Correlations of correlations are not reliable statistics: implications for multivariate pattern analysis.

Bertrand Thirion, Fabian Pedregosa, Michael Eickenberg, Gaël Varoquaux

Correlations of correlations are not reliable statistics: implications for multivariate pattern analysis.

Bertrand Thirion BERTRAND.THIRION@INRIA.FR
Parietal team, INRIA, Saclay and CEA, Neurospin France

Fabian Pedregosa
CEREMADE/Chaire Havas-Dauphine Économie des Nouvelles Donées

Michael Eickenberg
Parietal team, INRIA, Saclay and CEA, Neurospin France

Gaël Varoquaux
Parietal team, INRIA, Saclay and CEA, Neurospin France

Abstract

Representational Similarity Analysis is a popular framework to flexibly represent the statistical dependencies between multi-voxel patterns on the one hand, and sensory or cognitive stimuli on the other hand. It has been used in an inferential framework, whereby significance is given by a permutation test on the samples. In this paper, we outline an issue with this statistical procedure: namely that the so-called pattern similarity used can be influenced by various effects, such as noise variance, which can lead to inflated type I error rates. What we propose is to rely instead on proper linear models.

1. Introduction

The use of machine learning in functional neuroimaging has been boosted in the recent years by the adoption of the so-called *multivariate pattern analysis* (MVPA) framework, in which brain activation signals are compared to stimuli using multivariate models such as (Haxby et al., 2001; Cox & Savoy, 2003; Haynes & Rees, 2006). More precisely, two settings have emerged to draw statistically meaningful conclusions regarding the statistical associations between experimental stimuli and brain activation measurements: On the one hand, *encoding models* define a mapping from possibly very high-dimensional features extracted from the stimuli to brain activity at a given location.

ICML Workshop on Statistics, Machine Learning and Neuroscience

On the other hand, *decoding models* test whether the activity in a given set of voxels –possibly the whole brain– are predictive of a certain feature of the stimuli used during the acquisition. These two settings have been clearly described in e.g. (Naselaris et al., 2011; Varoquaux & Thirion, 2014). The popularity of this framework is notably driven by the premise to offer a more sensitive detection of task-related brain activations, due to the pooling effect on voxels (decoding) or on stimulus features (encoding). In the present work we focus on functional Magnetic Resonance Imaging (fMRI) data.

An alternative to these two models has been proposed, which borrows from both ideas: it consists in quantifying the between-sample similarity of the stimuli on the one hand and of the evoked activation signals on the other hand, in order to find whether there is some common structure between these two sets of similarities. This approach has been called *Representational Similarity Analysis* (RSA) (Kriegeskorte et al., 2008; Kriegeskorte, 2009) and is also very popular. A striking aspect is that, unlike encoding and decoding models that require algorithmically or computationally involved estimators, RSA simply relies on descriptive statistics of the data, making it conceptually simple and affordable.

A discussion of the motivation for RSA can be found in (Nili et al., 2014): On the one hand, this approach offers a great flexibility in terms of experimental design, and is easy to implement and use. It is also viewed as a sensitive statistic to detect associations between stimuli and brain activity (see e.g. (Borghesani et al., 2014)). On the other hand, this approach does not rely on any signal model; unlike more classical encoding and decoding, it actually avoids defining

explicit associations between brain patterns and combinations of stimuli. In that sense, it can be viewed as a *black box* model.

It is fair to consider that there are two main parts to RSA: one is the representation of similarities of the input stimuli and the second part compares it to neuroimaging data correlation. In this work, we discuss the second part, namely the comparison between RSA and linear encoding models. We outline the convergence between the two approaches, but also some important differences that should be taken into account when discussing the results of statistical analysis based on RSA. Our contribution consists of a discussion of the model, followed by illustrative experiments on simulated and experimental data. Our main finding is that, in spite of the use of non-parametric statistics, RSA-based inference is not reliable, because it can be sensitive to effects that are not stimulus-related signals increase (or decrease). To give a simple perspective, we only consider the simplest setting where RSA can be compared to alternative encoding schemes.

