Functional Neuroimaging Group Studies

Bertrand Thirion

Functional neuroimaging group studies

Bertrand Thirion

December 19, 2016

Abstract

Multi-subject statistical analysis is an essential step of neuroimaging studies, as it makes it possible to draw conclusions that hold with a prescribed confidence level for the population under study. The use of the linear assumption to model activation signals in brain images and their modulation by various factors has opened the possibility to rely on relatively simple estimation and statistical testing procedures. Specifically, the analysis of functional neuroimaging signals is typically carried out on a per-voxel basis, in the so-called *mass univariate framework*. However, the lack of power in neuroimaging studies has incited neuroscientists to develop new procedures to improve this framework: various solutions have been set up to take into account the spatial context in statistical inference or to deal with violations of distributional assumptions of the data. In this chapter, we review the general framework for group inference, the ensuing mixed-effects model design and its simplifications, together with the various solutions that have been considered to improve the standard mass-univariate testing framework.

1 Introduction

Most neuroimaging statistical problems deal with the comparison of sets of images that embed some features of brain structure or function across a group of subjects, with the purpose to detect some common effects across individuals or some differences across sub-populations. The most standard framework consists in comparing images on a voxel-by-voxel basis (or a vertex-by-vertex basis if the data is sampled on a mesh), after resampling in a common spatial referential, such as the referential defined by the Montreal NeuroImaging Institute (MNI) template [18], or a coordinate system on the cortical surface [7]. This resampling is assumed to correct for pose and shape differences across individuals, so that possible remaining differences are related to the anatomical or functional feature of interest and not to a mere residual of between-subject registration. Note that this is an assumption, and that the limitations of the brain image coregistration procedures used for this purpose are key to understanding the difficulty of statistical inference in the context of neuroimaging group studies.

The simplest statistical inference procedure for image-based data, a.k.a. mass-univariate inference, relies on the computation of a statistic in each voxel. This statistic is then compared to a reference distribution, that represents its likely values when the hypothetical effect is absent. If the actual statistic value is extreme with respect to this distribution –*its p-value is low*– one can conclude that it would not likely be observed under the null hypothesis, hence it is more likely explained by some alternative hypothesis: *the null hypothesis is rejected*. This inference scheme

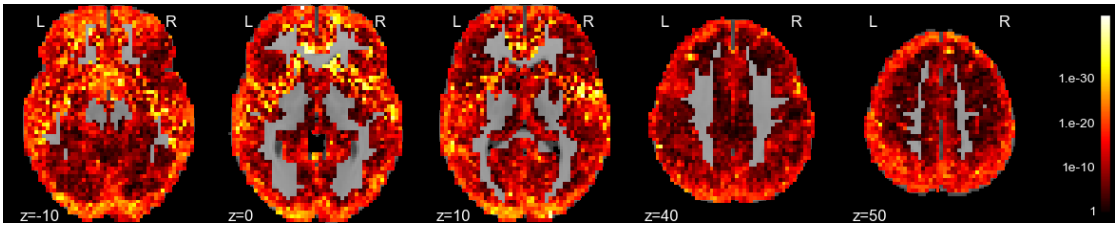

Figure 1: Illustration of the deviation from normality of functional neuroimaging datasets: the map shows the p-value of a test [5] rejecting the normality of the distribution of the z-transformed activation statistics across subjects using the dataset presented in section 7 ($n = 573$). Note that the Gaussian hypothesis is significantly rejected in *all* cortical regions.

is known as *classical statistics*. In practice, the question of how low the p-value should be to support the conclusion is dealt with arbitrarily, and the corresponding choice ($p < .05$ is commonly accepted in the literature) cannot be justified within the framework of classical statistics [23].

The choice of the decision statistic is very important. As many studies involve comparing the mean value of a given image-derived feature across populations, parametric tests such as Student's t-test or Fisher's F-test are natural choices. However, they are limited in two regards:

- In order to yield accurate p-values, they involve a *normality hypothesis* which is most often violated (see e.g. Fig. 1). Depending on how whether standard parametric statistics are robust enough to these deviations or other non-parametric statistics should be used instead is an important question. In practice, however, the price to pay in terms of computation cost or loss of sensitivity when using non-parametric inference is most often too high, while accurate p-values for t or F tests can be obtained under weak assumptions by *permutation testing* [2, 21].
- The individual values that are compared in the test are not directly observed; instead they are estimated with some level of uncertainty from the acquired data. Taking into account the uncertainty in the statistical evaluation leads formally to a *mixed-effects model* [19]. Such a model is more accurate, but also arguably more expensive than simple random effects model. We review the mixed-effects formulation and its simplifications in detail in section 4.

Then comes an issue that is standard in statistics, yet particularly problematic neuroimaging, namely that of *multiple comparisons*: since many tests are performed simultaneously, the risk of observing low p-values by chance, hence of making false detections, is high. Depending on the statistical guarantees required for the analysis, a suitable correction can be implemented, such as family-wise error control (of which the *Bonferroni correction*, that corrects the significance level by the number of tests performed, is the simplest example) or false discovery rate control. Again, the most reliable procedure is given by permutation tests (see section 6).

One of the major issues with image-based statistical works is that those are often carried out on small samples of subjects; for instance, most cognitive neuroimaging findings are based on cohorts of no more than 20 subjects, due to the cost of data acquisition and processing. This situation leads to a degradation of the ability to separate signal from noise; in practice, enforcing a strict control on type I errors jeopardizes the analyst's ability to detect the actual signal. As a result, neuroimaging studies typically suffer from a *lack of power* and in a *low reproducibility* of the

findings [3].

For the sake of sensitivity, the mass-univariate setting can be enhanced by taking into account the image context in statistical inference. Smoothing is a relatively standard image analysis procedure, but it strongly biases the shape of the signal of interest and thus can only be used sparingly. However, on can also consider the continuous structure of the signals of interest embedded in the images by focusing on the size of the connected components of supra-threshold areas for a given detection threshold. Assuming that the distribution of such sizes under the global null hypothesis (that there is no effect present in any region of the image) is known, observing larger region sizes typically indicates the presence of an effect in these regions. This type of cluster-level inference has become a standard in neuroimaging [26]; it has also been extended to general procedures that avoid the prior selection of a cluster-forming threshold [31]. This and other related procedures are discussed in section 5. We give a brief account of permutation testing approaches in Section 6 and conclude in Section 7 with illustrating examples.

2 Variability of brain shape and function

Variability of brain organization and brain imaging Neuroimaging group analyses test the effect of some external variables of interest on the image signal, i.e. they compare the amount of signal explained by the variables of interest to the residual signal after fitting a linear model. They do thus measure the fraction of between-subject signal variance that is correlated with the target variable. However, this variability has a complex nature, as it encodes functional and structural features unique to each individual, together with limitations or various signal corruptions in the imaging process; in any case, it cannot be simply conceptualized as an additive random noise in the observations. Building suitable imaging features that achieve some robustness against the observed between-subject variability is thus in important challenge. It requires some efforts to understand and capture the information of interest in the presence of distortions and structured noise.

Variability of brain shape The variability of brain size and shape is mostly observed through anatomical imaging and various computational geometric procedures that measure the thickness, the regularity of the cortical surface, or some of its singularities (sulcal pitts, sulci fundi etc.). The variability of such features readily poses a challenge for the comparison of brains from different subjects: what makes each brain location unique from an anatomical perspective ? Or put differently, how to warp each individual brain such that the localized individual features can be considered as corresponding to each other ? The best approaches so far consist of first compensating for differences in image pose and brain size through linear transformations, then using high-dimensional diffeomorphic registration to align individual gray matter outline approximately [1], and then to perform statistical analysis, yet in the absence of further guarantee on the identity of the tested structures. This framework results in an uncertainty of about 10mm on the actual voxel correspondences, which can be taken as a blur on the results of any group analysis [17, 33]. Current attempts to improve upon this situation rely on surface-based mapping [8] –yet without any formal guarantee of accurately aligning cyto-architectural areas nor functional areas– or using functional localizer experiments to define individual regions of interest [22].

