Statistica Sinica **18**(2008), 1357-1374

DEALING WITH SPATIAL NORMALIZATION ERRORS IN fMRI GROUP INFERENCE USING HIERARCHICAL MODELING

Merlin Keller[1,2], Alexis Roche[2], Alan Tucholka[1,2]

and Bertrand Thirion[1,2]

[1] *CEA, Neurospin, Gif-sur-Yvette, France and* [2] *INRIA, Saclay, France*

Abstract: An important challenge in neuroimaging multi-subject studies is to take into account that different brains cannot be aligned perfectly. To this end, we extend the classical mass univariate model for group analysis to incorporate uncertainty on localization by introducing, for each subject, a spatial "jitter" variable to be marginalized out. We derive a Bayes factor to test for the mean population effect's sign in each voxel of a search volume, and discuss a Gibbs sampler to compute it. This Bayes factor, which generalizes the classical t-statistic, may be combined with a permutation test in order to control the frequentist false positive rate. Results on both simulated and experimental data suggest that this test may outperform conventional mass univariate tests in terms of detection power, while limiting the problem of overestimating the size of activity clusters.

Key words and phrases: Group analysis, hierarchical modeling, mixed effects, spatial uncertainty, Bayes factor, Metropolis within Gibbs, permutation test.

1. Introduction

In a typical cognitive study in functional magnetic resonance imaging (fMRI), several subjects are recruited from a population of interest and scanned while submitted to the same series of stimuli. A sequence of three-dimensional (3D) images of the brain is thus acquired for each subject, measuring over time a vascular effect of neural activity known as the blood oxygenation level dependent (BOLD) effect. From the time series recorded in each voxel, and the occurrence times for each stimulus, one may compute an estimate of the BOLD effect in response to any given stimulus, and more generally to any difference or combination of stimuli (contrast) (Friston (1997) and Worsley et al. (2002)).

Activation maps associated with a given contrast are obtained in this fashion for each subject, and used as input data for inference at the between-subject level, where the goal is to evidence a general brain activity pattern. A major issue of multi-subject cerebral studies lies in in the high morphological variability

of the human brain (Brett, Johnsrude and Owen (2002)). A traditional way to compensate for this is to register each individual anatomical image onto a common brain template (Ashburner and Friston (1999)), such as the widely used Montreal Neurological Institute (MNI) template. A comparative study of several normalization methods can be found in Hellier et al. (2003).

Any location in the brain can then be marked in a standard coordinate system, such as the one developed by Talairach and Tournoux (1988). However, registration is prone to errors (even assuming the existence of point-to-point correspondences between different brains), hence it does not seem reasonable to assume that homologous points are aligned across subjects. To date however, most methods for group analysis compare individual images on a voxelwise basis, thus making an implicit assumption that each subject is in perfect match with the template. Consequently, they tend to produce a stretching effect on group activity patterns due to the "jitter" induced by inaccurate registration. This effect can only be reinforced by preliminary linear spatial smoothing of the data, as is the traditional heuristic.

Alternatives to voxel-based methods have been developed recently, as in Thirion et al. (2007b), Xu, Johnson, and Nichols (2007) and Kim, Smyth, and Stern (2006). These techniques are feature-based in the sense that they extract key features from the individual images (critical points, activation regions centers), which are then matched across subjects. While providing a way to address imperfect registration of individual images, these methods rely crucially on the segmentation step, as well as on the ensuing feature-matching algorithm.

In this paper, we advocate a "low-level", as opposed to feature-based, approach that generalizes existing voxel-based methods while relaxing the assumption that the effects are well localized in the standard space. The method is developed in Section 2, where we start by extending the hierarchical model developed in Beckmann, Jenkinson and Smith (2003), Worsley et al. (2002) and Mériaux et al. (2006) by incorporating a simple model of spatial uncertainty. We then derive a Bayes factor to test for the mean population effect's sign in a given voxel, and justify a Metropolis-within-Gibbs sampling scheme to effectively compute this Bayes factor. Our approach is illustrated in Section 3 on both simulated and real data, and further discussed in Section 4.

2. Method

2.1. Classical two-level model

Considering a particular voxel $\boldsymbol{v} \in \mathbb{R}^3$ in the standard space, let X_i be the BOLD effect in $\boldsymbol{v}$ for subject i in response to a certain contrast of experimental conditions. A noisy estimate of X_i, denoted Y_i, is available from a within-subject

analysis on the fMRI time series that typically uses a general linear model (Friston (1997) and Worsley et al. (2002)). Under sufficient degrees of freedom, it is reasonable to consider Y_i as being normally distributed around X_i with known standard deviation s_i.

To address questions regarding the variability of the effect in a population, the unobserved effects $X_1, \ldots, X_n$ are further modeled as independent random variables drawn from an unknown distribution which characterizes the across-subject variability of BOLD responses. When this distribution is assumed Gaussian with unknown mean and variance (μ, σ^2), we obtain the same hierarchical model as in Beckmann, Jenkinson and Smith (2003), Worsley et al. (2002) and Mériaux et al. (2006).

