Very large fMRI study using the IMAGEN database: sensitivity - specificity and population effect modelling in relation to the underlying anatomy

Benjamin Thyreau, Yannick Schwartz, Bertrand Thirion, Vincent Frouin, Eva Loth, Sabine Vollstädt-Klein, Tomas Paus, Eric Artiges, Patricia J. Conrod, Gunter Schumann, Robert Whelan, Jean-Baptiste Poline

PII: S1053-8119(12)00275-3
DOI: doi: 10.1016/j.neuroimage.2012.02.083
Reference: YNIMG 9296

To appear in: *NeuroImage*

Accepted date: 26 February 2012

Please cite this article as: Thyreau, Benjamin, Schwartz, Yannick, Thirion, Bertrand, Frouin, Vincent, Loth, Eva, Vollstädt-Klein, Sabine, Paus, Tomas, Artiges, Eric, Conrod, Patricia J., Schumann, Gunter, Whelan, Robert, Poline, Jean-Baptiste, Very large fMRI study using the IMAGEN database: sensitivity - specificity and population effect modelling in relation to the underlying anatomy, *NeuroImage* (2012), doi: 10.1016/j.neuroimage.2012.02.083

Very large fMRI study using the IMAGEN database: sensitivity - specificity and population effect modelling in relation to the underlying anatomy

Benjamin Thyreau[a], Yannick Schwartz[a], Bertrand Thirion[b], Vincent Frouin[a], Eva Loth[c], Sabine Vollstädt-Klein[d], Tomas Paus[e,f,g], Eric Artiges[h], Patricia J. Conrod[c], Gunter Schumann[c], Robert Whelan[i], Jean-Baptiste Poline[a], The IMAGEN Consortium

[a] *Neurospin, Commissariat à l'Energie Atomique, Gif-sur-Yvette, France*
[b] *INRIA Parietal, Gif-sur-Yvette, France*
[c] *Institute of Psychiatry, King's College London, United Kingdom*
[d] *Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany*
[e] *Rotman Research Institute, University of Toronto, Toronto, Canada*
[f] *School of Psychology, University of Nottingham, Nottingham, United Kingdom*
[g] *Montreal Neurological Institute, McGill University, QC, Canada*
[h] *Institut National de la Santé et de la Recherche Médicale (INSERM), UMR1000, Orsay, France*
[i] *Trinity Centre for Bioengineering, Trinity College Dublin, Dublin, Ireland*

Abstract

In this paper we investigate the use of classical fMRI Random Effect (RFX) group statistics when analysing a very large cohort and the possible improvement brought from anatomical information. Using 1326 subjects from the IMAGEN study, we first give a global picture of the evolution of the group effect t-value from a simple face-watching contrast with increasing cohort size. We obtain a wide "activated" pattern, far from being limited to the reasonably expected brain areas, illustrating the difference between statistical significance and practical significance. This motivates us to inject tissue-probability information into the group estimation, we model the BOLD contrast using a matter-weighted mixture of Gaussians and compare it to the common, single-Gaussian model. In both cases, the models parameters are estimated per-voxel for one subgroup, and the likelihood of both models is computed on a second, separate subgroup to reflect models generalization capacity. Various group sizes are tested, and significance is asserted using a 10-fold cross-validation scheme. We conclude that adding matter information consistently improves the quantitative analysis of BOLD responses in some areas of the brain, particularly those where accurate inter-subject registration remains challenging.

Keywords: Brain Mapping: methods, Likelihood Functions, Linear Models, Magnetic Resonance Imaging, Sensitivity and Specificity

1. Introduction

Functional Magnetic Resonance Imaging (fMRI) is a reference modality to explore the structure of the human brain and its functional organization. It offers a spatial and temporal resolutions that allow to characterize many brain processes, and benefits from an established set of statistical methods, making it the modality of choice for number of studies in both cognitive neurosciences and translational neurosciences.

Most of the published fMRI studies involve 15 to 25 subjects. While this is sufficient to detect a large range of contrast effects in carefully controlled experiment or to establish group differences when the number of category is low, it quickly becomes insufficient to detect more subtle effects, such as phenomena involving some trait difficult to control or just plainly unknown. An example of this are studies linking neuroimaging with genetic data, where the need to find relevant correlations from a large pool of genomic variables benefits from having a larger cohort. These large cohorts are currently few, but are becoming more common, and the number of such studies will probably increase in the future.

This raises the question of how classical fMRI statistical methods behave when applied to group comprising several hundreds of subjects, or more. In fMRI group experiments, the classical method relies on a spatial normalization to gain voxel-to-voxel matching of the individual effect across all subjects. The group detection step uses the General Linear Model and associated statistics to compute voxel-wise estimation of the strength of some contrast of interest across the population, and eventually tries to reject the null hypothesis that the effect occurred by chance using significance testing on parametric distributions (Worsley and Friston, 1995). A common way to choose the threshold for significance is to set it at the *p-value* of 5% (using some correction for multiple comparisons). For most simple cases, it comes down to a single t statistic assessed on a student law distribution with a relatively low number of degrees of freedom. Considering the fact that the t-value increases with the square-root of the

group size N, we question whether it remains meaningful to look at a binary pattern obtained by simply thresholding the map at some standard *p-value* to answer a neurological or cognitive neuroscience question.

In this paper, we make use of an unique fMRI dataset comprising more than a thousand subjects. We apply a standard fMRI group analysis using the Statistical Parametric Mapping (SPM) method, and illustrate the behaviour of this approach on this unusually large cohort. We show that in doing this, a very wide area of the brain, including white matter areas, is shown as activated - or more accurately, a large number of voxels are found above the statistical threshold used for rejecting the null-hypothesis (corrected for multiple comparison). Although it is in fact not a surprise at all from a statistical point of view given the large number of observations, it is important to quantify this and convey this information practically, as it may have important consequences in actual experimental setups and analyses.

As some regions may show statistical significance while not biologically relevant, we were interested to see, in a second part of this paper, whether a more relevant and precise estimation of the BOLD signal may be obtained by considering the underlying anatomical matter. Even though spatial warping algorithms have significantly improved, especially those that include structural segmentation (Ashburner and Friston (2005)), the classification of every voxel as belonging to a certain kind of structure (e.g. Grey Matter or White Matter) is at best fuzzy. It is expected that, for some voxel location of the MNI space, the underlying anatomy will be different across subjects. Models that tolerate a certain variation of the location across subjects alleviate this issue but they are usually computationally expensive and not widely used yet (Keller et al. (2009); Thirion et al. (2006)). Therefore, we tested the effect of including individual subject's structural information in the model. Following the hypothesis that neuronal activity is the original cause of the BOLD signal, we expect that BOLD values will be higher in grey matter compared to white matter or cerebrospinal fluid. Hence, the use of the information provided by the tissue segmentation, as obtained from a T1-weighted image, may be useful to obtain a better estimation and detection of the BOLD signal.