A point relevant for all neuroimaging studies is that these differences are not

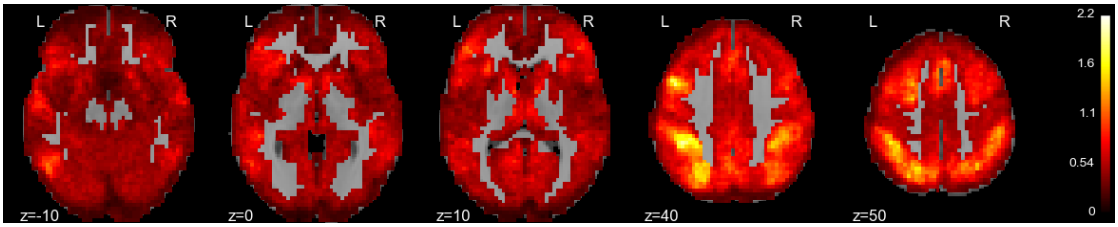

Figure 2: Illustration of the relative magnitude of within- and between-subject variability, through the average between/within variance ratio of the dataset presented in section 7. While the ratio is close to 1 in many regions, it is larger in regions that display a non-zero mean effect across the population (compare with Fig. 5).

modeled, because current imaging contrasts are not sufficient to disambiguate the nature and organization of all brain areas; it is also clear that human cortical folding cannot be mapped diffeomorphically across individuals [28]. The corresponding variability is thus inaccurately handled as unstructured noise.

Variability of brain function Besides the variability of brain shape and anatomical organization, different subjects may display different brain activation patterns, which can be interpreted as differences in functional organization, cognitive strategies, attention, or more simply signal-to-noise ratio of the imaging data related e.g. to presence of motion or various acquisition artifacts. In the absence of external data, these sources of variability cannot be identified easily, nor can they be removed. They are often considered as an additional additive noise. Note that functional MRI is not a quantitative modality, in the sense that the corresponding measurements of the BOLD (blood oxygen-level dependent) signal are not expressed in absolute physical units. Practitioners have found that expressing the signal fluctuations as a percentage of the baseline level was a practical measure, yet it is unclear how invariant the resulting quantification is to parameters of no interest in statistical analysis, such as MR sequence parameters (see an illustration in Fig. 5(b)).
A rough measure of the between-subject variability can be given by comparing the amount of between-subject variance to average amount of within-subject variance (related to observation noise). An example of such a ratio is displayed in Fig. 2.

3 Mixed-effects and fixed-effects analyses

Fig. 2 illustrates a situation that arises frequently in functional neuroimaging, namely the presence of two or more heterogeneous sources of variance in the data. One is related to the observation process and can be viewed as noise, while the other is related to the difference between individuals, as discussed in Section 2. We refer to the corresponding variance estimates as *first-level* and *second-level* variance respectively. This sources of variability are jointly embedded in the observations, yet the terms are separated in the so-called mixed effects modeling framework [19], which we will discuss in detail in section 4.

It should be noted that one has the possibility to neglect the second-level variance, which leads to a *fixed-effects model*: such a model typically assumes that a given effect has been observed in all subjects as if it were a repeated measurement on the same individual, and thus aims at deriving the mean effect and variance

of this common effect without considering cross-subjects fluctuations. Such an inference cannot be used as a population-level inference, given that it ignores the variability that stems from the subject effect [10]. The difference between fixed- and mixed-effects inference is illustrated in Fig. 3.

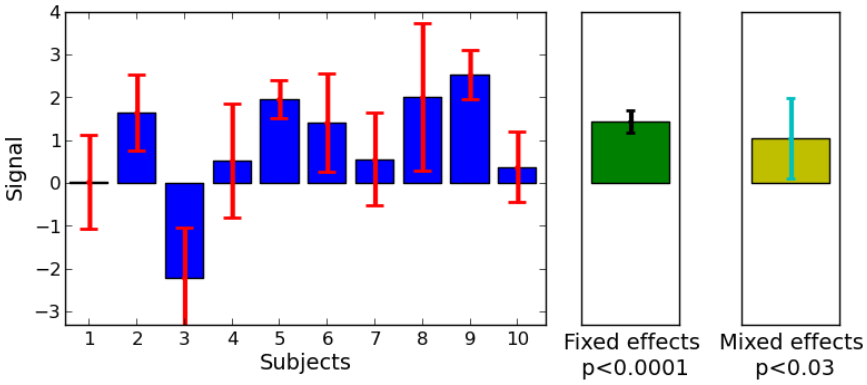

Figure 3: Illustration of the difference between fixed- and mixed-effects inference: Given $n = 10$ observations associated with a given level of uncertainty (left), one can perform a fixed-effects inference that ignores cross-subject variability of the observations and thus leads to population effects with tight uncertainty (middle), or consider this variance and then obtain wider uncertainty estimates (right). Only the mixed effects model yields a valid inference on the population from which the observations were sampled.

4 Group analysis for functional neuroimaging

In this section, we first review the two-level linear model for functional neuroimaging, then discuss the estimation of the mixed-effects model and ensuing statistical tests.

4.1 Problem setting and notations

As a preliminary note, it is assumed here that some functional data are observed at a given set of brain locations across multiple subjects, be these locations cortical surface nodes or voxels (see e.g. [36]).

To clarify the notations, we use bold capital letters for matrices (e.g. $\mathbf{A}$), bold letters for vectors (e.g. $\mathbf{a}$) and small letters for scalars (e.g. a). We denote $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Given n scalars $(\sigma_1^2, .., \sigma_n^2)$, we denote by $\mathrm{diag}(\sigma_1^2, .., \sigma_n^2)$ the $(n \times n)$ diagonal matrix with these scalars on the diagonal. We use the following conventions: $n_{subjects}$ denotes the number of subjects, p the number of voxels and $m(s)$ the number of scans of a given dataset in a given subject s.

Within-subject model We consider a study performed over $n_{subjects}$ subjects. For any subject $s \in [n_{subjects}]$, let $\mathbf{Y}_s$ denote a set of observations (fMRI scans) obtained in this subject. To simplify the statistical formalism, each fMRI scan is *flattened* to a vector representation, where each coordinate of the vector represents the fMRI activity in a given voxel within a suitable brain mask. Let us denote $m(s)$ the length of the time series in subject s. $\mathbf{Y}_s$ is thus a matrix of shape $(m(s) \times p)$ and the value $\mathbf{Y}_s$ (i, j) for $1 \leq i \leq m(s)$ and $1 \leq j \leq p$ represents the fMRI signal

at voxel j acquired at time i. Let $(\mathbf{X}_s)_{s\in[n_{subjects}]}$ represent the design matrices that model experimental and nuisance effects that are likely to be reflected in brain signals. The first level general linear model (GLM) takes the form

$$\forall s \in [n_{subjects}], \mathbf{Y}_s = \mathbf{X}_s\mathbf{B}_s + \boldsymbol{E}_s, \tag{1}$$

where $(\mathbf{B}_s)$ for each $s \in [n_{subjects}]$ is a matrix of shape $(n_{reg} \times p)$ that yields the coefficients associated with the columns of the design matrix and $(\boldsymbol{E}_s)$ is the unmodeled signal, considered as observation noise.

In practice, only a certain combination of the parameters is of interest and will be the subject of further inference. For instance, the effects associated with motion parameters may not be considered in population-level analyses, while the activation signal associated with a combination of experimental conditions quantifies a cognitive response of interest that one wishes to compare with individual characteristics. To keep the setting clear yet comprehensive, we thus rewrite equation 1 with split design and parameter matrices to introduce a distinction between the coefficients of interest and those modeling other effects. Note that singling out a parameter of interest amounts to defining a *contrast*, i.e. a linear combination of the effects $\mathbf{B}_s$, that represents it in the data. Here we assume that the design matrix is written in a form such that the contrast of interest corresponds to the first column $\mathbf{x}_s^i$ of $\mathbf{X}_s$, i.e. $\mathbf{X}_s = [\mathbf{x}_s^i, \mathbf{X}_s^r]$; correspondingly, the effects of interest and residual effects are written written $\mathbf{b}_s$ and $\mathbf{B}_s^r$, i.e. $\mathbf{X}_s\mathbf{B}_s = \mathbf{x}_s^i\mathbf{b}_s + \mathbf{X}_s^r\mathbf{B}_s^r$, thus yielding

$$\forall s \in [n_{subjects}], \mathbf{Y}_s = \mathbf{x}_s^i\mathbf{b}_s + \mathbf{X}_s^r\mathbf{B}_s^r + \boldsymbol{E}_s \tag{2}$$