- First level (within-subject):

$$Y_i | X_i \overset{ind.}{\sim} N(X_i, s_i^2), \tag{2.1}$$

- Second level (between-subject):

$$X_i | (\mu, \sigma^2) \overset{i.i.d.}{\sim} N(\mu, \sigma^2), \tag{2.2}$$

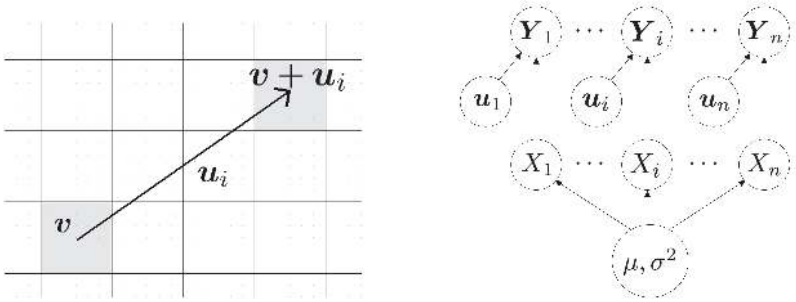
where independence sampling assumptions at both levels imply that the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are mutually independent conditionally on the population parameters (μ, σ^2). By integrating out the hidden variables X_i, we see that the observed effects are drawn independently but, in general, non-identically from the Gaussian distributions:

$$Y_i | (\mu, \sigma^2) \overset{ind.}{\sim} N(\mu, s_i^2 + \sigma^2). \tag{2.3}$$

That is to say, the observations are generally heteroscedastic unless all first-level deviations s_i are equal. In this special case, the model boils down to a simple sampling model, that is computationally attractive but lacks robustness against unreliable observations.

2.2. Incorporating spatial uncertainty

An important limitation of the two-level model, however, is that it describes the data separately in each voxel, thus making an implicit assumption that images from different subjects are comparable on a voxelwise basis. We now relax this assumption by incorporating localization uncertainty into the model. Given a voxel of interest $\boldsymbol{v} \in \mathbb{R}^3$ in the standard space, we consider that its homologous voxel in subject i is shifted according to an unknown displacement $\boldsymbol{u}_i \in \mathbb{R}^3$ which



(a) (b)

Figure 1. (a) Illustration of the spatial displacement $\boldsymbol{u}_i$, representing the unknown registration error for subject i in a certain voxel $\boldsymbol{v}$. (b) Graphical representation of the local hierarchical model.

reflects the registration error, as illustrated in Figure 1 (a). We thus generalize the within-subject model (2.1) as follows:

$$\boldsymbol{Y}_i(\boldsymbol{v}+\boldsymbol{u}_i)|(X_i,\boldsymbol{u}_i) \sim N(X_i, \boldsymbol{s}_i^2(\boldsymbol{v}+\boldsymbol{u}_i)), \tag{2.4}$$

where $\boldsymbol{u}_i$ is the discretization on the image grid of a 3D zero-mean Gaussian variable $N(0,\nu_i^2\boldsymbol{I}_3)$ assumed independent from the subject's response X_i. Discretization is simply achieved by rounding towards the nearest voxel coordinates. We restrict ourselves to the case of a scalar covariance matrix $\nu_i^2\boldsymbol{I}_3$, where ν_i therefore represents an isotropic standard registration error. In practice, we may use a subjective estimate for ν_i as registration procedures generally do not provide introspective performance measures.

Importantly here, we use bold letters for both $\boldsymbol{Y}_i$ and $\boldsymbol{s}_i^2$ to stress that they are spatial maps, as opposed to $Y_i = \boldsymbol{Y}_i(\boldsymbol{v})$ and $s_i^2 = \boldsymbol{s}_i^2(\boldsymbol{v})$, respectively the values of $\boldsymbol{Y}_i$ and $\boldsymbol{s}_i^2$ at location $\boldsymbol{v}$, which may not be in correspondence with $\boldsymbol{v}$ given spatial uncertainty. While (2.4) models the displaced effect $\boldsymbol{Y}_i(\boldsymbol{v}+\boldsymbol{u}_i)$, a generative model of the whole image $\boldsymbol{Y}_i$ is needed since our analysis now involves data from potentially any voxel location. We address this issue by specifying other voxels than the homologous of $\boldsymbol{v}$ as being drawn independently from uniform distributions, yielding

$$p(\boldsymbol{Y}_i|X_i,\boldsymbol{u}_i) = N\Big(\boldsymbol{Y}_i(\boldsymbol{v}+\boldsymbol{u}_i); X_i, s_i^2\Big) \times \prod_{\boldsymbol{v}'\neq\boldsymbol{v}+\boldsymbol{u}_i} \frac{1}{2r}\mathbf{1}_{|\boldsymbol{Y}_i(\boldsymbol{v}')|<r}, \tag{2.5}$$

where the radius r is an arbitrarily large constant. The hierarchical structure of our model can be summarized in a graph, as illustrated in Figure 1 (b).

This model is arguably not a realistic generative model of the data for at least two reasons. First, it ignores spatial correlations in the images, which

are potentially informative about the displacements $\boldsymbol{u}_i$. Second, this is only a *local* model in the sense that it is defined conditionally on a particular voxel of interest $\boldsymbol{v}$, and we cannot exhibit an unconditional density that is compatible with (2.5) for all $\boldsymbol{v}$. It is important, however, to realize that our aim here is not to perform a multivariate model-based analysis, but rather to guide the selection of a decision statistic for hypothesis testing (the test itself is to be calibrated under a less restrictive set of assumptions, as discussed in Section 2.6). With this in mind, the simplicity of the model may be regarded as a key advantage from a computational perspective.

2.3. Bayes factor as a decision statistic

Based on our "spatially noisy" two-level model, we now design a test of the presence of a positive mean population effect in a given voxel, that is, we test the null hypothesis $H_0 : \mu \leq 0$ against the alternative $H_1 : \mu > 0$. To that end, we may use the following Bayes factor as a decision statistic:

$$K = \frac{p(\boldsymbol{Y}|H_0)}{p(\boldsymbol{Y}|H_1)} = \frac{\int_{\mathbb{R}_-}\int_{\mathbb{R}_+^*} \pi(\mu,\sigma^2)L(\mu,\sigma^2)\,d\sigma^2 d\mu}{\int_{\mathbb{R}_+^*}\int_{\mathbb{R}_+^*} \pi(\mu,\sigma^2)L(\mu,\sigma^2)\,d\sigma^2 d\mu}, \tag{2.6}$$

where $L(\mu,\sigma^2) = \prod_{i=1}^n p(\boldsymbol{Y}_i|\mu,\sigma^2)$ is the likelihood function associated with the model given by (2.5) and (2.2). K compares the respective integrated likelihoods of both H_0 and H_1, and definition relies on a prior distribution $\pi(\mu,\sigma^2)$, an issue that we postpone to Section 2.4. The smaller K, the higher the evidence against H_0, hence the critical region of the test will be of the form $K \leq k$.