Several approaches to improve fMRI signal estimation by including anatomical information have been proposed. One of the simplest is to include anatomical priors from an atlas to conduct analysis on specific region of interest. However, the structural reliability of this method is limited to the accuracy of the atlas being considered, and currently most anatomical atlases are mutually inconsistent, as described by Bohland et al. (2009). Furthermore, atlas-based methods rely on a correct initial registration to atlas space and are probabilistic in their target labelling, as they do not use subject-specific anatomy.

An increasingly popular method for analysis consists in projecting the fMRI data onto the cortical surface. In that framework, a mesh of the cortical surface is first computed, usually from a structural, segmented T1-weighted image (Fischl et al. (1999), Van Essen et al. (2001)) and the fMRI data, coregistered to the structural image, are projected onto this surface where the analysis occurs. In Andrade et al. (2001), the projection relies on the interpolation of the BOLD data at the mesh nodes, plus some signal from the normal direction. The group-wide functional statistics in surface space benefit from the associated strong structural input to the spatial-normalization process. Some drawbacks include the difficulty of getting a working surface from noisy T1 images even when tissue classification is mostly correct, the restriction of the method to cortical brain structures, ignoring sub-cortical nuclei, and the fMRI data distortion correction. Another major practical drawback, however, is the heavy computational cost associated with all the processings, which currently limits the popularity of the technique.

Other methods aim at increasing the detection efficiency by inserting *a-priori* information on the shape of the expected result in the estimation process. The most common one is simply to filter the BOLD data to an average expected size of activations, usually 8 to 12 mm in diameter. In Penny et al. (2005), a Bayesian scheme is used to include regression coefficients of neighboring voxels in the model. Alternatively, regularisation based on Markov Random Fields has been suggested (Held et al. (2002); Descombes et al. (1998)) as a way to increase signal from noise. A fast implementation proposed by Ou et al. (2010) further incorporates anatomical information into the MRF-based detection framework. In Van De Ville et al. (2007), some statistical testing associated with wavelet-based processing is proposed, and in a different manner, in order to overcome inter-subject variability, Thirion et al. (2006) proposed a data-driven approach to create functional parcels, constraining them to be spatially-connected. Some other models uses spatial regularization to improves the estimation of the data autocorrelation. In Woolrich et al. (2001), the estimated autocorrelation parameters are spatially smoothed in a nonlinear, edge-sensitive way, hopefully within matter type. And while Worsley et al. (2002a) focuses on smoothing the autocorrelation as a way to gain degrees of freedoms, he suggests that adding cortical-surface information would improve the estimation. However, all these fMRI regularisation schemes do not make explicit use of the structural information such as provided by an additional, higher resolution image.

When dealing with a large group of subjects, it is increasingly difficult to fully guarantee the common spatial alignment of structures via, e.g. MNI standard-space normalization. Even though spatial warping algorithms have significantly improved (Klein et al. (2009)), especially those that include structural segmentation (Ashburner and Friston (2005)), the classification of every voxel as belonging to a certain kind of structure (e.g. Grey Matter or White Matter) is at best fuzzy and the matching imperfect. It is expected that, for some voxel location of the MNI space, the underlying anatomy will be different across subjects.

More advanced and more accurate models of the BOLD signal accounting simultaneously for signal variances originating from tissue-, subjects-, group-, or even scanner-level effects - are likely to emerge. However, such methods are not yet widely available nor practically usable in neuroimaging studies.

We therefore tested the effect of the inclusion of individual subject's structural information into the group BOLD model, and tested whether this improved the statistical model. Following the hypothesis that neuronal activity is the original cause of the BOLD signal, we expect that BOLD values will be higher in grey matter compared to white matter or cerebrospinal fluid. Hence, better use of the information provided by the tissue segmentation, as obtained from the T1-weighted image, would reveal useful to obtain a better estimation and detection of the BOLD signal.

Objectives of the study

The objectives of the study are three-fold.

First, we wanted to characterize the impact of using very large N while using classical inference, and whether an increased sensitivity leads to a situation for which activated regions are not likely to be biologically plausible, demonstrating and quantifying the difference between statistical and practical significance.

Second, for large N, it may be that some regions are significantly activated while only a subgroup of the population is indeed showing some effect, leading to erroneous interpretation. This is even clearer when looking at regions that contain grey matter only for part of the population. One goal of the study is therefore to see to which extent some effects could be detected in brain areas most likely to be composed of white matter. Recent work have reported that activity could be detected in the white matter (Mazerolle et al. (2008)).

Third, we wanted to establish whether the model used to detect BOLD activity in a population would benefit from knowledge of the underlying subject-per-subject anatomy, and if the use of this knowledge could lead to more sensitive or specific analyses.

We investigated these issues with a simple contrast computed from a face-viewing task, from the multi-centre IMAGEN neuroimaging-genetic database.

2. Material and Methods

2.1. Dataset and preprocessings

The fMRI paradigm

A large number (> 1500) of adolescents (13 to 15 years old), part of the IMAGEN sample (Schumann et al. (2010)), underwent fMRI BOLD recording during a simple passive face-viewing task (Grosbras and Paus (2006)). The task comprises alternating blocks containing videoclips of faces or control stimuli. Each block lasts 18-seconds, for a total duration of about six minutes. Face blocks comprise short greyscale videos of male or female faces, (emotionally neutral or angry) whereas control blocks are made of expanding or contracting grey circles. The functional contrast of interest in the present study corresponds to the face viewing condition (blocks of neutral and angry faces combined) minus the control condition: this contrast exhibits strong response in the fusiform gyrus and the amygdala, and several other brain regions, such as those along the superior temporal sulcus and in the frontal cortex (see probabilistic maps of the face network in Tahmasebi et al. (2011)).

Acquisitions parameters

Data were acquired from each of 7 acquisition centers on 3 Tesla scanners of three different manufacturers (General Electric (2), Siemens (4), Philips (1)), transferred to the data processing center at Neurospin, and made available to the consortium in a central database. The Imagen study gathers data from 8 acquisition centers, but at the time of this study, data from one of them had not been fully available.

For each subject, the BOLD time series are recorded using Echo-Planar Imaging, at the spatial resolution of 3.4mm isotropic, and temporal resolution of 2.2 seconds. The total length is 160 volumes. A separate high resolution (1mm isotropic) structural T1-weighted image is subsequently acquired using a MPRAGE (Mugler and Brookeman (1990)) sequence. For a complete description of the IMAGEN project, see Schumann et al. (2010).