Group-subject model At the population level, it is expected that the contrasts of interest, observed across subjects, could potentially be explained by subject-dependent variables, such as age, behavioral tests or genetic variables. We note $\mathbf{B} = [\mathbf{b}_1, .., \mathbf{b}_{n_{subjects}}]^T$ the $(n_{subjects} \times n_{voxels})$ matrix that represents the contrasts of interest measured across individuals. The explanatory variables of interest are grouped in a second level design matrix $\mathbf{Z}$, that has a shape $(n_{subjects} \times n_{factors})$

$$\mathbf{B} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}, \tag{3}$$

where $\boldsymbol{\beta}$ is a matrix of shape $(n_{factors} \times n_{voxels})$ that represents the population-level effects. Without loss of generality, Eq. 1-3 can thus be rewritten

$$\mathbf{Y}_s = \mathbf{x}_s^i\mathbf{b}_s + \mathbf{X}_s^r\mathbf{B}_s^r + \boldsymbol{E}_s, \ \forall s \in [n_{subjects}] \tag{4}$$

$$\mathbf{B} = [\mathbf{b}_1, .., \mathbf{b}_{n_{subjects}}]^T = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}} \tag{5}$$

Finally, the question of interest is where in the brain a certain combination of the factors of interest yields a positive effect on average in the population, i.e., whether $\mathbf{c}^T\boldsymbol{\beta} > 0$, where $\mathbf{c}$ is a suitable vector of contrast on the population-level effects: if $\mathbf{Z}$ contains three variates related to the sex, the age of the subjects and an intercept, setting $\mathbf{c} = (0, 1, 0)$ will display the effect of age on the observed BOLD signal. Different types of contrasts correspond to different statistical questions, see appendix A for an overview. The problem consists thus in estimating with which confidence one can reject the null hypothesis $\mathbf{c}^T\boldsymbol{\beta} = 0$. Since the model defined in Eqs. 4-5 can be handled at each brain location independently, we proceed with voxel-level analysis (estimation and statistical analysis). We keep focusing on voxel-level probabilistic assessment till the end of this section; taking into account the joint signal distribution over voxels is deferred to section 5: such an approach is typical of a mass-univariate modeling frameworks.

Interpretation as a mixed effects model It is straightforward to observe that Eqs. 1-3 can be concatenated into one equation:

$$\forall s \in [n_{subjects}], \mathbf{Y}_s = \mathbf{x}_s^i \mathbf{Z}|_s \boldsymbol{\beta} + \mathbf{x}_s^i \boldsymbol{\mathcal{E}}|_s + \mathbf{X}_s^r \mathbf{B}_s^r + \boldsymbol{E}_s. \quad (6)$$

where $\mathbf{Z}|_s$ and $\boldsymbol{\mathcal{E}}|_s$ denote the restrictions of matrix $\mathbf{Z}$ and $\boldsymbol{\mathcal{E}}$ to their row s. This means that the observed data $\mathbf{Y}_s$ are actually composed of four different effects:

- the effect of interest, to be tested: $\mathbf{x}_s^i \mathbf{Z}|_s \boldsymbol{\beta}$
- A random subject effect $\mathbf{x}_s^i \boldsymbol{\mathcal{E}}|_s$
- Some within-subject effects of no interest or nuisance effects, handled as fixed effects $\mathbf{X}_s^r \mathbf{B}_s^r$
- Some observation noise $\boldsymbol{E}_s$.

The model thus comprises both fixed and random effects, hence the reference to a *mixed effects* model.

4.2 Estimation

It is then generally assumed that the two random components are Gaussian distributed, with unknown variance: let $(\boldsymbol{\Lambda}_s)$ be the $(m \times m)$ variance-covariance matrix of the noise $\boldsymbol{E}_s$ in a subject $s \in [1..n_{subjects}]$ and $\boldsymbol{\Delta}$ be the variance-covariance matrix $(n_{subjects} \times n_{subjects})$ of the random effects. While $\boldsymbol{\Lambda}_s$ are often taken as the covariance matrix of a (temporal) auto-regressive process scaled by an unknown variance, $\boldsymbol{\Delta}$ can be assumed to be diagonal, as $(\boldsymbol{\mathcal{E}}|_1, .., \boldsymbol{\mathcal{E}}|_{n_{subjects}})$ have been sampled independently. In the absence of additional information on the population structure, $\boldsymbol{\Delta} = \gamma^2 \mathbf{I}_{n_{subjects}}$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. The estimation of the parameters of the mixed effects model, is carried out, generally following the maximum likelihood principle. The parameters to estimate are

$$\Theta = \left((\boldsymbol{\Lambda}_s, \mathbf{B}_s^r), \boldsymbol{\beta}, \gamma^2 \right), \quad (7)$$

given $(\mathbf{Y}_s, \mathbf{X}_s)_{s=1..n}$ and $\mathbf{Z}$:

$$\hat{\Theta} = \text{argmax}_\Theta \, \mathcal{N}(\mathbf{B}; \mathbf{Z}\boldsymbol{\beta}, \gamma^2 \mathbf{I}_{n_{subjects}}) \prod_{s=1}^{n_{subjects}} \mathcal{N}(\mathbf{Y}_s; \mathbf{x}_s^i \boldsymbol{\beta}_s + \mathbf{X}_s^r \mathbf{B}_s^r, \boldsymbol{\Lambda}_s)$$

For the sake of simplicity and computation efficiency, it is frequent to use a two-step method: first, the estimation of the effects of interest and their covariance, independently in each subject, and then the estimation of the population parameters. Standard solutions include: the EM algorithm [41, 29, 36], Bayesian methods [39] or the Gauss-Newton method or some variant thereof [35]. Here we follow the EM approach. First level model fit (for notational simplicity, we assume that $\boldsymbol{\Lambda}_s$ is given, while it is voxel-specific and data-dependent in practice). The first-level estimation procedure yields estimates $(\hat{\mathbf{b}}_s, \hat{\sigma}_s^2)$ of the individual effects and associated variance. These estimates have a large number of degrees of freedom $m(s) - n_{reg}$. The group model (Eq. 4-5) then boils down to

$$\begin{aligned} \hat{\mathbf{b}}_s &= \mathbf{b}_s + \mathbf{e}_s, \ \forall s \in [n_{subjects}] \\ \mathbf{B} = [\mathbf{b}_1, .., \mathbf{b}_{n_{subjects}}]^T &= \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}, \end{aligned}$$

where for each subject $s \in [n_{subjects}], \sigma_s^2 = var(\mathbf{e}_s)$ is estimated with a large number of degrees of freedom, hence we assume that it is exact, while $\gamma^2 = var(\boldsymbol{\mathcal{E}})$ is

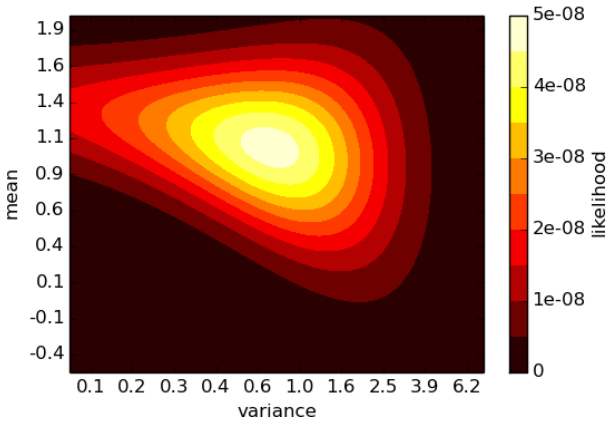

Figure 4: Likelihood of the observations displayed in Fig. 3 as a function of the parameters $(\boldsymbol{\beta}, \gamma^2)$ (which are 2 scalars in that case). It can be seen that the likelihood function has a unique maximum.