A frequentist alternative to K is the maximum likelihood ratio,

$$R = \frac{\sup_{\mu\leq 0,\sigma^2\in\mathbb{R}_+^*} L(\mu,\sigma^2)}{\sup_{\mu>0,\sigma^2\in\mathbb{R}_+^*} L(\mu,\sigma^2)},$$

as proposed in Mériaux et al. (2006) for the two-level model discussed in Section 2.1, that corresponds to the special case of our model in which the localization errors vanish ($\nu_i \equiv 0$). R may be seen as a prior-independent variant of K using likelihood maximizations instead of integrations.

The computation of either K or R raises algorithmic challenges as no closed form exists. We discuss in Section 2.5 a Markov Chain Monte-Carlo (MCMC) technique to compute K numerically. In its simple version under no spatial uncertainty ($\nu_i \equiv 0$), this algorithm may be seen as a stochastic version of the expectation-maximization (EM) algorithm described in Mériaux et al. (2006) and Roche et al. (2007) for R. While it might be feasible to extend the EM algorithm

to the more general case of spatial uncertainty, one advantage of the stochastic approach is that it is relatively robust to local convergence problems inherent to iterative deterministic procedures.

2.4. Prior specification

We now address the choice of a prior distribution for the population parameters (μ, σ^2), which may be seen as a third level in our hierarchical model. Under limited information about the possible values of (μ, σ^2), it is natural to use a weak prior. The Jeffreys prior turns out to be intractable in our case, thus we consider the scale invariant (improper) prior, $\pi(\mu, \sigma^2) \propto \sigma^{-2}$, which is also the Jeffreys prior associated with the second level of the model (2.2) or, equivalently, the special case of exact observations ($\nu_i \equiv 0$, $s_i \equiv 0$). However, the scale invariant prior is known to lead to an improper posterior in linear mixed models (Natarajan and Kass (2000)).

We instead adopt a proper prior for (μ, σ^2) that is conjugate to the second level of the model, yielding a Normal-Inverse Gamma distribution, as shown in Bernardo and Smith (2000):

$$\mu|(\sigma^2, \lambda, m) \sim N\big(m, \sigma^2/\lambda\big)$$
$$\sigma^2|(\alpha, \beta) \sim IG(\alpha, \beta),$$

where $IG(\alpha, \beta)$ is the Inverse-Gamma distribution with parameters (α, β), and density function

$$IG(z; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} \exp\Big(\frac{-\beta}{z}\Big).$$

$m \in \mathbb{R}$, λ, α, and $\beta > 0$ are hyperparameters. In order to tune α, and β, we followed the guidelines in Spiegelhalter et al. (1996), which suggests the use of a "just" proper prior for σ^2 in absence of prior knowledge, defined by $\alpha = \beta = 10^{-3}$. This means that the precision parameter $\tau = 1/\sigma^2$ has a prior mean of 1, and a prior variance of 10^3. From practical experience, we consider that in real fMRI datasets, σ^2 ranges from 10^{-2} to 10^2, hence our prior is roughly flat on the range of realistic values for σ. The prior mean is set to $m = 0$ so as not to bias the Bayes factor towards positive or negative effects. Finally, the scale parameter is set to $\lambda = 10^{-3}$, which may be interpreted as the weight given to the prior mean m with respect to one observation.

2.5. Monte-Carlo estimate of the Bayes factor

A characterization of K at (2.6) is

$$K = \frac{1}{p(\mu > 0|\boldsymbol{Y})} - 1,$$

given that the prior assigns equal probabilities to $H_0 : \mu \leq 0$ and $H_1 : \mu > 0$. Therefore, K may be computed by sampling from the posterior distribution of μ and counting the frequency of positive values. This can be done by means of a Gibbs sampler, as detailed hereafter.

We start with introducing the auxiliary variable $\tilde{\boldsymbol{Y}}_1, \ldots, \tilde{\boldsymbol{Y}}_n$ such that, for any subject i,

$$p(\tilde{\boldsymbol{Y}}_i | X_i, \boldsymbol{u}_i) = \delta\Big(\tilde{\boldsymbol{Y}}_i(\boldsymbol{v} + \boldsymbol{u}_i) - X_i\Big) \times \prod_{\boldsymbol{v}' \neq \boldsymbol{v} + \boldsymbol{u}_i} \frac{1}{2r} \mathbf{1}_{|\tilde{\boldsymbol{Y}}_i(\boldsymbol{v}')| < r}.$$

By letting $r \to \infty$, we see that the map $\tilde{\boldsymbol{Y}}_i(\boldsymbol{v}') + \varepsilon_i$, where $\varepsilon_i \sim N(0, \boldsymbol{s}_i^2(\boldsymbol{v}'))$ is an independent noise, has the same distribution as $\boldsymbol{Y}_i$, therefore $\tilde{\boldsymbol{Y}}_i$ may be interpreted as the badly localized effects before they are corrupted with observation noise.

From there, we design a Gibbs sampler to generate a sequence of samples from the joint posterior density $p(\mu, \sigma^2, \boldsymbol{u}, \boldsymbol{X}, \tilde{\boldsymbol{Y}} | \boldsymbol{Y})$ by sampling successively one of the following three blocks conditionally on the others: the population parameters (μ, σ^2), the auxiliary "badly localized effects" $\tilde{\boldsymbol{Y}}$, and the displacements $\boldsymbol{u}$. We now briefly summarize each step (the details are given in Appendix A).

