Representational fMRI analysis: an introductory tutorial

Alexander Walther[1,2], Naveed Ejaz[2], Nikolaus Kriegeskorte[1], Jörn Diedrichsen[2]

1. MRC Cognition and Brain Sciences Unit, Cambridge
2. Institute of Cognitive Neuroscience, University College London

Jörn Diedrichsen
Institute of Cognitive Neuroscience
17 Queen Square
London
WC
United Kingdom

Abstract

Multivariate pattern analysis has become an important approach in the analysis of fMRI data, as it allows inferences - not only about the relative size of activation in different regions - but also about their representational content. The field currently uses a wide variety of multivariate analysis techniques, including classification, representational similarity analysis, MANOVA, and pattern component modeling. This makes it very hard for the novice to judge the relative strengths and weaknesses of these approaches, and to understand their deeper relationship. Here, we present a generative model framework, which unifies various approach and analysis techniques under a single roof, and provides a set of simple, yet powerful, ways of testing representational hypotheses on distributed activity patterns.

1. Introduction

What is the difference between multivoxel pattern analysis (or representational analysis) and traditional fMRI analysis? Let us start with a simple example: the question of how single fingers are represented in primary motor cortex. The traditional approach would be to measure the cortical activity of a number of subjects while they move each of the digits. We would then align these brains to each other using a normalization algorithm that superimposes these brains based on the anatomical structure and then average the activity patterns for each of the fingers. Statistical inferences would then be drawn on the resultant group activity maps. Using such analysis we indeed can see a orderly spatial arrangement of the fingers from thumb to pinkie (Fig. 1a), replicating earlier results (Indovina and Sanes, 2001; Wiestler et al., 2011).

However, when we look at the activation patterns for each individual subjects (Fig. 1b), we can see striking inter-individual variability in the spatial shape and arrangement of these patterns. The differences in activation patterns between subjects are usually larger than the differences between fingers of the same subject. Indeed, we can estimate that only 20% of the reliable aspect of the activity pattern in M1 is explained by consistent spatial arrangement that is shared across subject, with the remaining 80% being reliable, but idiosyncratic to each individual subject. Averaging activation patterns in a group space simply destroys these details. In contrast, representational analysis is interested in *relationship* between the patterns *within each individual*. For example, the patterns for the ring and middle finger are, within each subject, quite similar to each other, while the patterns for the thumb assume a very distinct shape. To quantify this, one could ask how well we can

classify between two finger based on the patterns (Haxby et al., 2001)or by calculating a distance metric between these patterns (Kriegeskorte et al., 2008). The distances can then be arranged in a representational dissimilarity matrix (see Fig. 1c), which shows the dissimilarity for each pair of fingers. The comparison of these matrices across subjects reveals that this representational structure, i.e. the way in which these patterns are arranged relative to each other, is highly invariant across individuals, even though the actual underlying activity patterns are much more variable. Representational fMRI analysis exploits this invariance by making inferences about characteristics of this representational structure, rather than directly about the more variable activation patterns.

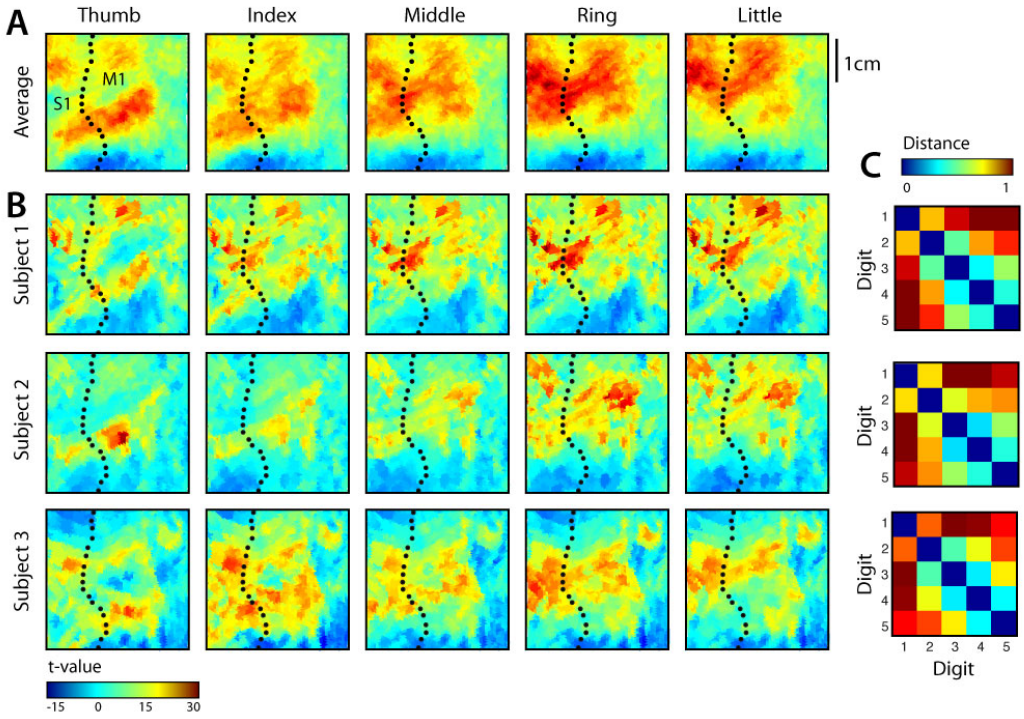

Figure 1. *Finger representations in primary motor cortex. (**A**) Group-average activity patterns shown on a flattened representations of a small area (~3x3cm) of primary motor cortex around the hand knob. The dotted line indicates the fundus of the central sulcus with the anterior bank to the right. (**B**) Activity patterns for three exemplary participants. (**C**) Representation dissimilarity matrices for each of these participants.*

2. A generative framework

A large number of different technique to quantify the representational structure of fMRI activity have been employed, including support vector machines (Ben-Hur et al., 2008; Misaki et al., 2010), linear discriminant analysis (Duda et al., 2001),

correlations , Mahalanobis and other pattern distances {}, multivariate analysis of variance (Kriegeskorte et al., 2006), canonical correlation analysis (Friston et al., 2008), and many others. All these measures get ultimately at the very same question, namely how different individual patterns related to each other. To understand their relationship it is useful to first a model of how the data came about – a so-called generative model.