unknown, and has to be estimated. $(\boldsymbol{\beta}, \gamma^2)$ can then be estimated so that they maximize the likelihood

$$\mathcal{L}(\hat{\mathbf{B}}; \boldsymbol{\beta}, \gamma^2) = \mathcal{N}\left(\hat{\mathbf{B}}; \mathbf{Z}\boldsymbol{\beta}, \gamma^2 \mathbf{I}_{n_{subjects}} + \text{diag}(\sigma_1^2, .., \sigma_{n_{subjects}}^2)\right) \tag{8}$$

using an EM algorithm: If we denote $\mathcal{N}(\tilde{\mathbf{b}}_s, \tilde{\mathbf{s}}_s^2)$ the variational distribution of $\mathbf{b}_s$, and $\tilde{\mathbf{B}} = [\tilde{\mathbf{b}}_1, .., \tilde{\mathbf{b}}_{n_{subjects}}]$, the EM algorithm consists in iterating the two steps:

$$\text{E-step:} \quad \forall s \in [1..n_{subjects}],\ \tilde{\mathbf{s}}_s^2 = \left(\frac{1}{\gamma^2} + \frac{1}{\sigma_s^2}\right)^{-1},\ \tilde{\mathbf{b}}_s = \tilde{\mathbf{s}}_s^2 \left(\frac{\hat{\mathbf{b}}_s}{\gamma^2} + \frac{\boldsymbol{\beta}}{\sigma_s^2}\right) \tag{9}$$

$$\text{M-step:} \quad \boldsymbol{\beta} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\tilde{\mathbf{B}},\ \gamma^2 = \frac{1}{n_{subjects}}\left(\sum_{s=1}^{n_{subjects}} \tilde{\mathbf{s}}_s^2 + \|\tilde{\mathbf{B}} - \mathbf{Z}\boldsymbol{\beta}\|^2\right) \tag{10}$$

It converges toward a local maximum of the likelihood function. Although there is no guarantee that the reached maximum is global, it can be observed in many cases that the likelihood function only has one maximum. See e.g. in Fig. 4 the likelihood as a function of the parameters $(\boldsymbol{\beta}, v)$ obtained for the toy data used in Fig. 3.

It should be noted that the above algorithm may not yield the optimal estimators for the variance parameters; for instance, more accurate estimates of the variance may be obtained by using Restricted Maximum likelihood approaches [14].

4.3 Statistical inference

The main question remains to test whether the effect of interest is different from zero, i.e. $\mathbf{c}^T\boldsymbol{\beta} > 0$ or $\mathbf{c}^T\boldsymbol{\beta} \neq 0$. It is not obvious what test should be used. A Wald statistic can be computed as

$$t = \frac{\mathbf{c}^T\boldsymbol{\beta}}{\sqrt{\mathbf{c}^T(\mathbf{Z}^T W^{-1}\mathbf{Z})^{-1}\mathbf{c}}}, \tag{11}$$

where $W(\gamma) = \gamma^2 \mathbf{I}_{n_{subjects}} + \text{diag}(\sigma_1^2, .., \sigma_{n_{subjects}}^2)$. However, it does not conform exactly to a student distribution under the null hypothesis [39, 4]. This means

that non-parametric methods have to be used to obtain an estimate of its null distribution. Given this requirement, the likelihood ratio statistic – that is known as the most powerful test for the model– is the best possible test. This is defined as follows

$$\Lambda = 2\left(\sup_{\boldsymbol{\beta},\gamma} \log \mathcal{L}(\hat{\mathbf{B}}; \boldsymbol{\beta}, W(\gamma)) - \sup_{\boldsymbol{\beta}: c^T\boldsymbol{\beta}=0,\gamma} \log \mathcal{L}(\hat{\mathbf{B}}; \boldsymbol{\beta}, W(\gamma))\right) \tag{12}$$

where $\mathcal{L}$ is defined as in equation 8. Λ is readily computed by running the EM algorithm twice, once in a constrained mode ($c^T\boldsymbol{\beta} = 0$), once in an unconstrained mode.

4.4 The random effects t-test

A crude approximation of the above model consists in neglecting the first-level variance; this amounts to assuming that the individual variances are identical ($\sigma_1^2 = \cdots = \sigma^2_{n_{subjects}}$) and clearly simplifies the ensuing statistical inference. The assumption yields a simple *random effects model*:

$$\mathbf{B} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}', \tag{13}$$

where $\boldsymbol{\mathcal{E}}' \sim \mathcal{N}(0, g^2\mathbf{I}_{n_{subjects}})$, for which the decision statistic is simply

$$t_{RFX} = \frac{\mathbf{c}^T\boldsymbol{\beta}}{\sqrt{g^2\mathbf{c}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{c}}}, \tag{14}$$

where $\boldsymbol{\beta}$ and g^2 are easily estimated with a least-squares fit. Note that it is the most widely used model for population level inference in neuroimaging. Similarly, an F statistic can be defined when one is interested in unsigned or multi-dimensional contrasts In that case, and under the Gaussian hypothesis, the log-likelihood ratio defined in equation 12 is a monotonous function of the Fisher statistic, or, if one is interested in signed effects, of the t statistic. We provide an empirical comparison of the t statistic with the full mixed-effects model in section 7. We also provide a table of the most frequently used statistical models (one-sample t-test, two-sample t-test, paired-t-test, two way ANOVA) in table 1. See [25] for a more complete discussion on this topic.

5 Taking into account the spatial context in statistical inference

Multiple comparisons in the mass univariate framework So far, we have considered each voxel independently, which is a requirement at the estimation stage, given the computation cost of the procedure, but comes at the price of low sensitivity when performing statistical inference. In the mass-univariate setting, one statistical test is performed per voxel, and the significance is then simply corrected for the complete set of tests: Corrected p-values can be obtained by family-wise error-rate (FWER) control, i.e. the probability of performing one false detection, or the generally more lenient false discovery rate (FDR) control, that controls for the proportion of false discoveries among the detections. We discuss the practical computation of FWER-corrected thresholds in section 6. For false discovery rate control, a simple reference procedure is presented in the neuroimaging context in

[13] and is most often used due to its simplicity, but non-parametric alternatives exist too [12].

This framework is usually adopted for its convenience and relatively simple interpretability. However, it is limited by its inability to take into account the image structure of the data: the observed cross-subject variability in brain organization limits the relevance of a voxel-by-voxel description of the data, and yields a reduced sensitivity to detect the true effects.

Spatial regularization The first way to take into account the image structure in statistical modeling is to regularize spatially the data: such a regularization is most often implemented with data smoothing in the volume or on the surface, with a 6 to 10mm full width at half maximum (fWHM) kernel. Choosing the adapted kernel directly leads to a bias-variance trade-off: larger kernels reduce the variance yet yield a large bias on the activation peaks position. Smoothing can instead be replaced by Markov Random field modeling [6] or anisotropic smoothing [32], but such procedures are not widely used by practitioners due to the lack of efficient implementations.

Inference on the regions size A second possibility to take into account the image structure in the inference procedure is to consider the size of the connected components of the set of supra-threshold voxels [11], assuming that the chosen threshold (e.g. t>3 for a t-test) separates correctly the peaks of the image from the background noise. The size statistic has to be compared with the reference distribution of the size of such structures obtained when no effect is present, i.e. under the null hypothesis. Such a distribution is in general not known a priori, but it can be approximated under the hypothesis that the data follows a known parametric distribution [26], or more simply by permutation [15] –see section 6. The latter approach is typically recommended. The underlying intuition is that, especially when looking at population-level statistics maps, narrow regions are unlikely to represent truly active brain structures. There are however two major drawbacks with such an approach: First, it depends on an arbitrary cluster-forming threshold, so that the result does not outline intrinsic characteristics of the data. Second, varying this threshold potentially yields abrupt changes in the detected areas, thus leading to a fundamental instability of the results.

It is however relatively popular in the neuroimaging community due to its enhanced sensitivity, which is however attributable to the weakness of the test: it rejects the global null hypothesis for the whole cluster, but does not make it possible to localize the activation within the supra-threshold cluster. Hence, it should only be used with high cluster-forming thresholds [38].

Cluster-mass testing and Threshold-free cluster enhancement (TFCE) An improvement upon these procedures consists in combining the size and height of peak regions in a common statistic [27, 16]. However, more general combinations of extent and height can be used alternatively, resulting in the so-called threshold-free cluster enhancement (TFCE) procedure [31]. Let $v \to \phi(v)$ be the statistical map obtained from the univariate inference step; the TFCE statistic is computed as follows:

$$\text{TFCE(v)} = \int_0^{\phi(v)} z^H a_v(z)^E dz, \tag{15}$$

where $a(z)$ is the size of the connected component of the map thresholded at level z that contains voxel v, H and E are two non-negative numbers. In general $H = 2$ and $E = .5$, as in the reference TFCE implementation in [31]. Note that $H = 1$ and $E = 0$ yields a cluster mass statistic, while $H = 0$ and $E = 1$ yields a generalized cluster size statistic integrated over thresholds. The setting $H = 2, E = .5$ was designed to represent optimally paraboloid-shaped activations.