Preprocessings and contrast maps

Each subject's data are processed using the SPM8 software (`http://www.fil.ion.ucl.ac.uk/spm/`) using the following steps. To correct for movements, each volume is spatially realigned to the mean over time. A linear model is then defined and fitted per-voxel to the realigned time series, using SPM8's standard General Linear Model routine. The design matrix defines the timing of Faces blocks and Control blocks, and also includes estimated motion parameters (3 translations and 3 rotations). Standard autoregressive noise model (AR(1)) and low frequency filters are set unchanged from the SPM8 defaults. This allows to compute an effect-size map for the "Faces vs Control" contrast, and its associated t significance map under normal hypothesis.

Separately, the T1-weighted image is segmented and warped to the MNI standard space using the "New Segmentation" algorithm from SPM8. This algorithm both classifies the input T1 image voxels as belonging to one of six classes of tissues (Grey, White, CSF, Meningeal, Skull, Background), and estimates the best spatial warping of six MNI-space tissue prior maps to match their respective T1 image tissues, in an iterative Bayesian framework. (Ashburner and Friston (2005)). This spatial normalization is based on a high-dimensional non-linear warping field (59 x 70 x 59 x 3 parameters), providing a possible relatively fine matching between images.

This yields probabilistic classification maps in both native and MNI space – of which we keep only the classes corresponding to Grey-Matter, White-matter, and Cerebrospinal fluid - and a deformation field.

The individual average BOLD image is then rigidly registered to the structural T1 image, and the resulting transformation, combined with the T1 deformation field, is applied to the fMRI statistics images using trilinear interpolation to obtain MNI-space, accurately registered contrast images and then, group significance maps. To avoid spreading the signal across anatomical structures, no smoothing is applied. However, the interpolation necessary for warping would inevitably add a limited amount of smoothing.

Additional checks of correct alignment

Due to the very high number of images involved, quality-checking has been made partially automatic. The functional signal intensity values were averaged over a few Regions of Interest (namely, in-brain mask, out-brain mask, Grey-matter mask, White-matter mask) for every subject images, and further manual, visual reviews have been limited to images whose values departed most from the mean values across the whole group. These manual reviews were mostly intended, and were successful at detecting obvious issues, such as problems due to warping, or severe artefacts. Additionally, contrasts images were binarized (non-null values set to 1) and summed over the group to detect misregistrations. In both cases, subjects are simply excluded if any problem occurred.

Data of 1326 subjects (from the initial ∼1500) successfully passed the pre-processing and checking steps and were included in the present study.

To create the tissue binary masks, thresholded tissues probability maps were used, at different thresholds due to the non-uniformity of the probabilities distribution, which were respectively 0.94, 0.82 and 0.80 for GM, WM and CSF, based on visual feedback, to ensure that masks were conservatively limited to their respective area.

2.2. Methods: 1 - sensitivity analysis as a function of brain matter

Description of the experiments

Empirical measures of sensitivity as a function of grey or white matter proportion are estimated from group results. For a specific population of subjects, a group significance map of the effect is obtained by computing the student t-value at each voxel. Additionally, a group probability for each tissue is obtained by summing the binary tissue masks of every subjects of the group.

We applied the following experiment on several subgroups of various sizes, obtained by selecting randomly subjects from the initial population.

First, we computed the distribution of the tissues as a function of the student p-value. The absolute values of t are discretized from 0 to 50 (with a 20 bins resolution), and for each bin, the average tissue probability over the image is measured for each tissue category. We report this for groups with different number of subjects (30,200,500,1326), and the significance threshold value for rejecting the family wise null hypothesis under normality assumption at the 5% level corrected level is computed with the bonferroni procedure to obtain a comparable statistic across groups of different sizes.

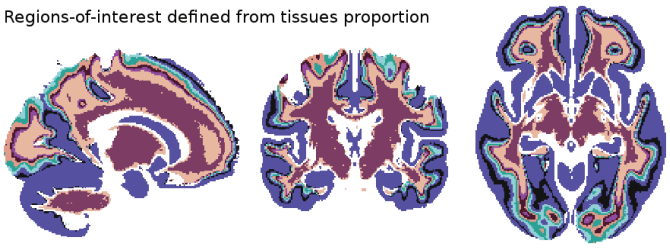

Figure 1: The 11 tissue-ratio Regions Of Interest, defined from isolines of the Grey/White probability ratio across the group. Color-coded levels use a random palette for better visual discrimination. The largest area ROI is defined from the upper bin of the grey-matter histogram

Second, the converse relationship is investigated. The average t-value is measured as a function of the Grey-White matter ratio. As we focus on grey and white matter, a special mask is computed which includes only voxels from where at least half of the subjects from the group has grey or white matter. For each voxel, the ratio of the two tissues across the group is computed and discretized (10 bins from 0 to 1). Figure 1 illustrates the regions of interest defined that way. The average t-values over those 10 iso-ratio ROIs is measured.

2.3. Methods: 2 - modeling the tissue difference

After studying the sensitivity of the standard analysis as a function of the amount of grey to white matter in a given voxel, we investigated whether the activation may come from a mixing of signals from different tissue type from different subjects. First, we checked whether at the same spatial locations, the variance coming from subjects having grey-matter tissue and the variance coming from subjects having white-matter tissue are equal. As the tissue classification algorithm is probabilistic, reflecting the uncertainty in labelling, we used a statistic adapted from the Levene test (Levene (1960)) for weighted data. The statistic formula is:

$$W = \frac{\{n_1.(\overline{r_1} - \overline{r})^2 + n_2.(\overline{r_2} - \overline{r})^2\}/(2-1)}{\{n_1.(\overline{r_1^2} - \overline{r_1}^2) + n_2.(\overline{r_2^2} - \overline{r_2}^2)\}/(n_1+n_2-2)},$$

where w_{ik} are the weights of tissue class k for subject i ($k \in \{1,2\}$, for resp. Grey or White matter) and $n_k = \sum_{i=1}^{N} w_{ik}$; $\overline{r}$ is the (weighted) average of the data deviation from their mean, the $\overline{r_k}$ are the (weighted) average of the deviation from the mean inside tissue class k and $\overline{r_k^2}$ are the (weighted) average of the squared deviation from the mean within tissue class. In essence, this ratio compares

the variances of the deviation from the mean within each of the two classes to the variance of the deviation from the mean between the two classes - as in the original Levene Test - except that we substituted standard moments by weighted moments in the computation (see Appendix 7 for details).