2. Statistical inference in RSA

2.1. Representational Similarity Analysis

Let $\mathbf{Y}$ be an fMRI dataset, written as an $n \times p$ matrix, where n is the number of samples and p is the number of voxels, possibly after reduction to a particular Region of Interest. Note that n can be the number of acquired images or the result of a deconvolution step. This dataset is associated with an experimental paradigm, represented by a succession of n stimuli presentations. One can represent it with a design matrix $\mathbf{X}$ of shape (n, q), where q is the number of stimulus features or directly with a kernel matrix $\mathbf{K}$ of size $n \times n$ that represents some kind of similarity between the stimuli. In this work,we explicitly assume that both representations are available, with $q < n$ and that $\mathbf{K} = \text{Corr}(\mathbf{X})$.

Representational similarity analysis proceeds by extracting the lower-triangular coefficients of the activation similarity matrix $\text{Corr}(\mathbf{Y})$, yielding $\mathbf{t}_Y = \text{Tril}(\text{Corr}(\mathbf{Y}))$ and of the kernel $\mathbf{t}_X = \text{Tril}(\mathbf{K})$. The decision statistic is Spearman correlation between $\mathbf{t}_X$ and $\mathbf{t}_Y$. In the following derivation, we consider Pearson correlations instead to simplify the analysis, but this is actually an arbitrary choice, and we did not observe any significant difference in a real dataset (note that in the experiments described later, we use Spearman correlation).

$$R_{RSA} = \text{Pearson}/\text{Spearman}(\mathbf{t}_X, \mathbf{t}_Y) \tag{1}$$

2.2. Statistical inference

This basic RSA setting can be used to detect significant associations between $\mathbf{Y}$ and $\mathbf{X}$ by using a permutation test: after shuffling either $\mathbf{Y}$ or $\mathbf{X}$ in the sample dimension J times, with e.g. $J = 10^4$, the distribution of the permuted Spearman correlation $(R^j_{RSA})_{j\in[J]}$ is computed and the initial value R_{RSA} in eq. 1 is compared with $(R^j_{RSA})_{j\in[J]}$, where the proportion of higher values in the permuted sample is the p-value.

2.3. Comparison with encoding model

The encoding formulation is obtained through the following mass-univariate model:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{2}$$

where $\mathbf{B}$ is $q \times p$ matrix that represents the response in each voxel. For instance, if $\mathbf{X}$ is the occurrence matrix of a stimulus belonging to a set of discrete classes, $\mathbf{B}$ are the univariate effects of an ANOVA model. Here we assume that $q \ll n$, so that one can resort to simple least-squares estimators. The natural decision statistic of the linear model (2) is the residual sum of squares of the residuals, which is a monotonous function of the following R^2 quantity:

$$R^2 = Tr(\mathbf{X}\hat{\mathbf{B}}\hat{\mathbf{B}}^T\mathbf{X}^T) \tag{3}$$

$$= Tr(\mathbf{XBB}^T\mathbf{X}^T) + Tr(\mathbf{XX}^\dagger\mathbf{EE}^T) + 2Tr(\mathbf{XX}^\dagger\mathbf{EB}^T\mathbf{X}^T),$$

where $\mathbf{X}^\dagger$ is the pseudo-inverse of $\mathbf{X}$; the third term can be neglected as it has a null expected value. The remaining error term is actually the squared norm of the projection of $\mathbf{E}$ on the span of the design matrix column vectors.

2.4. Statistical issues with RSA

Using the previous generative model, a rudimentary data kernel, without centering or normalization, is given by:

$$\mathbf{YY}^T = \mathbf{XBB}^T\mathbf{X}^T + \mathbf{EE}^T + \mathbf{XBE}^T + \mathbf{EX}^T\mathbf{B}^T$$

from which one can compute the correlation matrix of $\mathbf{Y}$, after centering and normalization of the voxel time series, which yields:

$$\widehat{\text{Corr}}(\mathbf{Y}) = \Delta_{\mathbf{Y}}^{-\frac{1}{2}}\mathbf{HYY}^T\mathbf{H}\Delta_{\mathbf{Y}}^{-\frac{1}{2}},$$

where $\mathbf{H}$ is the centering matrix $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{uu}^T$, $\mathbf{u}$ being the unit vector, and $\Delta_{\mathbf{Y}}$ is the diagonal matrix with the same diagonal as $\mathbf{HYY}^T\mathbf{H}$