- *Population parameters.* This step is straightforward, exploiting the conjugacy of our prior. The conditional posterior distribution of (μ, σ^2) is again from the Normal-Inverse Gamma class, with hyperparameters α', β', m', and λ' given by simple functions of the effects $X_1, \ldots, X_n$.
- *Badly localized effects.* For each subject i, $\tilde{\boldsymbol{Y}}_i(\boldsymbol{v}')$ is Gaussian, with mean $\boldsymbol{Y}_i(\boldsymbol{v}')$ and variance $\boldsymbol{s}_i^2$ in each voxel $\boldsymbol{v}'$, except in $\boldsymbol{v} + \boldsymbol{u}_i$ where the mean and variance also depend on the population parameters (μ, σ^2). Notice that this step is straightforward in the special case $\boldsymbol{s}_i \equiv 0$, as we then have $\tilde{\boldsymbol{Y}} = \boldsymbol{Y}$ almost surely.
- *Spatial displacements.* There is no simple way of sampling directly from the conditional distribution of the spatial displacements $\boldsymbol{u}_i$. Instead, we use an independent Metropolis-Hastings step, wherein the discretized Gaussian prior $N(0, \nu_i^2 \boldsymbol{I}_3)$ is used as a proposal distribution to draw a new displacement $\boldsymbol{u}_i$ which is accepted with a certain probability. This means that our sampling scheme is actually a Metropolis-within-Gibbs algorithm (see for instance Tierney (1994)). In our experiments, the average acceptance rate was around 20%, reflecting a limited match between the proposal distribution and the actual posterior. This observation suggests that the data provides strong, possibly anisotropic information on the displacements, and that the proposal distribution could be improved.

We note that under no spatial uncertainty ($\nu_i \equiv 0$), our method yields an alternative to the procedure in Woolrich et al. (2004) for sampling the posterior distribution of parameters in a mixed-effect linear model. By further assuming exact observations ($s_i^2 \equiv 0$), our model becomes similar to that developed in Friston et al. (2002) up to the prior. In this case, $p(\mu|\boldsymbol{Y})$ enjoys an explicit expression, hence providing a useful ground truth for the Gibbs sampler.

2.6. Thresholding

Depending on the value of the Bayes factor K in a particular voxel, our final task is to decide whether the voxel is active (in the sense that $H_1 : \mu > 0$ holds) or inactive ($H_0 : \mu \leq 0$). From a Bayesian perspective, this is a straightforward problem given the very definition of the Bayes factor. For instance, according to Jeffreys' scale of interpretation (Jeffreys (1961)), all voxels for which $K \leq 1/3$ exhibit a *substantial* evidence in favor of H_1, which becomes *strong* for those statisfying $K \leq 1/10$.

However, this simple thresholding may not be completely satisfactory in our context. First, as discussed earlier in Section 2.2, our Bayes factor does not rely on a proper multivariate model of the data, but rather on different voxel-specific models, that make it difficult to interpret the results in rigorous Bayesian terms. Importantly too, the neuroimaging research community has had a long tradition of classical hypothesis testing, and researchers are used to reporting functional brain regions in terms of frequentist risk. At least for the purpose of comparison with existing detection methods, we may consider tuning the threshold k so as to control the type I risk $P(K \leq k|H_0)$ below a fixed level α. The idea of using a Bayes factor for frequentist hypothesis testing is not new; see for instance Good (1992) and Aerts, Claeskens and Hart (2004) and, in a neuroimaging context, Woolrich et al. (2004).

While the type I risk may be calibrated under the model underlying K (up to the above conceptual warnings), a strong advantage of the frequentist approach is that it is easy to work under much more general assumptions. It is justified in Nichols and Holmes (2002) under mild nonparametric assumptions regarding the multivariate distribution of the data (in particular, the effects are symmetrically distributed in the population), as well as natural conditions regarding the decision statistic, that the type I risk is conservatively approximated by the probability of $K \leq k$ computed by sign permutations, i.e., by resampling K across all possible sign flips $\boldsymbol{Y}_i \to -\boldsymbol{Y}_i$ of the effect maps (for a total of 2^n possible permutations). This argument was extended to the case of noisy observations in Mériaux et al. (2006) and can easily be further extended to the case of spatial uncertainty by assuming that, for each subject i, the registration errors are independent of the well-localized effects.

This justifies using a sign permutation test to calibrate the type I risk associated with K. In practice, in order to keep computation time within reasonable bounds, we can pool the values of K across a limited number of randomly selected voxels, and generate a few random permutations for each. As shown in Mériaux et al. (2006), this trick makes it possible to control the overall false positive rate (the type I risk averaged across voxels) when using a uniform threshold k.

3. Results

3.1. Implementation

The above method was implemented in Python language, as part of the *NiPy* project (Neuroimaging in Python). The source code is freely downloadable from a *bazaar* repository at `https://launchpad.net/nipy`.

This is a rather computer intensive method. Using a parallel implementation over ten processors, the computation time for a typical analysis in 3D is of the order of one day (see Section 3.3), which is huge compared to a standard analysis as implemented e.g., in the Statistical Parametric Mapping (SPM) software (Friston (1997)), but still very small compared to the time required to design an fMRI experiment and acquire a complete group dataset. We believe that code optimization may enable us to divide computation time by one order of magnitude.

3.2. Simulations

We now illustrate from a simplistic simulation that standard voxelwise techniques (not accounting for spatial uncertainty) may lead to overestimating the size of positive effected regions, an undesirable "stretching effect" that our technique has the potential to reduce.

