matrices at different time points, it is not clear how much id rate changes as a function of n. For instance, if a sample of 100 subjects has an id rate of 90% and a second sample of 200 subjects has an id rate of 80%, it is not clear how much of this difference is due to sample size vs. other sample characteristics, such as the task under which the data was collected, or demographic patterns.

Second, it is not necessarily true that correlation between matrices is the best measure of similarity. Similarity between matrices can be calculated in many ways, several of which are discussed in the following sections. An optimal similarity measure will give less emphasis to those variations found between repeated scans of the same person and will give more emphasis to those features that are consistently different between subjects.

To solve the first problem, we seek an alternative measure of identifiability that is independent of sample size. More formally, if there is a sample A with scans of a cohort at two time points, and the similarity measure equals *k*, then the expected similarity value for a random subset B should also equal *k* (so long as the subset contains at least two subjects).

In other words, we seek a measure that is unbiased. Note that id rate does not necessarily satisfy this property. Subsets of a sample will, on average, have higher id rates. As an alternative, we propose the *pairwise id rate*, which is the rate at which a randomly chosen individual *x* will be correctly identified when compared against another randomly chosen individual . Pairwise id rate is

This definition leads into the second problem, which is determining the appropriate function s for calculating similarity. Previous studies of id rate have used the correlation between corresponding elements of matrices to calculate s. Here we will discuss some alternatives.

**Metrics to Establish Similarity**

A similarity measure is a function between pairs of objects that produces a rating of how alike the two objects are. This function should satisfy some constraints so that it does not violate an intuitive understanding of similarity. For example, for any object x, the most similar object should be itself, and so it should be the case that for any object y.

Similarity has much in common with the idea of distance. Specifically, similarity and distance are inversely related. If two objects are close to each other according to some idea of distance, then there is a corresponding similarity measure that rates these objects as being highly similar. Any strictly decreasing function of distance may be interpreted as a similarity measure.

A metric is a function between pairs of points that makes explicit the idea of distance. Multiple metrics may be applied to the same set of objects, and each metric may endow the set with a different interpretation. For instance, consider calculating the distance between the Earth’s north and south poles. If paths through the center of the Earth are allowed, then the standard Euclidean metric, where the shortest path is a straight line, gives the distance, which is about 8,000 miles. However, if paths are restricted to the surface, then the shortest path is any of the lines of longitude, and the distance is about 12,000 miles. If paths are again restricted to roads that can be traveled on, then the distance becomes infinite, since it is not possible to drive from one pole to the other. Each of these metrics is closely related to the paths that are available when going from one pole to the other.

While every metric produces a similarity measure, one must be careful that the measured distance fits with the expected notion of similarity, otherwise the constructed similarity measure will make no sense. As an example, two houses located right next to each other may look very different. To rate visible similarity, it would not be reasonable to use distance in physical space as a corresponding metric; a distance measure over an abstract space of visual properties would be needed.

Metrics may establish such abstract distances. For instance, consider two strings of bits, 0011 and 1101. The minimum number of bits that would need to ‘flip’ from 0011 to create 1101 is known as the Hamming distance. For these two bitstrings, the Hamming distance is three since they differ at three locations. A decreasing function of to the Hamming distance would grow with the number of bits these strings have in common. So, the Hamming distance as a metric corresponds to the number of bits in common as a similarity measure.

To sum up, an inverse function of a metric can be interpreted as a similarity measure, but care must be taken to show that the metric corresponds to the correct idea of similarity. Metrics are constrained by the set of possible paths that may be taken through a space, and in particular metrics govern the shortest paths between points, known as geodesics.

**Metrics over the Space of Covariance Matrices**

It would not be possible to list all metrics over the space of correlation matrices here. So, this discussion will focus on three broad categories of metric. The first category treats the matrix as a vector, and the second as a descriptor of a multivariate normal distribution.

**Matrices-as-Vectors and their Metrics**

One way to view a matrix is as a list of values that may be concatenated into a vector. For a correlation matrix of size , this would be a vector of length . And, given that a correlation matrix must have 1 on the diagonal and must be symmetric, many of these values may be omitted from calculations due to their redundancy. The resulting upper triangular matrix has entries, which is the length of the vector of interest.