Randomized Parcellation-Based Inference (RPBI) A recent alternative to TFCE has been proposed in [20], and consists in using multiple parcellations of the brain volume. A *parcellation* is a partition of the brain volume into connected regions *(parcels)*, for instance 1000 or 2000 parcels; taking parcel-based averages of the input signal can be understood as a dimension reduction procedure well suited for smooth images. The parcel-based signal estimation can be carried in a group of subjects in parallel, and thus used in a population model, where standard statistics (e.g. Eq. 11) are computed, resulting in a parcel-based statistical map. While the signal compression brought by parcel-based analysis is certainly beneficial to sensitivity, the resulting map is heavily biased by the initial parcellation specification: to avoid such biases, one has to marginalize out the parcellation choice. The RPBI statistic is thus defined as the voxel-based sum of binary variables testing whether the decision statistic in the parcel including the voxel was above a given threshold or not, computed over a large enough number of initial parcellations. Let $\mathcal{P}$ be a set of parcellations, and V be the set of voxels under consideration. Given a voxel v and a parcellation P, the parcel-based thresholding function θ_t is defined as:

$$\theta_t(v, P) = \begin{cases} 1 \text{ if } F(\Phi_P(v)) > t \\ 0 \text{ otherwise} \end{cases} \tag{16}$$

where $\Phi_P : V \to P$ is the mapping function that associates each voxel with a parcel from the parcellation P. For a predefined test, F returns the F-statistic associated with the average signal of a given parcel (a t or other statistic is also possible). Finally, the aggregating statistic at a voxel v is given by the counting function C_t:

$$C_t(v, \mathcal{P}) = \sum_{P \in \mathcal{P}} \theta_t(v, P). \tag{17}$$

$C_t(v, \mathcal{P})$ represents the number of times the voxel v was part of a parcel associated with a statistical value larger than t across the folds of the analysis conducted on the set of parcellations $\mathcal{P}$. The parameter t should be set to ensure a Bonferroni-corrected control at $p < 0.1$ in each of the parcel-level analyzes. In practice, the results are weakly sensitive to mild variations of t.

6 Type I error control with permutation testing

Permutation testing is the reference approach for statistical inference, as it provides valid and accurate p-values under a typically restricted set of hypotheses. In practice, many statistical procedures described previously (cluster size, TFCE, RPBI) dot not have a well-defined distribution under the null hypothesis. Even the family-wise error control in the presence of correlations does not have a perfectly known distribution and is correctly approximated only in some peculiar settings [40]. While some reference distributions could be estimated or even simulated under the assumption of Gaussian noise with pre-defined covariance – a framework popularized as Gaussian Random Field Theory [11]– the hypotheses involved in such a

procedure are relatively strong, hence easily proved wrong. In practice, the smoothness of the signal can vary across regions of an image, challenging the stationarity of the random field [30]. In general, permutation-based inference is thus the reference solution to obtain accurate control on type I errors.

As the statistics in (11) and (12) are pivotal, the advised procedure is to use Westfall-Young kind of correction for multiple comparisons [37]: family-wise error rate (FWER) corrected p-values are obtained by sampling the null distribution of the maximal statistic through a permutation scheme: let us consider a statistical map $v \rightarrow \phi(v)$ in which the values follow the same distribution under the null hypothesis (this is called the *pivotality* hypothesis). We assume that the data are exchangeable under the null hypothesis: for instance, we compare the mean effect across two populations of subjects, and do not have any other covariate in the data model. Then for any permutation π of the data (i.e. permutation of the rows of the second-level design matrix $\mathbf{Z}$), one can compute the corresponding map $v \rightarrow \phi^{\pi}(v)$, by applying the usual estimation procedure. Assuming that J such permutation-based estimations are carried out the critical threshold for the family-wise-error-corrected p-value α is given by

$$\phi^c(\alpha) = \mathcal{Q}_{1-\alpha}\left(\left(\max_{v\in[1..p]}\phi^{\pi_j}(v)\right)_{j\in[1..J]}\right), \tag{18}$$

Where $\mathcal{Q}_{1-\alpha}$ stands for the $(1-\alpha)$ quantile of the values (in this case, across permutations $(\pi_1, \cdots, \pi_J)$). The significance at level α of the non-permuted statistics $\phi(v)$ is thus assessed by thresholding of these values against ϕ^c_α. Slightly more complex strategies can be used to control the False discovery rate instead [12].

A particular case is when there is no regressor in the model, i.e. the test only consists in the intercept, meaning that the inference is about the mean effect being larger than zero. In that case, permuting the data does not change the statistic. What is done instead is to further assume that the distribution under the null hypothesis is symmetric, and then swap the sign of the observations: for n observations, this generates up to 2^n resamplings.

Limitations of permutation testing. There are however some cases where the use of permutation tests requires some care, because simplistic implementations lead to inaccurate results. One such case is when models include some covariates, on top of the effects of interest: due to the possible correlations between the contrasts of interest and other covariates, the permutation test needs to be handled with specific procedures [9, 38].

Another obvious limitation of permutation testing is its cost: to guarantee a stable estimate of ϕ^c_α in Eq. 18, it is necessary to use $J \gg \frac{1}{\alpha^2}$ permutations.

7 Illustration of various inference strategies on an example

The difference between the output of a mixed-effects and a fixed-effects model is usually large in neuroimaging settings, where inter-subject variability is large with respect to intra-subject variability (see e.g. Fig. 3). As neuroimaging studies are concerned with population-level validity of the findings, fixed-effects inference is ruled out.

By contrast, the difference between the mixed-effects statistic and the random effects statistic (i.e. the test in Eq. 14 as opposed to the test in Eq. 12) is more subtle

and deserves some consideration. Neglecting first-level variance is actually equivalent to assuming that it is equal in all subjects, i.e. that all individual observations are equally reliable (see section 4.4). Hence any difference between truly mixed-effects are random-effects model clearly outlines the impact of the difference in reliability across samples in the population.

Two examples are described next.

- The first example in Fig. 5(a) shows the results of permutation-controlled $\alpha = .05, J = 10^4$ mixed-effects and random-effects tests to detect the mean effect of a computation task (from which a simple reading effect has been subtracted [24]), based on a sample of $n = 30$ subjects.
- The second example in Fig. 5(b) is a two-sample test that targets the differential effect of two MRI scanners on the activation obtained after completing the same task as the first example. More precisely, it aims at detecting difference between a Siemens Trio 3T scanner (data acquired in 2007-2008 [36]) and a 3T Bruker scanner (data acquired in 2003-2006 [24]). The acquisition parameters of the Bruker dataset are: TR= $2400ms$, TE= $60ms$, the matrix size is 64×64, FOV = $24cm \times 24cm$. Each volume consisted of n_a 3-mm- or 4-mm-thick axial slices without gap, where n_a varied from 26 to 40 according to the session. A session comprised 130 scans. The acquisition parameters of the Trio dataset are: TR= $2400ms$, TE= $30ms$, the matrix size is 64×64, FOV = $19.2cm \times 19.2cm$ and the number of slices is 40, while the session duration is unchanged. Note that the two sets of images have been preprocessed in the same way. The cohort studied comprises $n_1 = 72$ and $n_2 = 78$ subjects from the Bruker and Siemens scanners respectively.

In both cases, random- and mixed-effects maps are presented, after thresholding at the $p < .05$ level, corrected for multiple comparisons by a permutation procedure. While the two kinds of maps are clearly similar, one can observe more sensitivity in the case of mixed-effects inference in the two settings:

- In the one-sample test, more active voxels are found in the ventral striatum, the anterior cingular cortex, the insula and the right intraparietal sulcus, that are regions of the dorsal attentional network. Moreover, the detected regions display more significant p-values.
- In the two-sample test, the effects observed in the mixed-effects analysis are wider and more significant in all regions, although this is more visible in the ventral striatum and the thalamus.

Note that the superiority of the mixed model in terms of sensitivity has already been reported in the literature [39, 29, 36, 4].