The core is a simple linear model, as used in traditional fMRI analysis (Fig. 2). The data (**Y**) is the product of a design matrix (**Z**) that contains the experimental design times the patterns plus some additive noise. The k^{th} row of the pattern matrix **U** contains the true activity pattern for the k^{th} condition across all voxels. There are two main differences to the normal univariate linear model analysis: First, we consider the correlation between voxels, both in the true patterns **U** (Σ_U) and in the noise (Σ_ε) – thus or model becomes truly multivariate. Secondly, we consider the patterns (**U**), which are estimated as the regression coefficients of our linear model, to be a random, rather than a fixed effect. This means that the activation values for each condition and each location are not thought to have a true value, which we test statistically, but rather that the activation is a random variable with a distribution across voxels. Given the high inter-subject variability of the spatial activity pattern and their randomly looking nature, this seems to be a natural choice. The consequence of this statistical viewpoint is that we are interested in some measure of the distribution of these patterns, rather than their mean value.

In the following, we will show here that the second moment on the true patters $\mathbf{G} = \mathbf{U}\mathbf{U}^T$ is the quantity of interest. The i,j^{th} element of the matrix G is simply the inner product of the i^{th} and the j^{th} pattern $\mathbf{G}_{i,j} = \left\langle \mathbf{u}_i, \mathbf{u}_j \right\rangle$. If we had subtracted the mean value (across all voxels) from each pattern, **G** would be proportional (with a constant *P*) to the true variance-covariance matrix of **U**. So, generally, we can think of the diagonal as the variance of the patterns, and about the off-diagonals as the covariances. Usually, however, we want to preserve mean activation differences between conditions, as these are clearly also neurally meaningful.

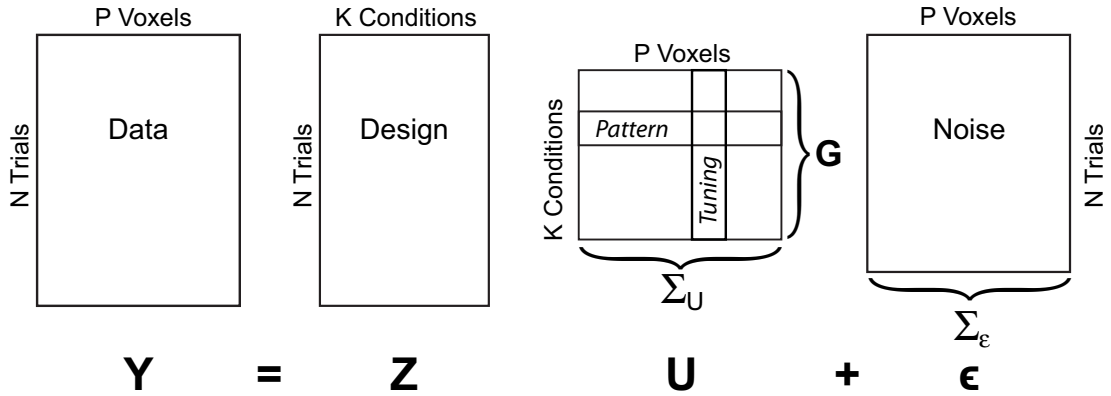

Figure 2. *Generative framework for representational fMRI analysis. The data* (**Y**) *consists of activity estimate for N trials for P voxels. In the simplest case, the design matrix* (**Z**) *simply contains a 1 if trial N belonged to condition K and a 0 otherwise (for more complicated designs, see x). In this setup, each row if the matrix of regression coefficients* (**U**) *simply shows the mean activity pattern (across voxels) for condition K. Each column shows the activity for a single voxel across all conditions, i.e. indicates the tuning of the voxel. The noise is thought to be independent across trials, but dependent across voxels.*

In the following we will show how an estimate of **G** plays a pivotal role in most kinds of representational analyses, including representational similarity analysis, pattern component modeling, corrected correlations, and crossvalidated MANOVA. So in the next section we will show the relationship between distances estimates and **G** and discuss issues in the estimation of **G** from the data.

3. Estimating distances and G

The first summary statistics of G that turns out to be universally useful is the Euclidean distance between the activity patterns of two conditions. This distance is simply the logical extension of the spatial distance between 2 points on a plane to the high-dimensional voxel space. Each pattern can be thought of as a point in a space, whose axes are formed by the activation value of each of the considered voxels. The squared distance between two patterns is:

$$d_{i,j}^2 = \left(u_i - u_j\right)\left(u_i - u_j\right)^T = \left\langle u_i - u_j, u_i - u_j \right\rangle \qquad \text{Eq. 1}$$

If two activity patterns are identical, then the distance between them becomes zero. Furthermore, by comparing distances between different pairs of stimuli allows us to make inferences about their relative similarity. For example the pattern for digit 1 in Fig. 1c is consistently more different from the digit 3 than from digit 5. This latter characteristic forms the foundation of representation similarity analysis (RSA, see section x, Kriegeskorte and Kievit, 2013). The squared Euclidean distance between two activity patterns $\mathbf{u}_i$ and $\mathbf{u}_j$ can also be expressed as a function of the second moment matrix **G**:

$$d^2_{i,j} = \langle \mathbf{u}_i, \mathbf{u}_i \rangle + \langle \mathbf{u}_j, \mathbf{u}_j \rangle - 2\langle \mathbf{u}_i, \mathbf{u}_j \rangle$$
$$= \mathbf{G}_{ii} + \mathbf{G}_{jj} - 2\mathbf{G}_{ij} = \mathbf{c}\mathbf{G}\mathbf{c}^T \qquad \text{Eq. 2}$$

Where **c** is a contrast vector with a 1 in the i^{th} position and a -1 in the j^{th} position. Thus, a distance matrix is a certain linear contrast on the **G** matrix, and can be always computed once we estimate **G**. The opposite is not true: **G** can't be recovered from a matrix of distances. The difference between them is that **G** also expresses on the diagonal the distance of each pattern from the origin (i.e. rest). By taking differences, this information gets lost. This is why it is often useful to use **G** as a sufficient statistics for the representational structure. Most measures that can be calculated on **G**, can also be calculated on the distance matrix.