Synthetic datasets of n subjects were generated as follows. We defined a volume of $20 \times 20 \times 20$ voxels, containing a single spherical activated region in its center, with uniform intensity value 5 (the background was set to 0) and a fixed diameter of 5 voxels. This idealized activation was then jittered according to a discretized random Gaussian vector with standard deviation 0.5 voxels along each axis, an optimistically low estimate of spatial uncertainty. Independent heteroscedastic Gaussian noise was then added to each voxel $\boldsymbol{v}$, the variance of which was taken equal to $1 + \boldsymbol{s}^2(\boldsymbol{v})$, with $\boldsymbol{s}^2(\boldsymbol{v}) \sim \chi^2(1)$. A total of n pairs $(\boldsymbol{Y}_i, \boldsymbol{s}_i^2)$ of effect and variance maps were generated in this fashion.

We then computed two Bayes factor maps as defined in Section 2, respectively without spatial uncertainty ($\nu_i \equiv 0$), and with spatial uncertainty (setting $\nu_i \equiv 0.5$ voxels). We performed this simulation for different values of n, and report the results for $n = 200$, as the discrepancies between the two techniques were less obvious for smaller n.

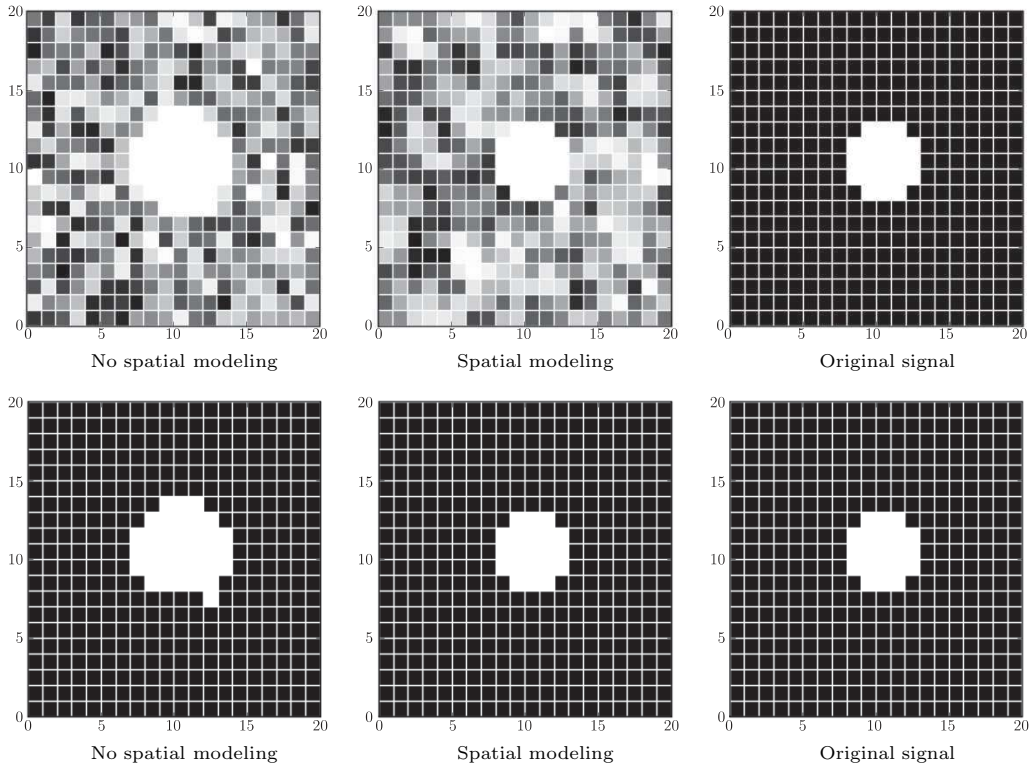


Figure 2. Top: Bayes factor maps on simulated data. Bright voxels correspond to small Bayes factor values (strong evidence for activation). Bottom: Oracle thresholding of each Bayes factor map.

Figure 2 displays the Bayes factor maps, showing as expected a noticeable stretching effect using the method without spatial uncertainty, which disappears when spatial uncertainty is accounted for. The method with spatial uncertainty achieves better detection at *all* thresholds, as shown on the receiving operator characteristic (ROC) curves in Figure 3. Those curves were plotted by counting for each possible threshold the number of detected voxels inside the original sphere (true positives) and outside (false positives).

For illustration, the bottom row in Figure 2displays the binary images obtained after thresholding the maps using the *Oracle* threshold (i.e., the threshold that minimizes bad classifications), confirming that the original sphere is better recovered under spatial uncertainty modeling.

As seen in Figure 3 from the posterior distributions of the mean population effect, both methods differ essentially near the sphere boundaries, where the method without spatial uncertainty finds a distribution whose support is far from the actual parameter value.