Next, to assess the impact of spatially-aware statistics, we perform a qualitative comparison of the maps obtained through the different spatial models, coupled with the mixed-effects statistic: peak significance (no spatial context), cluster size significance, TFCE and RPBI. The results are shown in Fig. 6, and are based on the one-sample test example.

They clearly shows that for a given significance level ($p < 0.05$, corrected for multiple comparison through an max-control permutation-based procedure), spatially informed approaches yield more active voxels than voxel-level inference, a gain that can be interpreted as a higher sensitivity. For instance, some temporal or frontal activation foci are not found in the voxel-based approach, but are detected with the other approaches. It should be reminded however that this increase in sensitivity is mitigated by two important drawbacks:

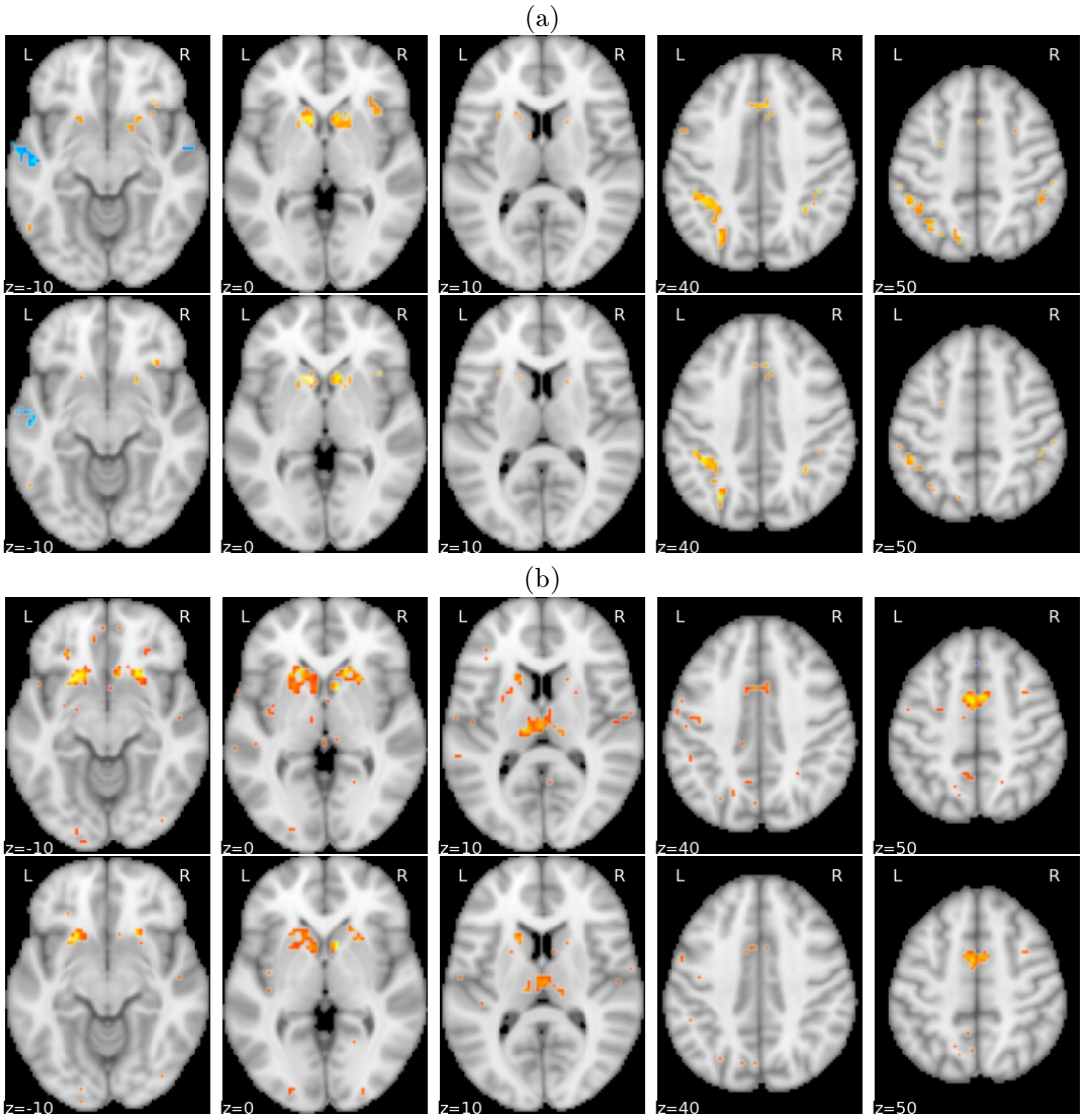

Figure 5: Difference between a mixed-effects model and a random effects model for statistical analysis. (a) One-sample test that aims at detecting the mean effect on a computation task (from which a simple reading effect has been subtracted [24]), based on a sample of $n_{subjects} = 30$ subjects; top: mixed-effects model; bottom: random-effects model. Both maps are thresholded at the $p < 0.05$ level, corrected for multiple comparisons, using permutation testing. (b) Illustration of the difference on a two-sample test, where the differential effect of two MRI scanners on the activation obtained in the same task as (a). The mixed- (top) and random- (bottom) effects maps, both thresholded at $p < 0.05$, FWER-corrected by permutation, show the same effects, but again the mixed-effects inference is more sensitive.

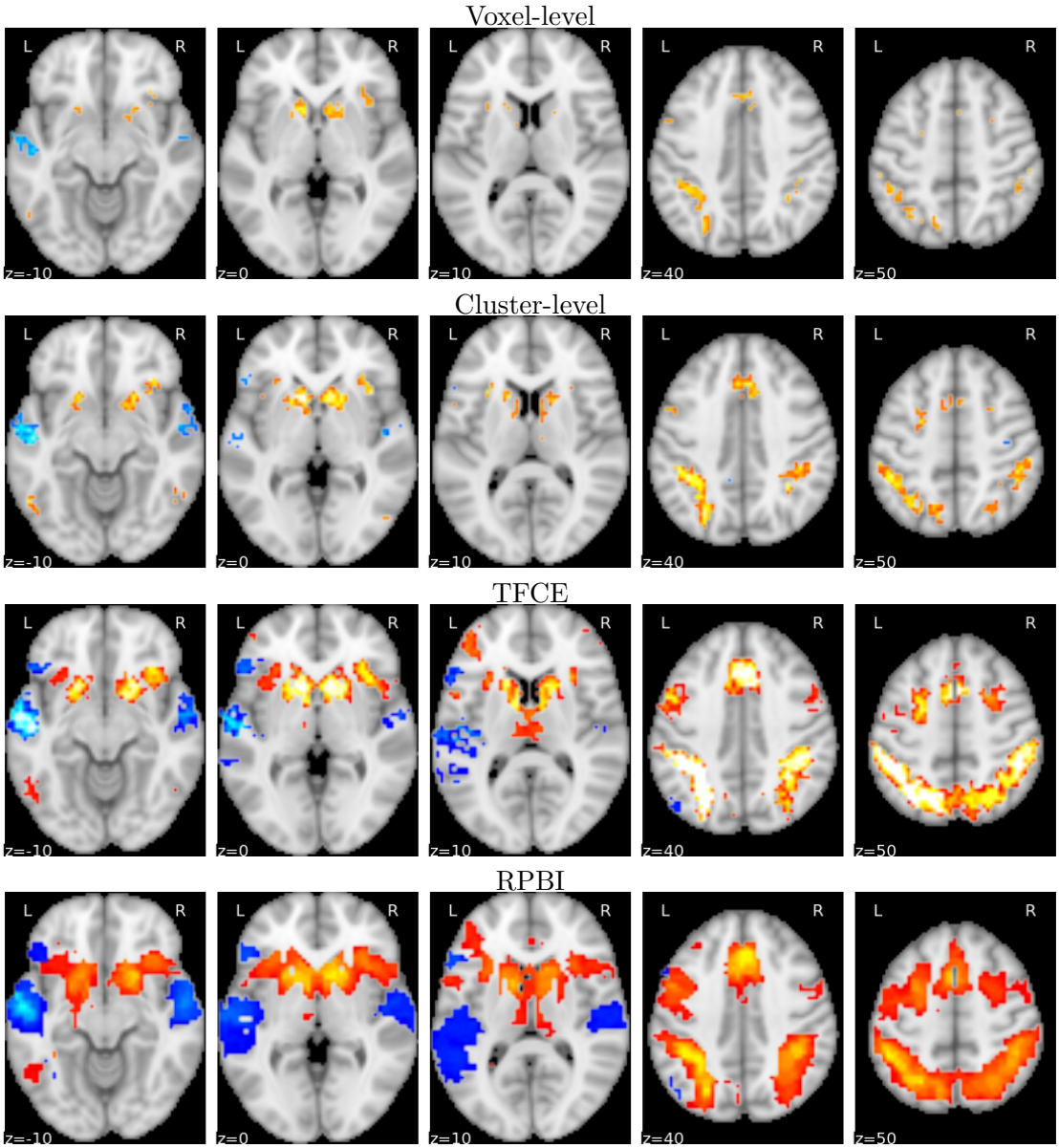

Figure 6: Impact of including the spatial context into neuroimaging statistical inference procedure: the four maps above represent the activation related to the one-sample test presented in Fig. 5(a), thresholded at a significance level of $p<.05$, corrected for multiple comparison through an F-max permutation scheme. Voxel-level, cluster-level, TFCE and RPBI present increasing amounts of activations.