Prewhitening

Noise in fMRI data shows spatial structure. First, different voxels show different noise levels – some will lie closer to the vascular supply and show highly variable signals, whereas others may contain mostly white matter that shows very little variability. Secondly, neighboring voxels share noise processes to some degree, not only because of interpolation though motion correction, but also because some of the underlying noise processes are spatially smooth. Therefore a Euclidean distance measure (which weighs each voxel equally independent of its variance) is suboptimal. Indeed, we have shown that the split-half reliability of distance measures can be increased by taking the noise structure into account (Walther et al., in preparation).

There are two identical ways of doing this: First we can calculate the Mahalanobis distance (Mahalanobis, 1936) between the mean patterns for each condition. A Mahalanobis distances is essentially a Euclidean distances that is weighted by the spatial structure of the noise, meaning noisy voxels, or highly correlated groups of voxels, are down-weighted:

$$\tilde{U} = Z^{+}Y$$
$$d_{i,j}^2 = \left(\tilde{\mathbf{u}}_i - \tilde{\mathbf{u}}_j\right)^T \hat{\Sigma}_\varepsilon^{-1}\left(\tilde{\mathbf{u}}_i - \tilde{\mathbf{u}}_j\right) \quad \text{Eq. 3}$$

Where $\hat{\Sigma}_\varepsilon$ is an estimate of the *PxP* noise-covariance matrix, estimated from the residuals of the regression:

$$\mathbf{r} = \mathbf{Y} - \mathbf{Z}\tilde{\mathbf{U}}$$
$$\hat{\Sigma}_\varepsilon = \mathbf{r}^T\mathbf{r} / \left(N - K\right) \quad \text{Eq. 4}$$

Estimation of the noise covariance-matrix is usually not performed on the whole brain, but – because we want to make inferences on the different representations in different areas of the brain - on local regions of interest or search-lights with restricted number of voxels. Despite this, it is not unusual that P>(N-K), in which case we need to regularize our estimate (Ledoit and Wolf, 2003). Equivalently, we can first prewhiten the beta estimates and then calculate the Euclidean distance:

$$\hat{\mathbf{U}} = \tilde{\mathbf{U}}\hat{\Sigma}_\varepsilon^{-1/2}$$
$$d_{i,j}^2 = \left(\hat{\mathbf{u}}_i - \hat{\mathbf{u}}_j\right)^T \left(\hat{\mathbf{u}}_i - \hat{\mathbf{u}}_j\right) \quad \text{Eq. 5}$$

, which yields the exact same distance as in Eq. 3.

The use of prewhitening demands that The bad news for multivariate fMRI analysis is that we have to conduct the multivariate analysis starting from the raw time-series (see Appendix 2 for details) – we cannot simply do a univariate first-level GLM first, and then subsequently conduct the multivariate analysis on the regression estimate from the univariate analysis. The good news is that for each ROI or search light, we simply need to store the sufficient statistics, the estimate of **G**, and we then can derive all other measures quickly from these estimates. The additional computational effort of going to the original time series is usually fully justified by the increased reliability and power of the resultant measures.

Cross-validation

The suggested distance measures, however, have one disadvantage: When an area does not distinguish between two stimuli, i.e. the two conditions have identical true patterns, then the true distance is zero. However, when we estimate **U** from noisy data, our estimates deviate from the true values by an estimation error, $\hat{\mathbf{U}} = \mathbf{U} + \eta$. Thus, the estimated patterns for condition 1 and 2 will be slightly different, and their estimated distances (Eq. 5) will be larger than zero. Indeed, the expected value of the estimated squared distances is $E\left(\hat{d}_{i,j}^2\right) = d_{i,j}^2 + 2\sigma_\eta^2$. This means we cannot

simply compare distance estimates against zero to test whether two patterns are significantly different. Furthermore, because all distance estimates will increase equally with increasing noise, the whole representational structure will be distorted (see section 5).

This dependence on noise can be fixed by using cross-validated estimates of the distances. In short, we divide out data set into *M* independent cross-validation folds. In the case of fMRI, it is common to let each fold be a separate imaging run, as the activation estimates across runs can be assumed to be independent. We then estimate **U** using prewhitening (Eq. 3, 5) on each fold separately, which results in *M* estimates $\hat{\mathbf{U}}^{(1)}...\hat{\mathbf{U}}^{(M)}$. These can then be used to compute the distance between condition *i* and *j* for each possible pairs of folds separately, and finally average across all *M(M-1)* pairs:

$$\hat{d}_{i,j}^2 = \sum_{l,m;l\neq m}^{M} \left(\hat{\mathbf{u}}_i^{(m)} - \hat{\mathbf{u}}_j^{(m)}\right)^T \left(\hat{\mathbf{u}}_i^{(l)} - \hat{\mathbf{u}}_j^{(l)}\right) / \left(M\left(M-1\right)\right) \qquad \text{Eq. 6}$$

Because the estimation noise is independent across folds, the expected value of $\eta^{(m)T}\eta^{(l)}$ is zero. Hence – the expected value of the crossvalidated distance estimate equals the true distance between the patterns.