To go one step further, one can further assume that the voxels in the region of interest share the same covariance matrix $\mathbf{\Sigma}$ (in the sample dimension). Let us also assume that the null hypothesis is true, i.e. that the effect $\mathbf{B}$ is null. It follows that $\mathbb{E}(\mathbf{YY}^T) = p\mathbf{\Sigma}$. Hence,

$$\widehat{\text{Corr}}(\mathbf{Y}) \to \Delta_{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{H\Sigma H}\Delta_{\mathbf{\Sigma}}^{-\frac{1}{2}} \tag{4}$$

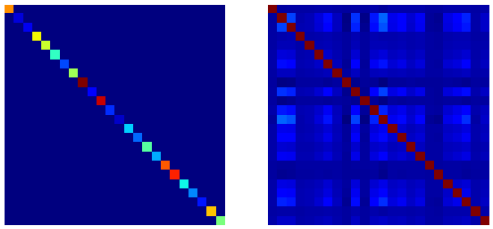

Figure 1. Heteroscedasticity and non-diagonal correlation matrix: (left) a generic diagonal matrix covariance $\mathbf{\Sigma}$ yields a non diagonal correlation matrix (right) due to centering of $\mathbf{Y}$ across samples and the normalization of correlation values.

in the large p limit, where $\Delta_{\mathbf{\Sigma}}$ is the diagonal matrix with the same diagonal as $\mathbf{H\Sigma H}$. Given that

$$R_{RSA} = \frac{1}{2} Tr\left((\widehat{\text{Corr}}(\mathbf{Y}) - \mathbf{I})\mathbf{K} \right), \qquad (5)$$

the RSA statistic asymptotically reflects the non-diagonal terms of $\Delta_{\mathbf{\Sigma}}^{-\frac{1}{2}} \mathbf{H\Sigma H} \Delta_{\mathbf{\Sigma}}^{-\frac{1}{2}}$.

The key point is that, even if $\mathbf{\Sigma}$ is diagonal, the centered and normalized matrix is not (see Fig. 1). Hence, owing to the structure of $\mathbf{\Sigma}$, it can be positively or negatively correlated with $\mathbf{K}$, leading to positive or negative correlations. By contrast, under the null hypothesis, the linear model statistic converges asymptotically to $Tr(\mathbf{X}\mathbf{X}^{\dagger}\mathbf{\Sigma})$ and thus measures the proportion of variance in $\mathbf{\Sigma}$ that is fit by span($\mathbf{X}$), without any additional bias. This means that the RSA decision is sensitive to heteroscedasticity of the noise, namely the fact that the noise varies across conditions/voxels. However, such variations are not unexpected, due to the artefactual effects that affect the BOLD signal (motion, tiredness, fluctuations in vigilance), yet they should not be confused with actual task-related BOLD signal increases.

3. Experiments and results

In a simulation experiment, we first exhibit the kind of issue that can affect RSA type inference in heteroscedastic noise. We then turn to a real dataset and show that the heteroscedasticity issue is actually not negligible.

3.1. Simulated data

We generate a simplistic dataset, where $p = 100$ voxels are observed during the occurrence of a certain paradigm with $n = 24$ samples. The paradigm is simply assumed to be a linear function of the sample number. We rely on two Gaussian noise models: one i.i.d., hence *homoscedastic*; the second one heteroscedastic, with variance increased by a factor of 2 for the second half of the samples. Note that these fluctuations are not perfectly correlated with the assumed paradigm and cannot be viewed as a functional signature. The R_{RSA} and R^2 statistics are evaluated, and the statistical significance is computed with a permutation test with $J = 10^4$ permutations. The experiment is repeated 10^3 times. We present a histogram of these p-values in Fig 2. Note that these histograms are expected to be flat, because no effect was simulated.