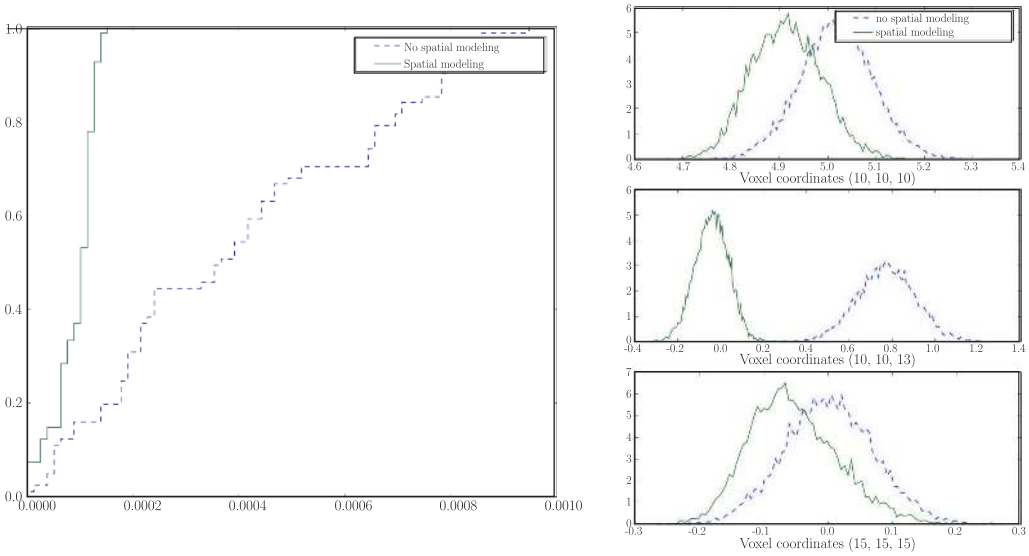


Figure 3. Left: ROC curves for detecting activations on simulated data. Right: Posterior distributions of the mean population effect in the sphere center (top), just outside (middle) and in the background (bottom). The solid line corresponds to the Bayes factor accounting for spatial uncertainty.

This simulation study shows that our method behaves according to intuition and therefore provides a consistency check. We remark that even a moderate amount of spatial uncertainty may yield a massive difference in the results of the two methods, provided that the number of subjects is large enough.

3.3. Data

We used an event-related fMRI protocol involving a relatively large cohort of 38 right-handed subjects. The participants were presented with a series of stimuli or were engaged in tasks such as passive viewing of horizontal or vertical checkerboards, left or right click after audio or video instruction, computation (subtraction) after video or audio instruction, sentence listening, and reading. Events occurred randomly in time (mean inter stimulus interval: 3s), with ten occurrences per event type, and ten event types in total.

The subjects gave informed consent and the protocol was approved by the local ethics committee. Functional images were acquired on a General Electric Signa 1.5T scanner using an Echo Planar Imaging sequence (time of repetition $= 2,400$ ms, time to echo $= 60$ ms, matrix size $= 64 \times 64$, field of view $= 24$ cm^2). Each volume consisted of 34 64×64 3 mm-thick axial contiguous slices. A session comprised 130 scans. Anatomical T1 weighted images were acquired on the same scanner, with a spatial resolution of $1 \times 1 \times 1.2$ mm^3. Finally, the

cognitive performance of the subjects was checked using a battery of syntactic and computational tasks.

First-level analyses were conducted using SPM5 (freely downloable from http://www.fil.ion.ucl.ac.uk). Data were submitted successively to motion correction, slice timing and normalization to the MNI template. For each subject, BOLD contrast images were obtained from a fixed-effect analysis on all sessions. Group analyses were restricted to the intersection of all subjects' whole-brain masks, comprising $43,367$ voxels.

As in the simulation (see Section 3.2), we computed two Bayes factor maps, respectively without and with spatial uncertainty modeling, setting this time $\nu_i \equiv 0.7$ voxels. This value was chosen by analogy with the size of the Gaussian kernel classically used to spatially smooth the data, and corresponds to a full width at half maximum (FWHM) of 5 mm, a standard value (the intuition underlying our choice is to interpret spatial smoothing as a crude heuristic to work around spatial uncertainty). 10^5 iterations of the Metropolis-within Gibbs sampler were used to approximate the Bayes factor in each voxel, following 10^4 preliminary iterations that were discarded, corresponding to the so-called "burn-in period".

We then used a randomized permutation test to threshold each statistical map. A total of $12,500$ permutations was generated by performing 500 random sign permutations of the individual effect maps and, for each permutation, computing the statistic in 25 randomly selected voxels. The computations were parallelized over ten processors. Computing the Bayes factor map took approximately 20 hours (3 hours without spatial uncertainty) and the permutation-based calibration required about 6 hours (respectively, 1 hour).

For comparison, we also included in our study the parametric t-test performed on the data after preliminary Gaussian smoothing (using $5 \times 5 \times 5$ mm^3 FWHM), as it is the reference method for fMRI group analysis. Note that, in this case, parametric thresholding yields very similar results to permutation-based thresholding, which comes as no surprise given the relatively large number of subjects and the asymptotic convergence of the permutation t test toward the Student distribution (Good (2005)).

We report results from the "calculation–sentences" contrast, which subtracts activations due to reading or hearing instructions from the overall activations detected during the mental calculation tasks. This contrast may thus reveal regions that are specifically involved in the processing of numbers. As seen in Figure 4, the three tested methods find qualitatively similar activation patterns, with large bilateral suprathreshold clusters in the parietal lobe, known to be involved in number processing.

When comparing the Bayes factor maps under $\nu_i \equiv 0$ and $\nu_i \equiv 0.7$ voxels (both maps being obtained without Gaussian smoothing), we do not observe a