This also means that, especially for small true distances, our distance estimate will sometimes become negative. This is no reason for concern, but rather an inevitable characteristic of an unbiased estimator. Because we sometimes overestimate the true distances, we also sometimes need to underestimate it. For a true distance of zero (the two patterns are equal), half the estimates should be negative. This means that we can use cross-validated distances – like cross-validated classification accuracies - to make inferences about differences between conditions (see section 4). By extension, we can also derive a cross-validated estimate of **G**:

$$\hat{\mathbf{G}} = \sum_{l,m;l\neq m}^{M} \left(U^{(m)}U^{(l)T}\right) / \left(M\left(M-1\right)\right) \qquad \text{Eq. 7}$$

While the expected value of $\hat{\mathbf{G}}$ is **G**, the estimate is not guaranteed to be a positive-definite matrix – i.e. the diagonal may contain negative elements, or the off-diagonals values may violate the condition $G_{i,j} > \sqrt{G_{ii}G_{jj}}$. This can occasionally cause some practical problems, for example estimated correlation coefficients can fall outside of [-1; +1]. In general, however, it allows us to use $\hat{\mathbf{G}}$ for inferences.

To summarize, these considerations suggest the general practical guide to multivariate analysis. (i) Define groups of voxels on which to conduct the analysis, for example using regions of interests, or volume-based (Kriegeskorte et al., 2006) or surface-based (Oosterhof et al., 2011) search lights. (ii) Estimate the regression-coefficients for each imaging run separately and prewhiten these using the estimate of the noise-covariance, obtained from the residuals of this first-level regression (Eq. 3-5). (iii) Calculate a cross-validated estimate **G** (Eq. 7). (iv) Obtain a suitable summary statistics calculated on $\hat{\mathbf{G}}$ for each person / region. (v) Make inference on the group level – using either traditional closed-form or permutation statistics (Stelzer et al., 2013).

4. Detecting encoding

The most basic use of multivariate analysis of fMRI data is to infer that a region encodes a certain variable of interest – i.e. that it shows significantly different patterns of activity between two or more conditions. This is traditionally done by using either LDA or SVM classifiers to test whether the patterns in a region allow for above-chance accuracy.

The use of cross-validated Mahalanobis distances allows for an equally simple, but more powerful test of encoding. Because the expected value of the distances is zero when two conditions are identical, we can simply test this difference against zero. Indeed, the crossvalidated Mahalanobis distances (Eq. 6) is very tightly related to the discriminant function of the LDA-classifier. This means classification accuracy is basically a discretization of a more continuous distance measure. In situations in which we are not interested in decoding per se, but would like to make inferences about the underlying distribution, a continuous measure is more reliable and hence provides a more powerful test of encoding (Walther et al., in preparation).

For more than 2 classes, we can use the average squared distance between any possible pair of conditions as a test-statistics.

$$H = \sum_{i \neq j}^{K} \hat{d}_{i,j}^{2} \; / \; K\left(K-1\right) \qquad \text{Eq. 8}$$

Given the relationship between the distances and **G,** some basic algebra shows that this average distance is proportional to difference between the mean of the diagonal of $\hat{\mathbf{G}}$ (variances) and the mean of the off-diagonal (covariances):

$$\frac{1}{2}H = \sum_{i}^{K} \hat{\mathbf{G}}_{i,i} \; / \; K - \sum_{i \neq j}^{K} \hat{\mathbf{G}}_{i,j} \; / \; K\left(K-1\right) \qquad \text{Eq. 9}$$

Thus, as a test for encoding, we can equivalently ask whether the average pair-wise distance is larger than zero, or if the covariance (or more precisely, the inner product) of patterns of the same condition is higher than those of pattern from different conditions. This emphasizes the equivalence on making inferences on the estimated second moment or the estimated distances, which will carry forward through basically all following examples.

5. Representational similarity analysis (RSA)

Rather than asking whether a region shows *any* differences between *any* pair of stimuli or conditions, RSA looks at the full structure of distances between conditions. This representational structure can be visualized by plotting the activity patterns of the conditions into a two or three-dimensional space such that the distances between stimuli are well preserved. For example, when applying this data visualization technique to the distance matrices presented in Figure 1, a stable and orderly arrangements of the digits in primary motor cortex becomes visible. Without cross-validation this structure would actually be highly dependent on the overall noise level (Diedrichsen et al., 2011). With crossvalidation, a distance of zero becomes meaningful and the ratios between distances becomes interpretable – thus statements of two fingers being 1.5 times as far away as another pair of digits becomes meaningful.

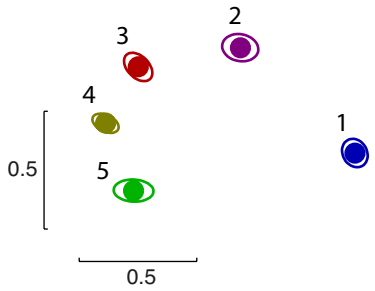

Figure 3. *Classical multidimensional scaling on the pattern distances between fingers in primary motor cortex. Digit 1 denotes the thumb, digit 5 the little finger. Ellipses denote between-subject standard error (Ejaz et al., in preparation).*

A simple technique to plot the similarity structure is classical multi-dimensionality of scaling. This technique uses the eigenvectors ($\mathbf{v}_i$) and eigenvalues (λ_i) of a centered version of $\hat{\mathbf{G}}$. To subtract out the mean pattern, we pre- and post-multiply $\hat{\mathbf{G}}$ with the *KxK* centering matrix **C**, which calculates each conditions against the mean of the remaining conditions ($\mathbf{C} = \mathbf{I}_K - \mathbf{1}/K$). The first eigenvector indicates are the value of

each condition on the hidden dimension that best separates the different classes, and the first eigenvalue tells us how much between-condition variance this vector explains. For a two-dimensional MDS, we would then simply plot $v_1\sqrt{\lambda_1}$ on the x-axis and $v_2\sqrt{\lambda_2}$ on the y-axis. Note that classical MDS, although it our case appropriate, not always provides the best visualization. Other techniques, such as non-classical distance scaling are also useful.

The real attraction of RSA, however, is to compare the obtained representation similarity structure to the prediction of various representational models. For example, we could compare the representational similarity of activity patterns evoked by visual stimuli in inferior-temporal cortex to the predictions of computer vision models. This idea has been extensively discussed and the interested reader in referred to these papers {}. One open question in this model comparison, however, is which metric should used to judge the correspondence between predicted and measured distances. The most cautious approach is to only assume that the rank-ordering of the distances is preserved (Kriegeskorte et al., 2008), suggestion the use of Kandall's-tau statistics (Nili et al., 2014).