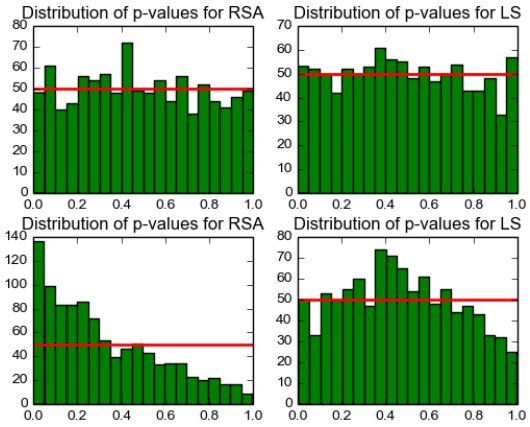

Figure 2. Results of the simulation under the null hypothesis: histogram of the p-values. (Top) under homoscedastic noise conditions, the distribution of the p-values under the null hypothesis is flat. (Bottom) under heteroscedastic conditions, the distribution becomes non-flat both for RSA and the linear model. However, the accumulation of very low p-values occurs only for RSA, meaning that the false positive rate is no longer under control.

Indeed, under homoscedastic noise conditions, the distribution of the p-values under the null hypothesis is flat. By contrast, under heteroscedastic conditions, the distribution becomes non-flat both for RSA and the linear model. However, the accumulation of very low p-values occurs only for RSA, meaning that the false positive rate is no longer under control, while the linear model does not yield too many low p-values.

3.2. Experimental data

We use the dataset described in (Haxby et al., 2001), in which subjects were viewing images from 8 different categories (shoes, bottles, places, faces, cats, scissors, scrambled pictures). We study the representation of these categories in the visual cortex by using a predefined parcellation using the Harvard-Oxford atlas, split between hemispheres (96 regions). We obtain the p-values of the association test between conditions by using either a linear model or an RSA approach, by computing the statistics defined previously and a permutation test (where the labels

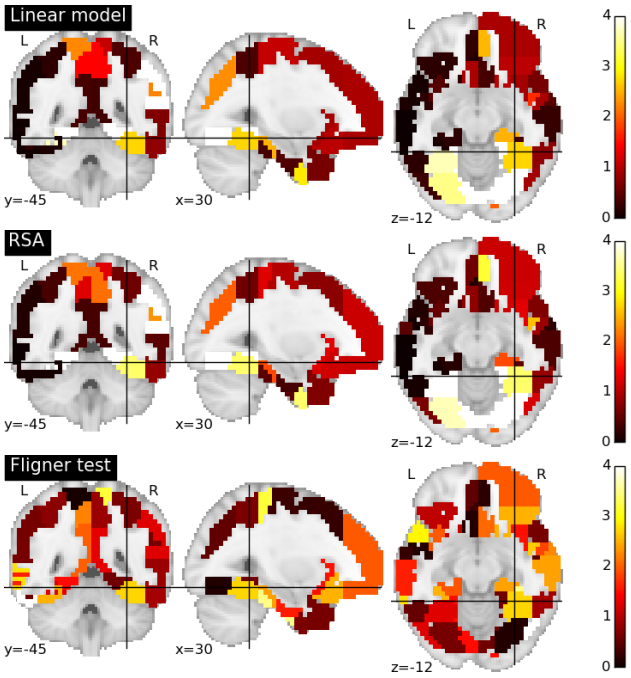

Figure 3. Differences between RSA and linear encoding models and noise heteroscedasticity. (Up) $\log_{10} p$-value of the encoding approach based on a linear model; (middle) $\log_{10} p$-value of the RSA test (down) $\log_{10} p$-value of the Fligner test that detects difference of variances across blocks. While the RSA and encoding model yield mostly similar results, there is a difference of significance e.g. in the anterior fursiform gyrus; in parallel, one observe significant variance fluctiations with the Fligner test.

are shuffled across sessions, but the within-session block structure is preserved), with $J = 10^4$ permutations. In parallel, we display the result of a Fligner test that detects variance difference across blocks. We used different chunks of 4 sessions in one subject, and obtained similar outcomes: we display in Figure 3 their result with sessions 1-4 of subject 1. We rely on Nilearn functions for fetching the data, the atlas and the visualization (Abraham et al., 2014).