- Only the voxel-level inference provides a guarantee on the active status of each detected voxel, as the other approaches only provide a rejection of the null hypothesis for some (unknown) voxel in the detected areas.
- Non-voxel-based methods require additional parameters. For instance, in our procedure, we relied on a cluster-forming threshold of $p < 10^{-4}$, the E and H parameters of the TFCE (we kept the default $E = .5$, $H = 2$), the internal threshold of the RPBI approach as well as the number of atlases and their resolution. See section 5 for a description of these parameters.

Importantly, it has been reported that the increased sensitivity afforded by spatially-aware method comes with higher reproducibility, i.e. a higher chance of reproducing the results on another sample of images (see e.g. [20]), which is acknowledged as a criterion for model selection in the neuroimaging and other communities [34]. This potentially means that in the low-sample/low-power regime common to most neuroimaging studies, the increase in sensitivity afforded by these methods offsets their drawbacks (see also [35]).

8 Conclusion

The flexibility afforded by the linear model has made it possible for neuroimagers to make inference on brain functional organization and its variability across groups of subjects in a relatively intuitive way. The central concept is arguably that of *contrast*, that corresponds to stating a precise question to identify the effect of variables on brain activation signals. Current gains in computational efficiency has led to the development of permutation-based approaches, that are typically more accurate and make it possible to draw inference from quantities that do not have any known or simple distribution under the null hypothesis.

In practice, neuroscientists often rely on the solution offered by the software that they use most often, such as SPM, FSL or AFNI, which make different choices: SPM relies mostly on parametric statistics, while FSL gives access to the popular TFCE statistic and relies more on permutation tools; AFNI yields a wider choice of statistics. The development and diffusion of methods in open-source software and more user-friendly environments will condition the adoption of the more advanced tools by the community.

Methodological research still has to address some hard challenges, such as the design of more efficient methods to take into account the spatial structure in the images while performing probabilistic inference. Also, the same level of statistical rigor needs to be achieved in more complex statistical analyses that consider the signals from multiple regions, such as multivariate pattern analysis and functional connectivity analysis.

Acknowledgment We are very thankful to Philippe Pinel who provided the data used in the comparison. This work was also possible thanks to the software developments carried out by Benoit da Mota, Virgile Fritsch and Gaël Varoquaux. Many thanks also to Ana Luisa Grilo Pinho who suggested the addition of the statistical tests table. We acknowledge funding from the Human Brain Project, the Digiteo project iConnectom and the Microsoft Research (MediLearn project).

References

[1] John Ashburner and Karl J Friston. Diffeomorphic registration using geodesic shooting and gauss-newton optimisation. *Neuroimage*, 55(3):954–967, Apr 2011. 3

[2] E. Bullmore, M. Brammer, S. C. Williams, S. Rabe-Hesketh, N. Janot, A. David, J. Mellers, R. Howard, and P. Sham. Statistical methods of estimation and inference for functional mr image analysis. *Magn Reson Med*, 35(2):261–277, Feb 1996. 2

[3] Katherine S Button, John P A Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma S J Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*, 14(5):365–376, May 2013. 3

[4] Gang Chen, Ziad S Saad, Audrey R Nath, Michael S Beauchamp, and Robert W Cox. Fmri group analysis combining effect estimates and their variances. *Neuroimage*, 60(1):747–765, Mar 2012. 8, 13

[5] R. D'Agostino and E. S. Pearson. Testing for departures from normality. *Biometrika*, 60:613–622, 1973. 2

[6] X. Descombes, F. Kruggel, and D. Y. von Cramon. Spatio-temporal fmri analysis using markov random fields. *IEEE Trans Med Imaging*, 17(6):1028–1039, Dec 1998. 10

[7] B. Fischl, M. I. Sereno, R. B. Tootell, and A. M. Dale. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp*, 8(4):272–284, 1999. 1

[8] Bruce Fischl, Niranjini Rajendran, Evelina Busa, Jean Augustinack, Oliver Hinds, B. T Thomas Yeo, Hartmut Mohlberg, Katrin Amunts, and Karl Zilles. Cortical folding patterns and predicting cytoarchitecture. *Cereb Cortex*, 18(8):1973–1980, Aug 2008. 3

[9] D. Freedman and D. Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–98, 1983. 12

[10] K. J. Friston, A. P. Holmes, and K. J. Worsley. How many subjects constitute a study? *Neuroimage*, 10(1):1–5, Jul 1999. 5

[11] K. J. Friston, K. J. Worsley, R. S. Frackowiak, J. C. Mazziotta, and A. C. Evans. Assessing the significance of focal activations using their spatial extent. *Hum Brain Mapp*, 1(3):210–220, 1994. 10, 11

[12] Y.C. Ge, S. Dudoit, and T.P. Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12:1–77, 2003. 10, 12

[13] Christopher R Genovese, Nicole A Lazar, and Thomas Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, Apr 2002. 10

[14] David A. Harville. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977. 8

[15] Satoru Hayasaka and Thomas E Nichols. Validating cluster size inference: random field and permutation methods. *Neuroimage*, 20(4):2343–2356, Dec 2003. 10

[16] Satoru Hayasaka and Thomas E Nichols. Combining voxel intensity and cluster extent with permutation test framework. *Neuroimage*, 23(1):54–63, Sep 2004. 10

[17] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, D. L. Collins, A. Evans, G. Malandain, N. Ayache, G. E. Christensen, and H. J. Johnson. Retrospective evaluation of intersubject brain registration. *IEEE Trans Med Imaging*, 22(9):1120–1130, Sep 2003. 3

[18] J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, and J. Lancaster. A probabilistic atlas of the human brain: theory and rationale for its development. the international consortium for brain mapping (icbm). *Neuroimage*, 2(2):89–101, Jun 1995. 1

[19] Robert A. McLean, William L. Sanders, and Walter W. Stroup. A Unified Approach to Mixed Linear Models. *The American Statistician*, 45(1):54–64, 1991. 2, 4

[20] Benoit Da Mota, Virgile Fritsch, Gaël Varoquaux, Tobias Banaschewski, Gareth J Barker, Arun L W Bokde, Uli Bromberg, Patricia Conrod, Jürgen Gallinat, Hugh Garavan, Jean-Luc Martinot, Frauke Nees, Tomas Paus, Zdenka Pausova, Marcella Rietschel, Michael N Smolka, Andreas Ströhle, Vincent Frouin, Jean-Baptiste Poline, Bertrand Thirion, and I. M. A. G. E. N. consortium. Randomized parcellation based inference. *Neuroimage*, 89:203–215, Apr 2014. 11, 16

[21] Thomas E. Nichols and Andrew P. Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp*, 15(1):1–25, Jan 2002. 2

[22] Alfonso Nieto-Castañón and Evelina Fedorenko. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *Neuroimage*, 63(3):1646–1669, Nov 2012. 3

[23] Regina Nuzzo. Scientific method: statistical errors. *Nature*, 506(7487):150–152, Feb 2014. 2

[24] Philippe Pinel, Bertrand Thirion, Sébastien Meriaux, Antoinette Jobert, Julien Serres, Denis Le Bihan, Jean-Baptiste Poline, and Stanislas Dehaene. Fast reproducible identification and large-scale databasing of individual functional cognitive networks. *BMC Neurosci*, 8:91, 2007. 13, 14