**Correlation and Angular Distance**

Given two vectors a and b, the cosine of the angle between them equals . The angular distance forms a metric, and so for forms a similarity measure due to the cosine’s downward trajectory in this region. This formula, , is commonly known as the cosine similarity. It is closely related to correlation and is exactly equal to the correlation coefficient if the means of a and b are zero. The distance metric corresponding to the correlation similarity is .

**The Fisher Transform**

Fisher noted that if two normally distributed variables are highly correlated, then the variance in the measured correlation between them is much smaller than it would be if the variables were completely uncorrelated. In other words, for uncorrelated variables, errors in the measurement of *r* tend to be large, and for highly correlated variables, errors are small. This difference may produce heteroscedasticity when comparing measured *r* values, which may violate assumptions of statistical tests. Fisher proposed a transform of r values, , that keeps the variance of z approximately constant for all values of r. The following elementwise metrics are tested both with and without Fisher-transformed r values.

**Euclidean Distance**

The best-known metric, Euclidean distance, is .

**Manhattan Distance**

The Manhattan distance, also known as the taxicab or L1 metric, is the sum of absolute deviations between two vectors: . The name comes from the street grids in Manhattan to which taxicabs are constrained.

**Chebyshev Distance**

The Chebyshev distance is the maximum distance between vectors over any of the coordinates: . In two dimensions it is the number of moves a king needs to make to travel between two squares on a chess board.

**Matrices-as-Multivariate-Gaussians and their Metrics**

Each correlation matrix C with positive determinant corresponds to a multivariate normal distribution centered at the origin with probability density function . Here, is the determinant and is a column vector representing a point in n-dimensional space. In terms of fMRI data, this pdf corresponds to an estimate of the probability density that the measured activation vector will equal at a randomly chosen time point during the task under which the data was collected.

In viewing the correlation matrix this way, C describes the set of likely activation vectors over the course of a task. Differences between measured C values correspond to changes in the statistical distribution of activations over time. The following metrics measure distances between these distributions.

**Earth Mover’s Distance and Wasserstein Metric**

Imagine that a distribution *p* is a pile of dirt, and the goal is to move this dirt into the shape of another distribution *q*. The Earth Mover’s Distance is defined as the minimum cost it takes to move the dirt from *p* to *q* if the cost of moving a speck of dirt is a function of how far it moves. If the speck-movement-cost is proportional to the standard Euclidean distance, then the Earth Mover’s Distance between *p* and *q* is known as the Wasserstein metric. “Close” distributions under the Wasserstein metric are those where much of the spread of probability is overlapping.

For normal distributions *p* and *q* with the same mean and covariance matrices and , the Wasserstein metric is .

**Fisher-Rao Metric**

A problem in many statistical questions is to determine how much information about a parameter (e.g., the true probability of heads when flipping a weighted coin) is available from an observation (e.g., the result of a single coin flip). This quantity is relevant for frequentist, Bayesian, and minimum description length formulations of statistics problems. The Fisher information gives a method for calculating this quantity. It is based on a measure of how sensitive the likelihood function is to changes in .

When a model has several parameters, such as a multivariate normal distribution, the Fisher information provides a method to calculate how much information is lost as a point moves through the parameter space. The minimum information difference between points is a natural metric in the space of parameters, known as the Fisher-Rao metric, and it is commonly used in the field of information geometry.

For normal distributions *p* and *q* with covariance matrices and , the Fisher-Rao metric is , where are the eigenvalues of .

**Numerical Stability Considerations**

Many correlation matrices from the samples studied in this paper have determinants that are very close to 0. As a result, calculating eigenvalues and inverses of matrices can be numerically unstable. Analyses in this paper rely on the numerically stable singular value decomposition algorithm (SVD) for eigenvalues, and on direct division of matrices rather than multiplying by a computed inverse (A/B instead of A \* B-1).

Additionally, some candidate metrics were eliminated due to requiring division by the determinant. These metrics produced unreliable results that were rarely better than chance. For instance, the KL-divergence (and all other f-divergences) and the Hellinger distance, which are commonly used measures of distance between probability distributions, and the Mahalanobis distance, which measures the length of a vector corrected for covariance in a sample, did not produce sensible values. It may be possible to improve the utility of these other metrics by, for example, reducing the dimensionality of the data prior to calculating distances. We leave this extension to future research.

**Data**

The IMAGEN and MLS data sets