Then, we investigated whether a model that makes use of the knowledge of the underlying tissue classification, and potentially different variance components, would help signal detection. We compared the standard model (that does not consider tissue classes) to the one that explicitly model the provenance of the BOLD signal and asked under which model the data are most likely.

Therefore, at each voxel of subject i, the first model considers that the contrast effect y_i is normally distributed around the group mean value μ, with variance σ^2:

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

The second model considers that y_i is drawn from a weighted mixture of tissue-specific normal distributions:

$$y_i \sim \sum_{k=1}^{3} w_{ik}\, \mathcal{N}(\mu_k, \sigma_k^2) \quad k \in \{GM, WM, CSF\}, \qquad (1)$$

where each Gaussian component with parameters (μ_k, σ_k^2) models the signal originating from tissue k (resp. Grey matter, White matter, and Cerebrospinal Fluid), and the weights w_{ik} $(0 \leq w_{ik} \leq 1)$ reflect the voxel-specific proportion of those tissues for the subject i.

Estimation and testing

The first model parameters (μ, σ^2) are trivially estimated for each subgroups by computing the sample mean $\hat{\mu}$ and sample variance $\hat{\sigma}^2$ at each voxel of the brain in the MNI space. For the second model, the estimation is also straightforward as we need to estimate the Gaussians parameters $(\hat{\mu}_k, \hat{\sigma}_k)$ only given that the tissues weights w_{ik} of each observations are known from the segmentated image. We use the sample weighted mean and sample weighted variance as maximum likelihood estimators, that is, for a specific voxel the model parameters for tissue k are:

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} w_{ik} x_i}{\sum_{i=1}^{N} w_{ik}} \qquad (2)$$

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{N} w_{ik} \left(x_i - \hat{\mu}_k\right)^2}{\sum_{i=1}^{N} w_{ik}}, \qquad (3)$$

where x_i is the BOLD contrast effect of the voxel for subject i, w_{ik} is the weight of the tissue at that voxel and N is the group size[1].

To assess whether one model fits better than the other, a ratio of the averaged likelihood Λ was computed. To ensure that the results hold, we used a cross validation scheme, for which the model parameters were estimated on one subgroup while the likelihood statistics was computed on another.

$$\Lambda = \frac{\frac{1}{n}\sum_i^N \left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(x_i-\hat{\mu})^2}{2\hat{\sigma}^2}\right)\right)}{\frac{1}{n}\sum_{i=1}^N \sum_{k=1}^3 \frac{w_{ik}}{\sqrt{2\pi\hat{\sigma}_k^2}} \exp\left(-\frac{(x_i-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2}\right)}$$

At every voxel, we tested for the null hypothesis that the two models equally fit, or equivalently, that the log-likelihood difference is zero. Significance was asserted using 44 folds, by computing the p-value associated with the binomial law that the difference distribution would have under the null hypothesis.

2.4. Methods: 3 - Application of the model

We studied whether using tissue information leads to better estimation of the activations. In this experiment, we compare the group-wide Gaussian model restricted to the grey-matter to the group-wide model ignoring tissue information. We compared the sensitivity of the two models, both estimated with a simple normal distribution. We selected two activation thresholds by using the 0.25 and 0.75 quantile of the contrast image histogram. The suprathreshold probability associated with the two Gaussian models is measured, and their difference is computed. This is repeated over ten independent samples for three group sizes: 15, 100, and 500 subjects. The difference between the two models is tested against the null using a binomial distribution over the ten different folds.

3. Results

3.1. Results-1: sensitivity analysis

The empirical relationship between the t-value and the tissue probability is shown in Figure 2 for various group size (100, 200, 500, and our maximum possible, 1326). It is apparent that when group sizes get larger, the effect of thresholding with a usual parametric threshold (e.g. from a Student t distribution, choosing a fixed p-value) will delimit larger and larger clusters. For the 1326-sized group, t-values as extreme as 50 can be reached. Moreover for t values of 10 (corresponding to a p-value $< 10^{-15}$), the underlying anatomy probability is roughly equally distributed between Grey and White matter, thus violating the common prior about the localization of BOLD activations. Also, we noted that even with a drastic multiple-comparisons correction (p=5% bonferroni-corrected, i.e. assuming iid. voxel signal), which is often considered too severe for practical use, about 6% of the suprathreshold voxels were still located inside of a stringent white matter mask (area where 95% of registered subjects agreed on labelling the most probable tissue as white matter). These

[1] Note that in this mixture model, the mean and variance can be estimated directly as the weights are known

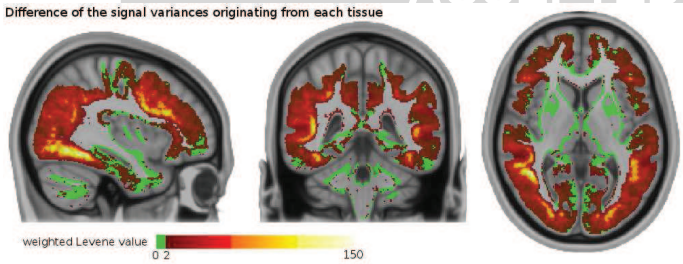

Figure 3: Inequality of BOLD variance components of the two tissues types. This map depicts the value of the adapted Levene test statistic. Stronger hot colors indicate areas where the variances of the Grey- and White-matter classes are the most different. Voxels where more than 90% of the subjects agree on the tissue class are not depicted as this statistic is mostly relevant in areas where both tissue proportion are non-null, i.e. voxels with strong inter-subject variability, partial volume effects or ambiguous tissue type.

exact values apply to the strongly activating contrast of interest (*faces-viewing vs visual control* contrast) and would certainly differ for more subtle contrasts.

The influence of the group size on a group t statistic is well known. As the group size increases, the t-value also increases following its square root. It is often assumed, however, that suprathresholds clusters of activations fully originate from a well-defined effect of interest consistent through all the subjects. The trend suggested by these plots is that by including more and more subjects into fMRI experiments, even effects of non-interest such as very faint background BOLD signal may be considered as relevant activations. In fact, in the simple linear model framework $y = x\beta + \epsilon$, ϵ normal iid, if an effect of magnitude m reaches the significance of $p = 0.05 \times 10^{-3}$ on a group of 30 subjects, (ie. if $\frac{m}{s/\sqrt{29}} = t^{-1}_{(29)}(10^{-5})$, where s is the sample standard deviation and $t^{-1}_{(\nu)}$ is the inverse Student probability function of ν degrees of freedom) then, when the group size is as high as 1300, an effect of magnitude about 7 times smaller can be detected at the same significance (since then $t^{-1}_{(1299)}(10^{-5}) = \frac{m/7.72}{s/\sqrt{1299}}$). This demonstrates that statistical significance is not necessary related to practical significance on an actual neuroimaging dataset.