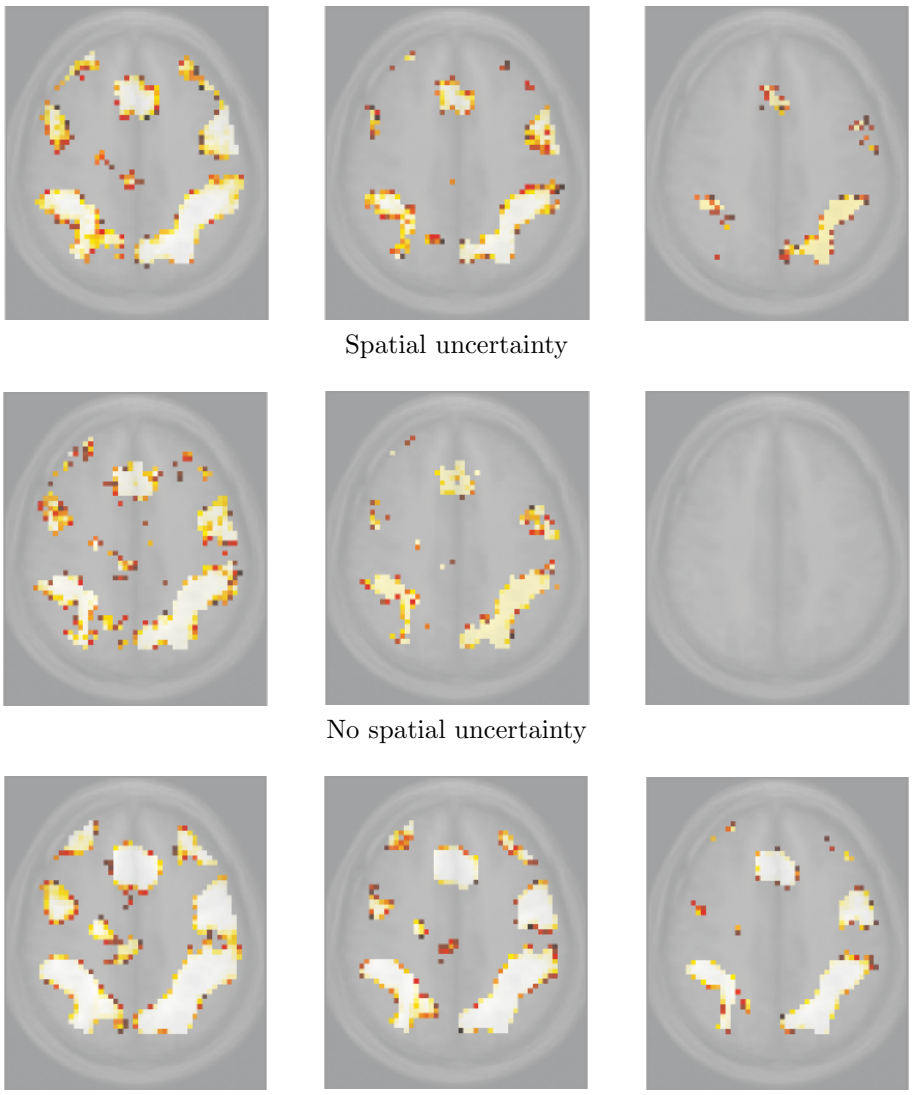


Figure 4. Bayes factor maps for the Calculation–Sentence contrast, thresholded for a false positive rate at 5% (left), 1% (center) and 0.1% (right) in an axial slice ($z = 45$ mm in Talairach space).

difference in the size of the detected clusters, contrary to what could be anticipated from the simulation results (see Section 3.2). However, the less noisy aspect of the map accounting for spatial uncertainty suggests higher detection power, along with the fact that the method with $\nu_i \equiv 0$ detected no cluster at 0.1% false positive rate. While still speculative, these findings are supported by a

large number of (moderately dependent) tests. Here, the lack of detection power might conceal the "stretching effect" observed in the simulations.

Not surprisingly, the standard t-test combined with preliminary smoothing yields wider suprathreshold clusters, as a predictable outcome of smoothing. On the other hand, we observe that most clusters survive longer to increasing thresholds in the standard analysis than in the first two analyses, suggesting a beneficial impact of smoothing on detection power. One obvious reason for smoothing the data is to increase the signal to noise ratio (under the terms of our model, reduce the first-level errors s_i) by exploiting the intrinsic spatial correlation of the signal. Such pre-processing may be useful considering the fact that the model underlying group inference ignores spatial correlations, mainly for computational reasons (see Section 2).

4. Discussion

In summary, we have introduced a new method for fMRI group data analysis that addresses the spatial variability of brain activation patterns. Contrary to previous feature-based approaches, our approach relies on a rather natural generalization of massively univariate voxel-based models, in which registration errors are treated as additional hidden variables.

In practice, registration is performed via T1-weighted images, hence the spatial mappings relating different brains are implicitly defined in an anatomical sense. This is to say that the method does not compensate for intrinsic functional variabilities in the extent or location of functional areas, but rather assesses variations in functional responses across homologous anatomical sites. Yet the fact that the analysis is carried out on a voxel-by-voxel basis (for computational reasons) implies that it can only provide limited geometrical characterization of functional areas. This should be kept in mind when interpreting group inference results. Typically, the shape of a multi-subject activity cluster (defined by local hypothesis testing) may not tell us much about the *average* shape of individual activation patterns.

Our preliminary experiments indicate that our method has a regularizing effect over conventional massively univariate inference (employed without preliminary image smoothing), and has the potential to reduce an artifactual "stretching effect" that arises in absence of spatial uncertainty modeling. The latter effect was demonstrated on a simulation but was not observed on the real dataset, probably due to the relatively limited number of subjects involved. We anticipate, however, that this effect would become prevalent in larger cohorts, which are increasingly used in neuroimaging (Thirion et al. (2007a)).

We do not dismiss traditional linear pre-smoothing, which may help boosting detection power, yet at the price of degraded spatial resolution. An ideal trade-off is still to be found. It might be useful, in practice, to use the proposed method after moderate image smoothing. We would however recommend nonlinear smoothing strategies, such as cortical surface-constrained filtering (Andrade et al. (2001)) or anisotropic diffusion (Kim et al. (2005)), in order to limit the blurring effect inherent to Gaussian and other linear filters.

Acknowledgement

We would like to thank Jean-Michel Marin (INRIA/Université Paris-Sud), Marc Lavielle (INRIA/Université Paris-Sud), and Philippe Ciuciu (CEA) for their valuable comments and suggestions, and Philippe Pinel (INSERM/CEA) for providing us with the data.