However, the use of cross-validated distances now allows us to interpret the ratio of the distances, and not only their rank-ordering. Furthermore, if a model predicts that a region should not differentiate between two conditions, then this is inherently meaningful, and should factor into the evaluation of the model. On the other hand, the overall scaling of the distances themselves is dependent on the signal-to-noise ratio in the data set, precluding the use of the squared prediction error $\left(\hat{d}^2_{i,j} - p^2_{i,j}\right)^2$, where $p_{i,j}$ is the predicted distance between i and j. This suggests the use of a correlation coeifficient between the measured distances and the predictions, with the intercept held fixed.

$$r = \frac{\left\langle \hat{d}^2, p^2 \right\rangle}{\left\| \hat{d}^2 \right\| \left\| p^2 \right\|} \qquad \text{Eq. 10}$$

More advanced methods of model evaluation, such as mixture models, etc, will be covered in future part of the tutorial paper.

6. Pattern component modeling (model-constrained RSA)

Often the design of an fMRI experiment is determined by different factors of interest that that have a hierarchical relation to each other. For example, visual stimuli belong categories. Instead of looking at the structure of all pairwise distances between

stimuli, we may then think of our data as a composition of pattern components associated with each category, and a component associated each individual stimulus.

Consider for example an fMRI experiment concerned with the perception of body-parts and inanimate objects. In the experiment, two visual objects from the body part category (e.g. a face and a hand) are presented, while the other two stimuli are inanimate objects (e.g. a visual scene and a fruit). We may then presume the representational geometry of the fMRI patterns reflects this categorical division (Fig. 4a). All stimuli will share on common activity pattern, which simply corresponds to the default response of the system to any visual stimulus. Additionally, a region may also have a common response to all stimuli from a certain category. For example, the extrastriate body area would have a strong category-specific response to any body-part, while the response to inanimate objects would be lower. Finally, a region can also be characterized by how well it discriminates between different stimuli within a certain category. For example, we would expect the extrastriate body area to have a larger stimulus specific response for body parts than for visual scenes.

We can exploit our knowledge about the experimental design to estimate the strength of these pattern components from **G**. The only constraining assumption we need to make is that at each level, the different patterns are orthogonal to each other – that is the response to visual scenes is independent from the reponse to body parts. This seems like a strong assumption – however, in a high-dimensional space (and we are using usually 20+ voxels) two random, unrelated vectors are guaranteed to be nearly orthogonal. Under this assumption, we can think about the second moment matrix G as being composed of multiple components or basis matrices $\mathbf{G}_c$ (Fig. 4b).

$$\mathbf{G} = \sum_c \mathbf{G}_c h_c \qquad \text{Eq. 11}$$

The strength of each individual component can then be estimated using linear regression: For this we need to stretch the matrices $\mathbf{G}_c$ and **G** into vectors, using the vector operator vec()**:**

$$\mathbf{X} = \left[\; vec\left(\mathbf{G}_1\right) \;\; vec\left(\mathbf{G}_2\right) \;\; ... \; \right]$$

$$\hat{\mathbf{h}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T vec\left(\hat{\mathbf{G}}\right)$$

The vector $\hat{\mathbf{h}}$ then gives us unbiased estimates of the strength of the common, the two category-specific, and the stimuli-specific activity patterns (Fig 3a).

It is instructive to compare the suggested estimation to a more direct way of estimating the variances of the experimental factors: We could simply use the empirical mean response to all stimuli and then measure the norm length of this vector (the sum-of-squares) to determine its strength. Then we would subtract the overall mean pattern from all stimuli and then calculate the mean within each category and compute its norm. Finally we would subject out the category mean pattern from each stimulus pattern to estimate its strength. In MVPA, this procedure has been dubbed "cocktail-blank removal" and, though widely used in fMRI (Op de Beeck, Brants, Baeck, & Wagemans, 2010; Op de Beeck, 2010; Williams et al., 2008; Williams, Dang, & Kanwisher, 2007), has lately been subject of criticism (Garrido, Vaziri-Pashkam, Nakayama, & Wilmer, 2013; Diedrichsen et al., 2011).

To compare both methods, we simulated an fMRI pattern ensemble as laid out in Fig. 4a, where all pattern components have a variance of one. We then estimated the pattern variance component of each factor either by calculating the mean patterns (Fig. 4c and e) or by using pattern component modeling on **G** (Fig. 4d and f). As can be seen, estimation over the mean patterns leads to significant distortions of the representational structure. At low noise levels, the stimulus- and category-specific effects are underestimated, whereas the mean activation is overestimated. This is because the category centroid is now assumed to lie exactly between the two stimuli, making the stimulus specific effects point in opposite directions (Fig 3c, induces an anti-correlation). The vector for the mean activity needs to be correspondingly larger. Increasing noise levels then mostly impacts the estimation of length of the stimulus-dependent component. This is because the estimates of the stimulus-specific patterns rely on fewer data sample than the estimates of the mean activity patterns (Fig. 4c).

Contrary to that, estimating the variance components through **G** does show two difference: In the noise-less case the mean activity pattern is estimated to be smaller, as category- and stimulus-specific effects are assumed to be uncorrelated, rather than anti-correlated (Fig. 4f). Furthermore, the cross-validation of the estimate prevents variance inflation from noise, but faithfully reflects the true representation.