3.2. Results-2: modeling with the tissues

The comparison of the variances coming from the two tissues classes, using the weighted Levene statistic, is illustrated in Figure 3, indicating that the variance at Grey Matter, White Matter and CSF cannot be assumed identical. This favors the hypothesis that those voxels contains a mixture of signal originating from grey matter for some subjects, and white matter for other subjects. This is a strong incitment to further exploit the variance components via a mixture model.

The possible improvement brought by adding anatomical information to the modelling of the BOLD effect is demonstrated by the likelihood ratio of two models which either makes use or ignores this information. Figure 4 shows three examples of the log of the estimated likelihood

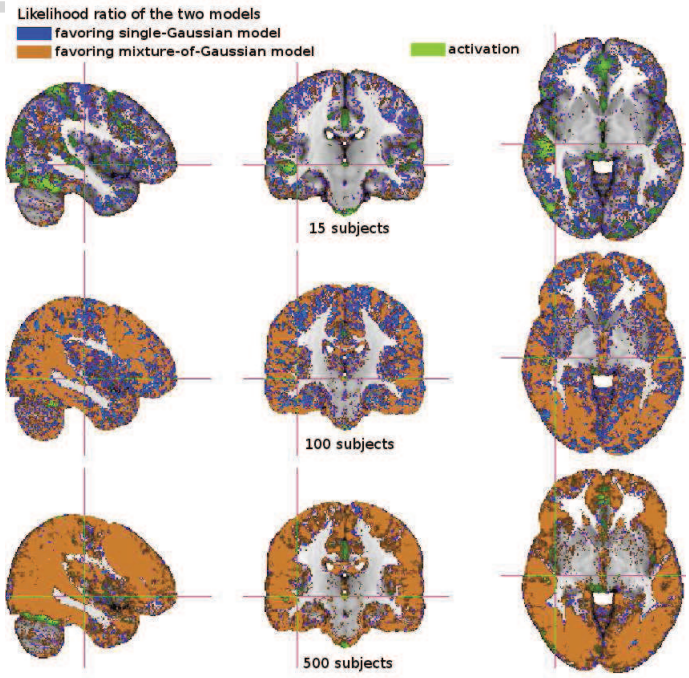

Figure 4: Example of the result of Likelihood ratios between the two models. For pairs of groups of three different size (15, 100 and 500, from top to bottom rows), the two models parameters are estimated on the first group and the likelihood computed for the other groups. The log-likelihood difference is depicted (Blue as favouring single-normal model, Orange as favouring matter-weighted normal model, not colored when the absolute log difference is less than 1.0). The estimated effect map for the contrast computed on 100 subjects is depicted as background overlay, thresholded at +/- 3 (Green). The improvement brought by the matter-weighted model becomes more obvious as the number of subjects increases.

ratio of both models. It should be noted that, although the tissue-weighted model has more parameters which would naturally better fit the training data set, the likelihood is evaluated on a separate testing data set, which removes non-specific bias and ensures generalization.

Although the model difference is typically small, it is significant and can be consistently reproduced. Figure 5 depicts areas where the likelihood ratio favors the anatomically-informed model significantly. These areas tend to correspond to voxels for which tissue classification varies across individual subjects, despite the registration.

3.3. Results-3: Using the selected model

The difference in sensitivity of the two models is assessed and depicted in Figure 6 to quantify the gain brought by the tissue-weighting model over the tissue-blind model. It is apparent that the GM-weighted model is more sensitive in areas with actual activations, and where tissue classification is ambiguous or suffers from partial-volume effects. This is the case both for positive and negative effect.

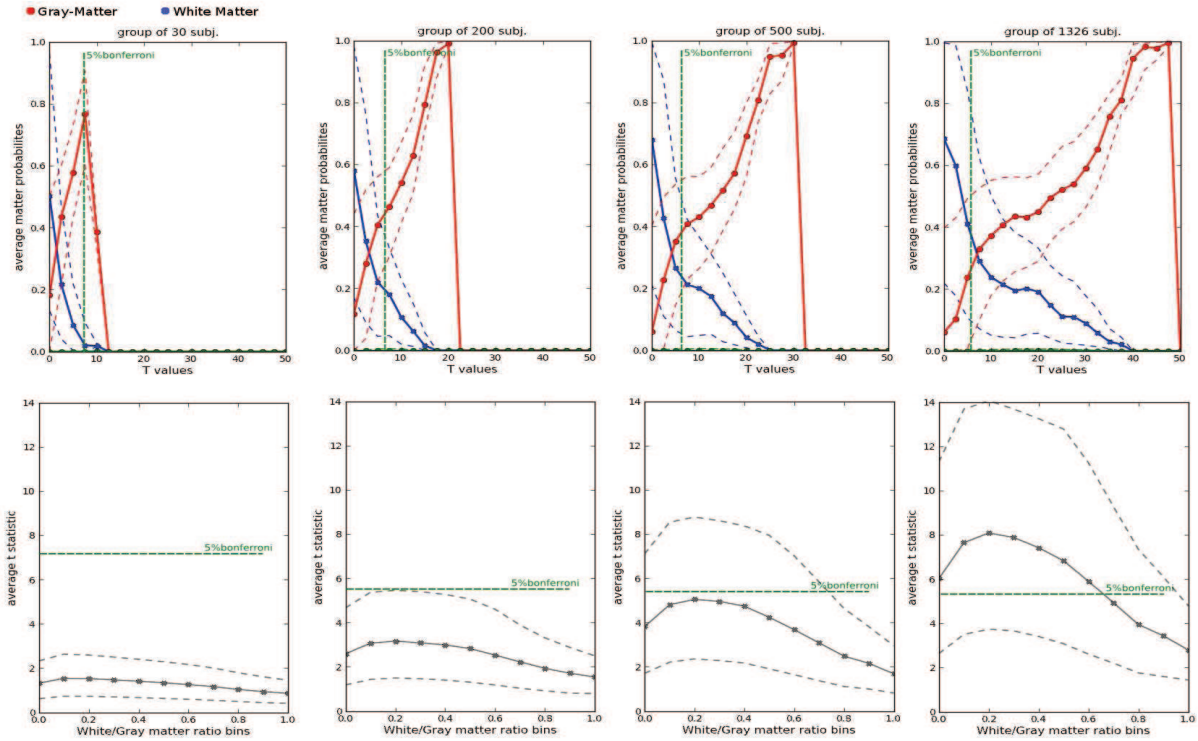

Figure 2: Relationship between the effect statistics and the anatomical structure, for different group sizes (100, 200, 500, 1326). Top: Tissue probability as a function of the t-statistic. Red is Grey-matter, Blue is White-matter. Plain color lines is the averages over the ROIs, dashed lines are their 25-75% quantiles. Bottom: Average effect t-value as a function of the White/Grey probability ratio.