[25] Russell A. Poldrack, Jeanette A. Mumford, and Thomas E. Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, Cambridge, New York, 2011. 9, 20

[26] J. B. Poline and B. M. Mazoyer. Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. *J Cereb Blood Flow Metab*, 13(3):425–437, May 1993. 3, 10

[27] J. B. Poline, K. J. Worsley, A. C. Evans, and K. J. Friston. Combining spatial extent and peak intensity to test for activations in functional imaging. *Neuroimage*, 5(2):83–96, Feb 1997. 10

[28] Denis Rivière, Jean-François Mangin, Dimitri Papadopoulos-Orfanos, Jean-Marc Martinez, Vincent Frouin, and Jean Régis. Automatic recognition of cortical sulci of the human brain using a congregation of neural networks. *Med Image Anal*, 6(2):77–92, Jun 2002. 4

[29] Alexis Roche, Sébastien Mériaux, Merlin Keller, and Bertrand Thirion. Mixed-effect statistics for group analysis in fmri: a nonparametric maximum likelihood approach. *Neuroimage*, 38(3):501–510, Nov 2007. 7, 13

[30] Gholamreza Salimi-Khorshidi, Stephen M Smith, and Thomas E Nichols. Adjusting the neuroimaging statistical inferences for nonstationarity. *Med Image Comput Comput Assist Interv*, 12(Pt 1):992–999, 2009. 12

[31] Stephen M Smith and Thomas E Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1):83–98, Jan 2009. 3, 10, 11

[32] Andres Fco. Solé, Shing-Chung Ngan, Guillermo Sapiro, Xiaoping Hu 0001, and Antonio López. Anisotropic 2d and 3d averaging of fmri signals. *IEEE Trans. Med. Imaging*, 20(2):86–93, 2001. 10

[33] Peter Stiers, Ronald Peeters, Lieven Lagae, Paul Van Hecke, and Stefan Sunaert. Mapping multiple visual areas in the human brain with a short fmri sequence. *Neuroimage*, 29(1):74–89, Jan 2006. 3

[34] Stephen C Strother, Jon Anderson, Lars Kai Hansen, Ulrik Kjems, Rafal Kustra, John Sidtis, Sally Frutiger, Suraj Muley, Stephen LaConte, and David Rottenberg. The quantitative evaluation of functional neuroimaging experiments: the npairs data analysis framework. *Neuroimage*, 15(4):747–771, Apr 2002. 16

[35] Bertrand Thirion, Philippe Pinel, Sébastien Mériaux, Alexis Roche, Stanislas Dehaene, and Jean-Baptiste Poline. Analysis of a large fmri cohort: Statistical and methodological issues for group analyses. *Neuroimage*, 35(1):105–120, Mar 2007. 7, 16

[36] Alan Tucholka, Virgile Fritsch, Jean-Baptiste Poline, and Bertrand Thirion. An empirical comparison of surface-based and volume-based group studies in neuroimaging. *Neuroimage*, 63(3):1443–1453, Nov 2012. 5, 7, 13

[37] P.H. Westfall and S.S. Young. *Resampling-based multiple testinging: Examples and methods for p-value adjustment*. Wiley-Interscience, 1993. 12

[38] Anderson M Winkler, Gerard R Ridgway, Matthew A Webster, Stephen M Smith, and Thomas E Nichols. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, May 2014. 10, 12

[39] Mark W Woolrich, Timothy E J Behrens, Christian F Beckmann, Mark Jenkinson, and Stephen M Smith. Multilevel linear modelling for fmri group analysis using bayesian inference. *Neuroimage*, 21(4):1732–1747, Apr 2004. 7, 8, 13

[40] K. J. Worsley. An improved theoretical p value for spms based on discrete local maxima. *Neuroimage*, 28(4):1056–1062, Dec 2005. 11

[41] K. J. Worsley, C. H. Liao, J. Aston, V. Petre, G. H. Duncan, F. Morales, and A. C. Evans. A general statistical analysis for fmri data. *Neuroimage*, 15(1):1–15, Jan 2002. 7

A Table of the most standard statistical tests in neuroimaging

Test Description	Order of Data	$\mathbf{Z}\boldsymbol{\beta}$	Hypothesis test
One-sample t-test, 5 observations	$\mathbf{B}_1$ $\mathbf{B}_2$ $\mathbf{B}_3$ $\mathbf{B}_4$ $\mathbf{B}_5$	$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} \beta_1 \end{pmatrix}$	H_0: Overall mean = 0 H_0: $\beta_1 = 0$ H_0: $c^T\boldsymbol{\beta} = 0$ $c = [1]$
Two-sample t-test, 5 observations $Y_1 \cdots Y_5$ in group 1 and 5 observations $Y_6 \cdots Y_{10}$ in group 2.	$\mathbf{B}_1$ $\mathbf{B}_2$ $\mathbf{B}_3$ $\mathbf{B}_4$ $\mathbf{B}_5$ $\mathbf{B}_6$ $\mathbf{B}_7$ $\mathbf{B}_8$ $\mathbf{B}_9$ $\mathbf{B}_{10}$	$\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$	H_0: equal mean in both groups H_0: $\beta_1 - \beta_2 = 0$ H_0: $c^T\boldsymbol{\beta} = 0$ $c = [1\ -1]$
Paired t-test, 5 paired measures of two responses $\mathbf{B}$ and $\mathbf{B}'$, corresponding e.g. to two experimental conditions observed in 5 subjects.	$\mathbf{B}_1$ $\mathbf{B}'_1$ $\mathbf{B}_2$ $\mathbf{B}'_2$ $\mathbf{B}_3$ $\mathbf{B}'_3$ $\mathbf{B}_4$ $\mathbf{B}'_4$ $\mathbf{B}_5$ $\mathbf{B}'_5$	$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_{\text{diff}} \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}$	H_0: $\mathbf{B}'$ is equal to $\mathbf{B}$ H_0: $\beta_{\text{diff}} = 0$ H_0: $c^T\boldsymbol{\beta} = 0$ $c = [1\,0\,0\,0\,0\,0]$
Two way ANOVA. Factor $\mathbf{B}$ has two levels and factor $\mathbf{B}'$ has 3 levels. There are 2 observations for each $\mathbf{B}/\mathbf{B}'$ combination.	$\mathbf{B}_1\mathbf{B}'_1(1)$ $\mathbf{B}_1\mathbf{B}'_1(2)$ $\mathbf{B}_1\mathbf{B}'_2(1)$ $\mathbf{B}_1\mathbf{B}'_2(2)$ $\mathbf{B}_1\mathbf{B}'_3(1)$ $\mathbf{B}_1\mathbf{B}'_3(2)$ $\mathbf{B}_2\mathbf{B}'_1(1)$ $\mathbf{B}_2\mathbf{B}'_1(2)$ $\mathbf{B}_2\mathbf{B}'_2(1)$ $\mathbf{B}_2\mathbf{B}'_2(2)$ $\mathbf{B}_2\mathbf{B}'_3(1)$ $\mathbf{B}_2\mathbf{B}'_3(2)$	$\begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_{\text{mean}} \\ \beta_{\text{Y1}} \\ \beta_{\text{Y'1}} \\ \beta_{\text{Y2}} \\ \beta_{\text{Y1 Y'1}} \\ \beta_{\text{Y1 Y'2}} \end{pmatrix}$	F-tests for all contrasts $c^T\boldsymbol{\beta} = 0$ H_0: Overall mean is 0 $c = [1\,0\,0\,0\,0\,0]$ H_0: Main $\mathbf{B}$ effect is 0 $c = [0\,1\,0\,0\,0\,0]$ H_0: Main $\mathbf{B}'$ effect is 0 $c = \begin{bmatrix} 0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0 \end{bmatrix}$ H_0 : $\mathbf{B}/\mathbf{B}'$ interaction is 0 $c = \begin{bmatrix} 0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,0\,1 \end{bmatrix}$

Table 1: Example of Statistical models used in neuroimaging. The first column describes the model, the second column describes how the data are ordered in the outcome vector, the third column shows the design matrix, and the last column illustrates the hypothesis tests and corresponding contrasts. Note, in the ANOVA example F -tests are used for all contrasts, whereas t-tests are used for the other examples. We use the notations from Eq. 13. This table is adapted from [25].