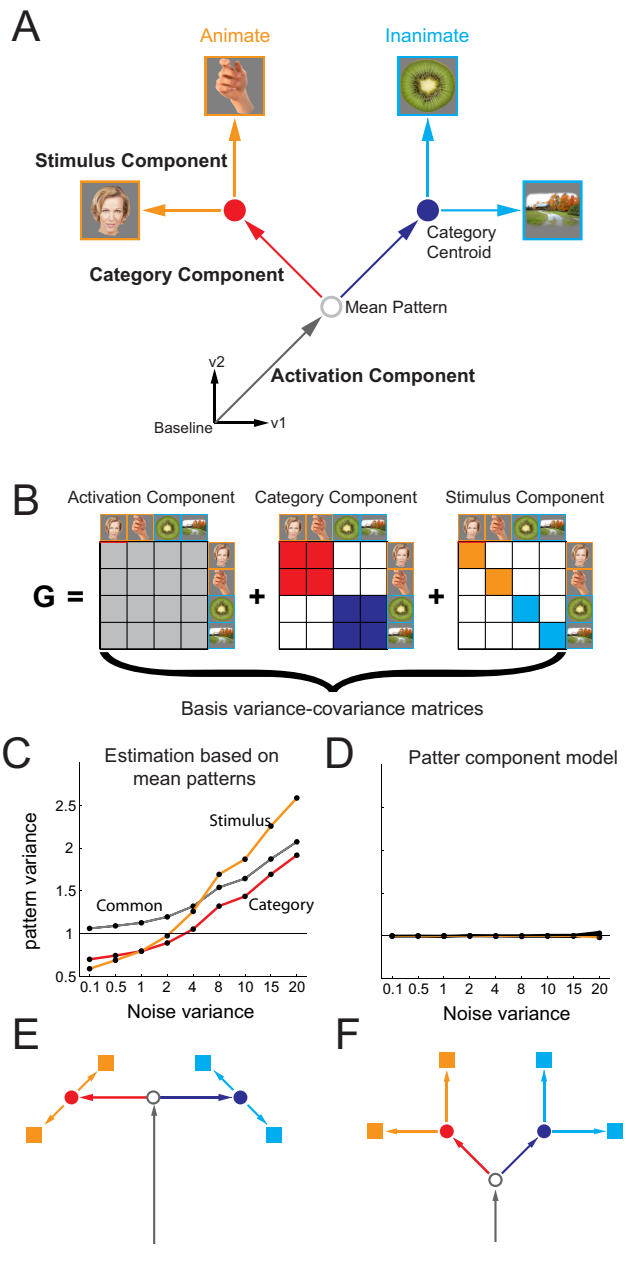

Figure 4. *Pattern component modeling of representational geometry.* ***A****. The representational geometry of an fMRI experiment with two categories (body-parts and inanimates), each of which contains two stimuli. For illustration purposes, the arrangement is depicted in 2D, although the patterns truly live in a P-D space. All stimuli share a common component of activation, and their respective category pattern. Each individual stimulus then is characterized by its own stimulus-specific pattern component.* ***B****. The second-moment matrix* **G** *can be decomposed into basis matrices associated with the common component, the category- and the stimulus-specific components (C, D) Estimation of the strength of the components for increasing noise levels, using the mean patterns or pattern component modeling. (E, D) VIsualisation of differences in the estimation procedures. Estimation over mean patterns assumes that the centroids lie exactly in-between the measured patterns and therefore assumes anti-correlation of stimulus patterns. Pattern-component modeling assumes independence of the category- or stimuli-specific patterns.*

7. Fully crossed designs (MANOVA) and pattern consistency

In the last section we have seen how the representation structure of a hierarchical experimental design can be understood as being composed of different pattern components, which can be estimated using simple linear contrast of the estimated **G** matrix. We will now show how this idea carries over to designs in which two or more experimental factors are fully crossed.

Consider an fMRI experiment in which the participants were asked to either observe and perform three different types of hand movements (e.g., Oosterhof et al., 2010). In this design, we may want to ask which regions encode different grasps when executing them without vision, and when observing them without execution. The main interest, however, is the question whether there is a region, in which executing a grasp yields similar activation patterns as observing the same grasp, which would constitute clear evidence for a mirror-neuron system. Such questions are in addressed in the context of classification by using first within-modality classifiers to determine modality-specific encoding and then a cross-modality classifier that is trained on data from the observation condition and then applied to the execution condition, or vice versa (Oosterhof et al., 2010). Designs of similar structure are commonly observed in multivariate analysis (Gallivan et al., 2013; Wiestler et al., 2014).

The balanced 2x3 design suggests using classical statistical methods, such as MANOVA, instead of classification. In the following we will show how the hypotheses that are commonly tested in multivariate experiments can be intuitively translated into a MANOVA framework by relying on linear contrasts on the pivotal statistical quantity $\hat{\mathbf{G}}$. The resultant test statistics are exactly identical to a the recently proposed cross-validated variant of the Barlett-Lawley-Hottellings trace statistics (Allefeld & Haynes, 2014), which has been indeed been shown to lead to more powerful inferences that classification analysis.

In MANOVA we can conduct classical statistical tests to test the significance of the main effects, or the interactions effect. Each of these tests can be defined by a contrast matrix **C** (figure 5). As for the F-test, a multivariate test has often numerous degrees of freedom, meaning that it test for multiple linear contrast (or any combination of these) at the same time {}. For example, testing the main effect of grasp type means that we would like to test for any differences between the three grasp types, averaged across “see” and “do” condition. The first row of **C** could therefore contrast grasp 1 and 2, and the second row grasp 2 and 3. Note that there are many different contrast matrices that encode exactly the same statistical test.

Allefeld et al. (2014) proposed the measure “pattern distinctness” D, a cross-validated version of the Barlett-Lawley-Hottellings trace:

$$D = \frac{1}{c}\,\mathrm{trace}\left(\tilde{\mathbf{U}}^{(A)\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\mathbf{H}\tilde{\mathbf{U}}^{(B)}\hat{\Sigma}_{\varepsilon}^{-1}\right)$$

Where $\tilde{\mathbf{U}}^{(A)}$ are the un-prewhitened patterns estimates from crossvalidation fold A, c is a normalization constant, $\Sigma_{\varepsilon}^{-1}$ is the estimated noise variance of from the residuals of the first-level regression, and **H** is a squared form the contrast matrix, the so-called hypothesis matrix: $\mathbf{H} = \mathbf{C}(\mathbf{C}^{\mathsf{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathsf{T}}$.