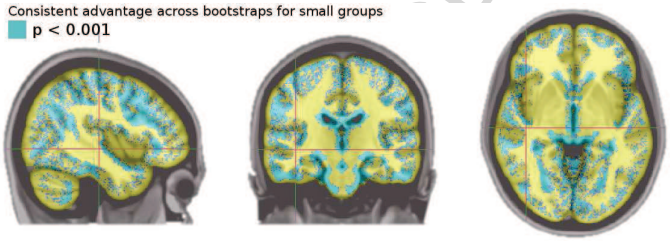

Figure 5: Significance from small groups. p-values of the binomial test assessing the advantage of the the anatomically informed model in terms of likelihood ratio. At each voxel, for 44 independent subgroups of 30 subjects, the two models are estimated on the first "training" half, and the log-ratio of their average likelihood on the "testing" second half is computed. The sign of that log-ratio value over the 44 subgroups is tested against a binomial distribution, and the corresponding p-value is depicted. Plain Cyan is $<< 0.001$, whereas Yellow is > 0.05, not significant.

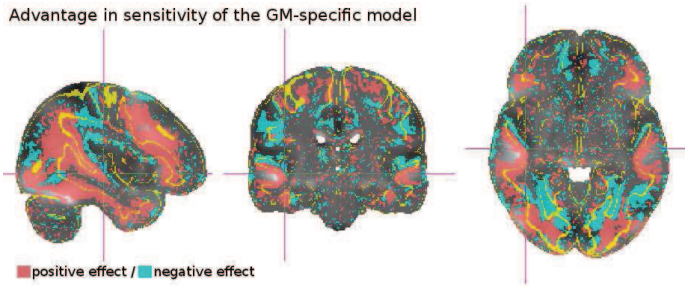

Figure 6: Group-wide average effect (greyscale background) and, as overlay, the significant difference of suprathreshold probability for the GM-weighted normal model minus the tissue-independent normal model. (Red when the effect is positive, Cyan when the effect is negative, estimated on N=500 subjects. Yellow iso-contour shows tissue probabilities of 50% grey, 50% white).

4. Discussion

In the above experiments, we described the behaviour of the thresholded t-test for significance assessment, as classically done in fMRI, on a database of unprecedented size. It can be emphasized that, as more and more subjects are included, the statistical test outcome becomes less and less anatomically relevant, as it tends to cross the threshold for significance however low the effect and whatever the underlying tissue might be. As strong activation in white matter is usually problematic from the physiological standpoint, we wanted to investigate whether it could partly be explainable by a subset of true signal combined with the undesired mixture of tissues. This led us to test whether tissue-dependent distributions could explain the voxel signal better than the classical Gaussian model. We then showed that the sensitivity of such a model is improved, even for small group sizes. A possible limitation of the current study is the lack of any modelling of a scanner effect. Whether the signal variance could be explained better with an additional scanner-specific variable remains to be verified, but appears most likely. This would strengthen our current results.

In the linear model framework, the standard error decreases as the square root of the number of subjects, and associated significance tests will therefore reject the null-hypothesis more and more easily. This is expected and well-known from a statistical point of view, but it has not been observed on actual neuroimaging data. Here we show that, using analyses performed with a popular fMRI analysis package (SPM) that can be used with large cohorts, biological significance may differ from statistical significance.

In this paper, we chose to use the high-dimensional SPM8 warping procedure and associated segmentation. We may expect the results to have differed quantitatively should we have used a different segmentation and/or warping algorithm, such as SyN (Avants et al. (2008)), ART (Ardekani et al. (1995)), or DARTEL (Ashburner (2007)). It is well known that spatial normalization is an approximative process. However the simultaneous influence of spatial normalization and tissue classification on the BOLD signal detection is not yet thoroughly documented. By exploiting the probabilistic nature of the tissue labelling, we showed that we can recover additional signal. Spatial registration is a complex procedure, which usually involves constrained deformations. The regularisation is necessary to protect against meaningless warping, but may also preclude perfect tissue alignment.

This problem can even corrupt white-matter areas, usually considered of lesser importance in cognitive studies. The activations in regions corresponding to white matter may arise from some BOLD signal originating in fact from cortical tissue as a consequence of misclassification and poor spatial normalization. Indeed, in another experiment (not presented in this paper) conducted using the white-matter ROIs set from the Johns Hopkins University atlas (Mori et al., 2008) as regions further intersected with each subject-level white-matter mask, the BOLD signal averaged over the defined regions was significantly non-null across subjects in most areas, at a 0.05 risk of error corrected for the number (50) of ROIs. At the voxel level, up to 6% of the activated voxels were found inside a very stringent white matter mask.

In an alternative approach, one may have the intuitive idea to simply fit some tissue confounding regressors (ie. each tissue probability) in the voxel signal model (e.g. using method from Casanova et al. (2007)); but as the variance components are unequal, a multilinear model with a single, normally-distributed residual variance model will not optimally explain the data. In a measurement of the 1326-sized group map, the average variance over the grey and white matter mask was 0.32, while it was 0.37 and 0.22 when separately restricted to Grey and White matter masks, respectively, indicating that BOLD signal measured on grey and white matter regions could be considered as originating from different distributions. This is reflected in our weighted-Levene statistic map depicting the difference of the signal variances between each tissues (fig. 3).

In neuroimaging fMRI analyses, it is usual to apply some smoothing to the data, and this is recommended because it increases the signal to noise ratio, mitigate inter-subject spatial variability, and enables the use of the Gaussian Random fields Theory for multiple comparisons correction. However, in this study, we did not apply any extra spatial smoothing to avoid mixing white and grey matter signals. Therefore, in real setting, one has to take care not to consider the effect illustrated in this paper as a simple undesirable consequence of smoothing.

Another limitation of the study is the use of a simple Random-Effect statistic that does not consider intra subject residual variance. Although it is a deliberate choice in this study to use that most common setting, it is possible that using the full Mixed-Effects statistics (Worsley et al., 2002b; Roche et al., 2007) would improve the estimation accuracy, by downweighting uncertain data, possibly even in a tissue-specific way. However, grey matter voxels have generally larger variance than white matter voxels, so the effect we observed may in fact be increased. It remains to be seen if a better mixed-effects model can be obtained by including anatomical information.