One can observe that in general, the RSA and encoding model yield qualitatively similar results, with higher response in the occipital cortex and right orbitofrontal cortex. There are also some differences, with more significant values for RSA, such as in the right anterior fusiform gyrus, but this is associated with significant variance fluctuations across blocks, which calls for some caution when interpreting the results in this area.

4. Conclusion

The present work aimed at *i)* recalling the fact that RSA analysis can in many settings be handled using explicit linear encoding models *ii)* uncovering some differences between the significance of associations observed using encoding models and RSA, where the main difference lies in the way to handle the neuroimaging data: an explicit linear fit in the case of encoding model (possibly regularized if necessary), against a comparison of correlation structures of RSA. The latter, implicit approach suffers from the ill-controlled behavior of correlations, e.g. when the data are not i.i.d in the time/sample dimension. For the sake of simplicity, we assumed a relatively simple structure of the set of stimuli: namely that of a low-rank model (i.e. a design matrix with few columns). However, we acknowledge that this does not represent all possible use cases and defer the investigation of more complex analyses (full rank square design matrices) to future work.

Finally, it should be emphasized that the cognitive problems to be addressed are usually much more subtle than those discussed here. For instance, they may involve the use of several concurrent variables, where the selective association of some of them with the neuroimaging data has to be established. However, permutation testing with multiple independent variables is notoriously hard to handle in the case of linear models (Anderson & Robinson, 2001), hence arguably more problematic with RSA.

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n 604102 (HBP), from the joint Inria-Microsoft Research lab (*MediLearn* project) and from the ANR funding agency, BrainPedia project ANR-10-JCJC 1408-01.

References

Abraham, Alexandre, Pedregosa, Fabian, Eickenberg, Michael, Gervais, Philippe, Mueller, Andreas, Kossaifi, Jean, Gramfort, Alexandre, Thirion, Bertrand, and Varoquaux, Gaël. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform*, 8:14, 2014. doi: 10.3389/fninf.2014.00014.

Anderson, Marti J and Robinson, John. Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88, 2001.

Borghesani, Valentina, Pedregosa, Fabian, Eger, Evelyn, Buiatti, Marco, and Piazza, Manuela. A perceptual-to-conceptual gradient of word coding along the ventral path. In *Pattern Recognition in Neuroimaging*, Tubingen, Germany, June 2014. IEEE.

Cox, David D. and Savoy, Robert L. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in

human visual cortex. *Neuroimage*, 19(2 Pt 1):261–270, Jun 2003.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, Sep 2001. doi: 10.1126/science.1063736.

Haynes, John-Dylan and Rees, Geraint. Decoding mental states from brain activity in humans. *Nat Rev Neurosci*, 7(7):523–534, Jul 2006. doi: 10.1038/nrn1931.

Kriegeskorte, Nikolaus. Relating population-code representations between man, monkey, and computational models. *Front Neurosci*, 3(3):363–373, 2009. doi: 10.3389/neuro.01.035.2009.

Kriegeskorte, Nikolaus, Mur, Marieke, and Bandettini, Peter. Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Naselaris, Thomas, Kay, Kendrick N., Nishimoto, Shinji, and Gallant, Jack L. Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410, May 2011. doi: 10.1016/j.neuroimage.2010.07.073.

Nili, Hamed, Wingfield, Cai, Walther, Alexander, Su, Li, Marslen-Wilson, William, and Kriegeskorte, Nikolaus. A toolbox for representational similarity analysis. *PLoS Comput Biol*, 10(4):e1003553, Apr 2014. doi: 10.1371/journal.pcbi.1003553.

Varoquaux, Gael and Thirion, Bertrand. How machine learning is shaping cognitive neuroimaging. *GigaScience*, 3:28, 2014. doi: 10.1007/s12021-008-9041-y.