As can be shown (see 7.1) this is equivalent to a linear contrast on $\hat{\mathbf{G}}$, in which we multiply every element of $\hat{\mathbf{G}}$ with the corresponding element of **H**, and sum them up. Thus, **H** has an intuitive interpretation as a weight matrix on a certain contrast for $\hat{\mathbf{G}}$.

For example the simple main effect of any differences between grasp (i.e. grasp encoding) in the observation conditions would be to test whether the first three diagonal elements are bigger than the mean of the corresponding of diagonals, as described in section 3. The main effect of grasp, averaged across see and do, would test the inner product between all corresponding pairs of grasp (no matter which modality) against all other pairs (see Figure 5).

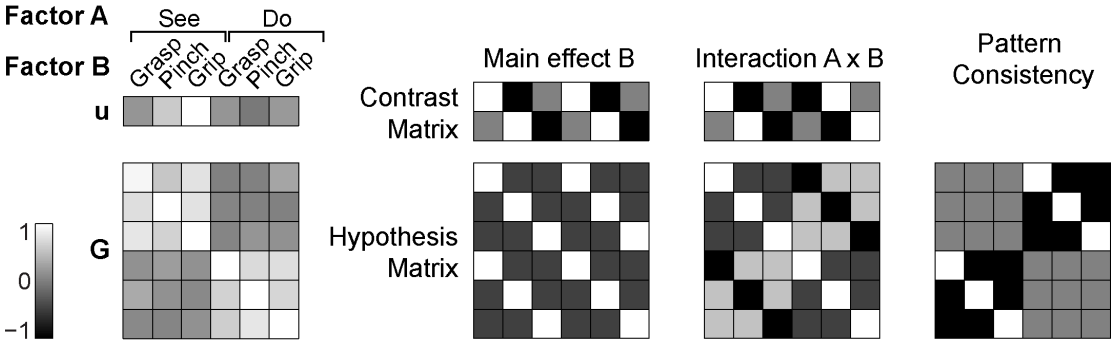

Figure 5. *Analysis of a 2x3 fully crossed experimental design. The second-moment matrix **G** consists of 4 quadrants: the inner products of the 3 grasp patterns in the see-condition (upper left), in the do-condition (lower right), and across condition (upper right, lower left). While the main effect of grasp type and the interaction effect can be specified in terms of traditional contrast matrices on the regression coefficients, the test for pattern consistency across conditions can only be expressed as a contrast of **G** as specified by the corresponding Hypothesis matrix.*

The interaction effect between the two factors is often of neuro-scientific interest, as an absence of an interaction indicates that the patterns associated with the two factors linearly superimpose in a region and thus are likely encoded in separate populations of neurons, whereas a strong non-linear interaction indicates integration of the two factors (Diedrichsen et al., 2013b; Fabbri et al., 2014; Kornysheva and Diedrichsen, 2014). In terms of a Hypothesis matrix, such an interaction effect, compares the size of within-modality encoding (see-see, do-do) to the size of the across-modality encoding (do-see and see-do) of a grasp.

But what of the test of main interest – i.e. the test whether a region encodes the grasp in similar fashion for observation and execution? Based on what we have learned so far, we can simply intuit the correct hypothesis matrix, namely testing the diagonal of the see-do and do-see block of the **G**-matrix to the off-diagonal elements (see Fig. 5). Thus, we need to test how consistent the patterns for the different grasp types are across "see" and "do" conditions. Interestingly, however, this particular contrast has no simple corresponding classical MANOVA contrast that could be specified in terms of a contrast matrix (Allefeld and Haynes, 2014). Rather, the hypothesis matrix for *pattern consistency* needs to be specified by subtracting the interaction from the main effect. However, if we view statistical tests as a linear combination of elements of $\hat{\mathbf{G}}$, the specification of the corresponding hypothesis matrix becomes quite intuitive.

9. Conclusion and outlook

Multivariate analysis of fMRI data is currently undergoing a rapid development into a mature sub-discipline of neuroscience. Traditionally, this analysis approach as practically relied heavily on cross-validated classification approaches. However, usually we are not interested in decoding external variables from brain states, but to learn how external variables are encoded in neural activity. For this purpose, classification is not an optimal method, as stressed by many different groups (Allefeld and Haynes, 2014; Diedrichsen et al., 2011; Kriegeskorte et al., 2008; Naselaris et al., 2011). On the other hand, traditional multivariate techniques such as MANOVA and CCA have not found widespread use, as the number of variables (voxels) is too large for the available data, which renders classical test statistics invalid.

Based on practical experience, we suggest here a middle approach that is powerful, computational efficient, and conceptually easy to understand. We have presented a range of techniques that are all based on the matrix of cross-validated inner products of the activity patterns, which serves as a central statistical quantity

for the representational analysis. Ongoing research in our laboratories also indicates that we can test and quantify the dimensionality of the representation (Diedrichsen et al., 2013a), and the question of the spatial arrangement of the representations within this framework. Furthermore, while the methods used here are inherently linear, the extension to non-linear methods are relatively straightforward by replacing the inner product in G with a non-linear kernel.

6. References

Allefeld, C., and Haynes, J.D. (2014). Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. Neuroimage 89, 345-357.

Ben-Hur, A., Ong, C.S., Sonnenburg, S., Scholkopf, B., and Ratsch, G. (2008). Support vector machines and kernels for computational biology. PLoS Comput Biol 4, e1000173.

Diedrichsen, J., Ridgway, G.R., Friston, K.J., and Wiestler, T. (2011). Comparing the similarity and spatial structure of neural representations: a pattern-component model. Neuroimage 55, 1665-1678.

Diedrichsen, J., Wiestler, T., and Ejaz, N. (2013a). A multivariate method to determine the dimensionality of neural representation from population activity. Neuroimage 76, 225-235.

Diedrichsen, J., Wiestler, T., and Krakauer, J.W. (2013b). Two distinct ipsilateral cortical representations for individuated finger movements. Cereb Cortex 23, 1362-1377.

Duda, R.O., Hart, P.E., and Stork, D.G. (2001). Pattern classification (Hoboken, NJ: Wiley).