Finally, although the main point of this study was to illustrate the behaviour of classical analyses on large datasets, the model proposed here could have important practical applications. First, it allows to compute the BOLD effect measured at a voxel or within a region of interest *accounting for variation of the underlying anatomical structure*, for each subject. This may be crucial for instance in imaging genetic studies that require precise individual endophenotypes. Second, it may be important to be able to compare the BOLD values of different brain regions. This requires removing the effect induced by the amount of grey or white matter in these regions. Our

model allows to compare areas near the tissue interface to areas in plain grey-matter cortex, for either the effect size or the t- or z-value. Jernigan et al. (2003) argue that differences in size effects between areas of interest should be required before interpreting an activation pattern in fMRI studies, and propose that this should be the standard way of data presentation.

We hope that the present paper emphasized the limitations of the simple Student t-test based maps when working on large fMRI cohorts. This work may also be used to obtain better BOLD estimates that account for anatomical information of individual subjects.

5. Acknowledgements and funding source.

Support was provided by the IMAGEN project, which receives research funding from the European Community's Sixth Framework Programme (LSHM-CT-2007-037286. The funding sources had no further role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

6. References

References

Andrade, A., Kherif, F., Mangin, J.-F., Worsley, K. J., Paradis, A.-L., Simon, O., Dehaene, S., Le Bihan, D., Poline, J.-B., Feb. 2001. Detection of fMRI activation using Cortical Surface Mapping. Human Brain Mapping 12 (2), 79–93.
URL http://doi.wiley.com/10.1002/1097-0193(200102)12:2<79::AID-HBM1005>3.0.CO;2-1

Ardekani, B. A., Braun, M., Hutton, B. F., Kanno, I., Iida, H., 1995. A fully automatic multimodality image registration algorithm. Journal of computer assisted tomography 19 (4), 615–23.
URL http://www.ncbi.nlm.nih.gov/pubmed/7622696

Ashburner, J., Oct. 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38 (1), 95–113.
URL http://www.ncbi.nlm.nih.gov/pubmed/17761438

Ashburner, J., Friston, K. J., Jul. 2005. Unified segmentation. NeuroImage 26 (3), 839–51.
URL http://dx.doi.org/10.1016/j.neuroimage.2005.02.018

Avants, B. B., Epstein, C. L., Grossman, M., Gee, J. C., Feb. 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis 12 (1), 26–41.
URL http://dx.doi.org/10.1016/j.media.2007.06.004

Bohland, J. W., Bokil, H., Allen, C. B., Mitra, P. P., Jan. 2009. The brain atlas concordance problem: quantitative comparison of anatomical parcellations. PloS one 4 (9), e7200.
URL http://dx.plos.org/10.1371/journal.pone.0007200

Casanova, R., Srikanth, R., Baer, A., Laurienti, P. J., Burdette, J. H., Hayasaka, S., Flowers, L., Wood, F., Maldjian, J. A., Jan. 2007. Biological parametric mapping: A statistical toolbox for multimodality brain image analysis. NeuroImage 34 (1), 137–43.
URL http://dx.doi.org/10.1016/j.neuroimage.2006.09.011

Descombes, X., Kruggel, F., von Cramon, D. Y., Nov. 1998. fMRI signal restoration using a spatio-temporal Markov Random Field preserving transitions. NeuroImage 8 (4), 340–9.
URL http://www.ncbi.nlm.nih.gov/pubmed/9811552

Fischl, B., Sereno, M. I., Dale, A. M., Feb. 1999. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. NeuroImage 9 (2), 195–207.
URL http://dx.doi.org/10.1006/nimg.1998.0396

Grosbras, M., Paus, T., 2006. Brain networks involved in viewing angry hands or faces. Cerebral Cortex 16 (8), 1087.
URL http://cercor.oxfordjournals.org/content/16/8/1087.short

Held, K., Kops, E., Krause, B., Wells III, W., Kikinis, R., Muller-Gartner, H., 2002. Markov random field segmentation of brain MR images. Medical Imaging, IEEE Transactions on 16 (6), 878–886.
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=650883

Jernigan, T. L., Gamst, A. C., Fennema-Notestine, C., Ostergaard, A. L., 2003. More "mapping" in brain mapping: statistical comparison of effects. Human Brain Mapping 19 (2), 90–95.
URL http://www.ncbi.nlm.nih.gov/pubmed/12768533

Keller, M., Lavielle, M., Perrot, M., Roche, A., 2009. Anatomically Informed Bayesian Model Selection for fMRI Group Data Analysis. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009, 450–457.
URL http://www.springerlink.com/index/U04520X4V0T17524.pdf

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G. E., Collins, D. L., Gee, J., Hellier, P., Song, J. H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R. P., Mann, J. J., Parsey, R. V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage 46 (3), 786–802.
URL http://dx.doi.org/10.1016/j.neuroimage.2008.12.037

Levene, H., 1960. ROBUST TESTS FOR EQUALITY OF VARIANCES. Contributions to probability and statistics: Essays in honor of Harold Hotelling 2, 278.

Mazerolle, E. L., D'Arcy, R. C. N., Beyea, S. D., Jan. 2008. Detecting functional magnetic resonance imaging activation in white matter: interhemispheric transfer across the corpus callosum. BMC neuroscience 9 (1), 84.
URL http://www.biomedcentral.com/1471-2202/9/84

Mori, S., Oishi, K., Jiang, H., Jiang, L., Li, X., Akhter, K., Hua, K., Faria, A. V., Mahmood, A., Woods, R., Toga, A. W., Pike, G. B., Neto, P. R., Evans, A., Zhang, J., Huang, H., Miller, M. I., van Zijl, P., Mazziotta, J., Apr. 2008. Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. NeuroImage 40 (2), 570–82.
URL http://dx.doi.org/10.1016/j.neuroimage.2007.12.035

Mugler, J. P., Brookeman, J. R., 1990. Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). Magnetic Resonance in Medicine 15 (1), 152–157.
URL http://dx.doi.org/10.1002/mrm.1910150117

Ou, W., Wells, W. M., Golland, P., Jun. 2010. Combining spatial priors and anatomical information for fMRI detection. Medical image analysis 14 (3), 318–31.
URL http://www.ncbi.nlm.nih.gov/pubmed/20362488

Penny, W. D., Trujillo-Barreto, N. J., Friston, K. J., Jan. 2005. Bayesian fMRI time series analysis with spatial priors. NeuroImage 24 (2), 350–62.
URL http://www.ncbi.nlm.nih.gov/pubmed/15627578

Roche, A., Mériaux, S., Keller, M., Thirion, B., 2007. Mixed-effect statistics for group analysis in fMRI: a nonparametric maximum likelihood approach. NeuroImage 38 (3), 501–510.
URL http://www.ncbi.nlm.nih.gov/pubmed/17890108

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Buchel, C., Conrod, P. J., Dalley, J. W., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Itterman, B., Lathrop, M., Mallik, C., Mann, K., Martinot, J.-L., Paus, T., Poline, J.-B., Robbins, T. W., Rietschel, M., Reed, L., Smolka, M., Spanagel, R., Speiser, C., Stephens, D. N., Strohle, A., Struve, M., Dec. 2010. The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Mol Psychiatry 15 (12), 1128–1139.
URL http://dx.doi.org/10.1038/mp.2010.4http://www.nature.com/mp/journa

Tahmasebi, A. M., Banaschewski, T., Barker, G., Büchel, C., Conrod, P. J., Flor, H., Garavan, H., Heinz, A., Ittermann, B., Lathrop, M., Loth, E., Martinot, J.-L., Poline, J.-B., Robbins, T. W., Rietschel, M., Smolka, M., Spanagel, R., Stephens, D. N., Ströhle, A., Schumann, G., Paus, T., Consortium, T. I., 2011. Probing the Face Network in a Multi-centre Study of the Adolescent Brain:

Sources of Variability and Probabilistic Maps. Human Brain Mapping in Press.

Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.-B., Aug. 2006. Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets. Human brain mapping 27 (8), 678–93.
URL http://www.ncbi.nlm.nih.gov/pubmed/16281292

Van De Ville, D., Seghier, M. L., Lazeyras, F., Blu, T., Unser, M., Oct. 2007. WSPM: wavelet-based statistical parametric mapping. NeuroImage 37 (4), 1205–17.
URL http://www.ncbi.nlm.nih.gov/pubmed/17689101

Van Essen, D. C., Drury, H. A., Dickson, J., Harwell, J., Hanlon, D., Anderson, C. H., Jan. 2001. An integrated software suite for surface-based analyses of cerebral cortex. Journal of the American Medical Informatics Association : JAMIA 8 (5), 443–59.
URL http://jamia.bmj.com/content/8/5/443.full

Woolrich, M. W., Ripley, B. D., Brady, M., Smith, S. M., Dec. 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. NeuroImage 14 (6), 1370–86.
URL http://dx.doi.org/10.1006/nimg.2001.0931

Worsley, K. J., Friston, K. J., Sep. 1995. Analysis of fMRI time-series revisited–again. NeuroImage 2 (3), 173–81.
URL http://dx.doi.org/10.1006/nimg.1995.1023

Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., Evans, A. C., Jan. 2002a. A general statistical analysis for fMRI data. NeuroImage 15 (1), 1–15.
URL http://dx.doi.org/10.1006/nimg.2001.0933

Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., Evans, a. C., Jan. 2002b. A general statistical analysis for fMRI data. NeuroImage 15 (1), 1–15.
URL http://www.ncbi.nlm.nih.gov/pubmed/11771969

7. Appendix

We derived here a statistic for weighted data inspired by the Levene statistic (Levene, 1960). The original Levene statistic, for two groups (1 and 2) is:

$$W = \frac{\{n_1(\overline{r_1} - \overline{r})^2 + n_2(\overline{r_2} - \overline{r})^2\}/(2-1)}{\{\sum_{j=1}^{n_1}(r_{j1} - \overline{r_1})^2 + \sum_{j=1}^{n_2}(r_{j2} - \overline{r_2})^2\}/(N-2)},$$

where

- n_k is the number of subjects in class (group) k,
- $r_{jk} = |y_{jk} - \bar{y_k}|$, where y_{jk} is the value of the jth observation in class k and $\bar{y_k}$ is the mean of observations in class k,
- $\overline{r_k}$ is the mean of the r_{jk} in class k,
- $\overline{r}$ is the global mean of the r_{jk}

To rearange the denominator, we use the expectation equality

$$E[(x - E[x])^2] = E[x^2] - E[x]^2,$$

which, in its discrete form applied to r_k, reads:

$$\frac{1}{n}\sum_{j=1}^{n}(r_{jk} - \overline{r_k})^2 = \overline{r_k^2} - \overline{r_k}^2,$$

where $\overline{r_k^2}$ is the mean of ${r_{jk}}^2$ over j. This equality still holds when the expectation is changed to a weighted mean. We can therefore rewrite the statistic as:

$$W = \frac{\{n_1.(\overline{r_1} - \overline{r})^2 + n_2.(\overline{r_2} - \overline{r})^2\}/(2-1)}{\{n_1.(\overline{r_1^2} - \overline{r_1}^2) + n_2.(\overline{r_2^2} - \overline{r_2}^2)\}/(n_1+n_2-2)},$$

We modified the above to account for weighted information, by replacing means with weighted-means; ie. we now redefine the above terms as:

$$\begin{aligned}
n_k &= \sum_{j=1}^{N} w_{jk}, \\
r_{jk} &= |y_j - \overline{y_k}|, \text{ where } \overline{y_k} = \frac{1}{n_k}\sum_{j=1}^{N} w_{jk} y_j, \\
\overline{r_k} &= \frac{1}{n_k}\sum_{j=1}^{N} w_{jk} r_{jk}, \\
\overline{r_k^2} &= \frac{1}{n_k}\sum_{j=1}^{N} w_{jk} r_{jk}^2, \\
\overline{r} &= \frac{1}{n_1+n_2}\sum_{j=1}^{N}(w_{j1} r_{j1} + w_{j2} r_{j2})
\end{aligned}$$

with y_j the observation for subject j, $k \in \{1,2\}$ the two classes, and w_{jk} the weight from class k for subject j.

In this manuscript, y_j is the BOLD effect of subject j at some voxel of interest, whereas w_{jk} is the estimated proportion of tissue k ($k \in \{1,2\}$ corresponding to Gray Matter and White matter) at that voxel for that subject.

Note that our notion of group is different in this later statistic than in the original Levene statistic, and is not expected to follow a known distribution under the null hypothesis. However, for our experiment, we tabulated this statistic using simulations under several weight distribution profiles (and the number of observations equals that of our experimental data) in order to confirm that voxels highlighted in hot colors in figure 3 were significant at a 5% level ($W \approx 2$ for $p = 0.05$, while $W \approx 4$ for a $p = 10^{-4}$ level).

Highlights

- The behaviour of classical fMRI statistical analysis for very large databases
- Thresholded activation maps get less relevant as the number of subjects grows
- Including anatomical information, through simple tissue types, improves accuracy.
- BOLD signal is better estimated using 3 tissue-dependent variance components