Ejaz, N., J., D., and Hamada, M. (in preparation). Natural statistics of hand use shapes activity patterns in the primary motor and sensory cortex.

Fabbri, S., Strnad, L., Caramazza, A., and Lingnau, A. (2014). Overlapping representations for grip type and reach direction. Neuroimage 94, 138-146.

Friston, K., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., and Ashburner, J. (2008). Bayesian decoding of brain images. Neuroimage 39, 181-205.

Gallivan, J.P., McLean, D.A., Flanagan, J.R., and Culham, J.C. (2013). Where one hand meets the other: limb-specific and action-dependent movement plans decoded from preparatory signals in single human frontoparietal brain areas. J Neurosci 33, 1991-2008.

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

Indovina, I., and Sanes, J.N. (2001). On somatotopic representation centers for finger movements in human primary motor cortex and supplementary motor area. Neuroimage 13, 1027-1034.

Kornysheva, K., and Diedrichsen, J. (2014). Human premotor areas parse sequences into their spatial and temporal features. Elife 3, e03043.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proc Natl Acad Sci U S A 103, 3863-3868.

Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci 17, 401-412.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci 2, 4.

Ledoit, O., and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance 10(5).
Mahalanobis (1936). On the generalized distance in statistics. Proceedings of the National Institute of Sciences 2, 49-55.
Misaki, M., Kim, Y., Bandettini, P.A., and Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage 53, 103-118.
Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fMRI. Neuroimage 56, 400-410.
Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. PLoS Comput Biol 10, e1003553.
Oosterhof, N.N., Wiestler, T., Downing, P.E., and Diedrichsen, J. (2011). A comparison of volume-based and surface-based multi-voxel pattern analysis. Neuroimage 56, 593-600.
Oosterhof, N.N., Wiggett, A.J., Diedrichsen, J., Tipper, S.P., and Downing, P.E. (2010). Surface-based information mapping reveals crossmodal vision-action representations in human parietal and occipitotemporal cortex. J Neurophysiol 104, 1077-1089.
Stelzer, J., Chen, Y., and Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. Neuroimage 65, 69-82.
Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and J., D. (in preparation). Reliability of dissimilarity measures for multivariate fMRI pattern analysis.
Wiestler, T., McGonigle, D.J., and Diedrichsen, J. (2011). Integration of sensory and motor representations of single fingers in the human cerebellum. J Neurophysiol 105, 3042-3053.
Wiestler, T., Waters-Metenier, S., and Diedrichsen, J. (2014). Effector-independent motor sequence representations exist in extrinsic and intrinsic reference frames. J Neurosci 34, 5054-5064.

7. Appendix / Footnotes

7.1 Cross-validated Bartlett-Lawley-hottelngs' trace

The test-statistics

$$D = \sum_{m \neq l} \text{trace}\left(\tilde{\mathbf{U}}^{(m)\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\mathbf{H}\tilde{\mathbf{U}}^{(l)}\hat{\Sigma}_{\varepsilon}^{-1}\right)$$

can be shown to be equivalent to a linear contrast on the $\hat{\mathbf{G}}$. By using the "trace-trick" trace(ABC) = trace(BCA) it can can easily rotate the matrix such that

$$D = \sum_{m \neq l} \text{trace}\left(\mathbf{H}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\mathbf{H}\tilde{\mathbf{U}}^{(l)}\hat{\Sigma}_{\varepsilon}^{-1}\tilde{\mathbf{U}}^{(m)\mathsf{T}}\right)$$

where we can see that the last term is one cross-validation fold paring of the estimate of **G** on prewhitened data (Eq. X). Thus, we arrive at the equivalence

$$
\begin{aligned}
D &= \text{trace}(\mathbf{H}^T\mathbf{Z}^T\mathbf{Z}\mathbf{H}\hat{\mathbf{G}}) \\
&= \text{trace}(\mathrm{H}\hat{\mathbf{G}}) \\
&= \sum_{i,j} \mathrm{H}_{i,j}\hat{\mathbf{G}}_{i,j}
\end{aligned}
$$

For balanced fMRI designs with $\mathbf{Z}^T\mathbf{Z} \approx \mathbf{I}c$ the new hypothesis matrix H is simply a scaled version of the old hypothesis matrix **H**. Since statistical test on the crossvalidated quantity need to be conducted through permutation test (Stelzer et al., 2013), or on the group level by using the inter-subject variability as a SE, we can simply ignore such arbitrary scaling.

7.2 Statistical properties of inner products

To obtain optimal summary statistics on $\hat{\mathbf{G}}$ it is useful to consider the statistical properties of inner products. First, we determine the probability distribution of one of the elements of G, calculated on one pair of folds. Assume you have two random vector a,b which are both noisy instantiations of a true vector with a=A+e and b=B+f, where e and f are independent random vectors with zero mean and variance σ_e^2 and σ_f^2. We first can decompose the inner product:

$$\langle \mathbf{a},\mathbf{b} \rangle = \langle \mathbf{A}+\mathbf{e},\mathbf{B}+\mathbf{f} \rangle = \langle \mathbf{A},\mathbf{B} \rangle + \langle \mathbf{A},\mathbf{f} \rangle + \langle \mathbf{B},\mathbf{e} \rangle + \langle \mathbf{e},\mathbf{f} \rangle$$

Because the expected inner product between the two noise vectors, and between the noise vectors and the true patterns is zero, we have the simple result:

$$\mathrm{E}\left(\langle \mathbf{a},\mathbf{b} \rangle\right) = \langle \mathbf{A},\mathbf{B} \rangle$$

The variance of the product of two random variances with zero mean is the product of their variance. From this we can conclude that:

$$\text{var}\left(\langle \mathbf{a},\mathbf{b} \rangle\right) = \langle \mathbf{A},\mathbf{A} \rangle \sigma_f^2 + \langle \mathbf{B},\mathbf{B} \rangle \sigma_e^2 + \sigma_e^2\sigma_f^2 \mathrm{P}$$