Appendix

A. Details of the Gibbs sampler

Our sampling algorithm is based on the model defined by Equations (2.5) and (2.2) after introduction of auxiliary variables, as described in Section 2.5. The joint posterior density of the hidden variables $(\boldsymbol{u}, \tilde{\boldsymbol{Y}})$ and the parameters (μ, σ^2) is proportional to

$$p(\boldsymbol{u}, \tilde{\boldsymbol{Y}}\mu, \sigma^2|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\tilde{\boldsymbol{Y}})p(\tilde{\boldsymbol{Y}}|\boldsymbol{u}, \mu, \sigma^2)p(\boldsymbol{u})\pi(\mu, \sigma^2).$$

The posterior conditional density of each block is deduced from this joint density by considering variables from all other blocks as fixed.

A.1. Population parameters

The posterior conditional density of the parameters (μ, σ^2) is given by

$$\begin{aligned} p(\mu, \sigma^2|\boldsymbol{u}, \tilde{\boldsymbol{Y}}, \boldsymbol{Y}) &\propto p(\tilde{\boldsymbol{Y}}|\boldsymbol{u}, \mu, \sigma^2)\pi(\mu, \sigma^2) \\ &\propto \prod_{i=1}^{n} \left\{ N(\tilde{\boldsymbol{Y}}_i\Big(\boldsymbol{v} + \boldsymbol{u}_i); \mu, \sigma^2\Big) \right\} \pi(\mu, \sigma^2), \end{aligned}$$

so that they depend on the other variables only through the hidden effects $X_i = \tilde{\boldsymbol{Y}}_i(\boldsymbol{v} + \boldsymbol{u}_i)$. Exploiting the conjugacy of our prior, the conditional posterior distribution of (μ, σ^2) is again from the Normal-Inverse Gamma class, as defined in Section 2.4, with hyperparameters m', λ', α', and β' given by

$$m' = \frac{n\overline{X} + \lambda m}{n + \lambda}; \quad \lambda' = n + \lambda; \quad \alpha' = \alpha + \frac{n}{2}; \quad \beta' = \beta + \frac{nS^2}{2} + \frac{n\lambda(m - \overline{X})^2}{2(n + \lambda)},$$

where $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$ and $S^2 = n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

A.2. Badly localized effects

Conditionally on $\boldsymbol{u}$ and (μ, σ^2), the badly localized effects $\tilde{\boldsymbol{Y}}_i$ are independent, with density

$$p(\tilde{\boldsymbol{Y}}_i|\boldsymbol{u}_i, \mu, \sigma^2, \boldsymbol{Y}_i) \propto p(\boldsymbol{Y}_i|\tilde{\boldsymbol{Y}}_i)p(\tilde{\boldsymbol{Y}}_i|\boldsymbol{u}_i, \mu, \sigma^2).$$

This expression can be further factored across voxels. Thus $\tilde{\boldsymbol{Y}}_i(\boldsymbol{v} + \boldsymbol{u}_i)$ is Gaussian, with mean m_i and variance γ_i^2 given by

$$m_i = \frac{\sigma^2 \times \boldsymbol{Y}_i(\boldsymbol{v} + \boldsymbol{u}_i) + \boldsymbol{s}_i^2(\boldsymbol{v} + \boldsymbol{u}_i) \times \mu}{\sigma^2 + \boldsymbol{s}_i^2(\boldsymbol{v} + \boldsymbol{u}_i)}; \qquad \gamma_i^2 = \frac{\sigma^2 \times \boldsymbol{s}_i^2(\boldsymbol{v} + \boldsymbol{u}_i)}{\sigma^2 + \boldsymbol{s}_i^2(\boldsymbol{v} + \boldsymbol{u}_i)}.$$

In any location $\boldsymbol{v}'$ other than $\boldsymbol{v} + \boldsymbol{u}_i$, $\tilde{\boldsymbol{Y}}_i(\boldsymbol{v}')$ is Gaussian with mean $\boldsymbol{Y}_i(\boldsymbol{v}')$ and variance $\boldsymbol{s}_i^2(\boldsymbol{v}')$. In the special case of no estimation errors ($s_i^2 \equiv 0$), $\tilde{\boldsymbol{Y}} = \boldsymbol{Y}$ almost surely, and this step can be dropped from the Gibbs sampler.

A.3. Spatial displacements

In this section, the $\boldsymbol{u}_i$ are discretized to the voxel grid only when evaluating $\tilde{\boldsymbol{Y}}_i(\boldsymbol{v} + \boldsymbol{u}_i)$, otherwise they are considered as continuous random variables. Conditionally on the parameters (μ, σ^2) and the badly localized effects $\tilde{\boldsymbol{Y}}$, the spatial displacements $\boldsymbol{u}_i$ are independent, with density

$$\begin{aligned} p(\boldsymbol{u}_i|\tilde{\boldsymbol{Y}}_i, \mu, \sigma^2) &\propto p(\tilde{\boldsymbol{Y}}_i|\boldsymbol{u}_i, \mu, \sigma^2)p(\boldsymbol{u}_i) \\ &\propto N(\tilde{\boldsymbol{Y}}_i(\boldsymbol{v} + \boldsymbol{u}_i); \mu, \sigma^2)N(\boldsymbol{u}_i; 0, \nu_i^2 \boldsymbol{I}_3). \end{aligned}$$

We use an independent Metropolis step to sample from this distribution, by drawing a proposal $\boldsymbol{u}_i$ from the 'prior' Gaussian $N(0, \nu_i^2 \boldsymbol{I}_3)$. If $\boldsymbol{u}_i^t$ is the current spatial displacement, the proposal is accepted ($\boldsymbol{u}_i^{t+1} = \boldsymbol{u}_i$) with probability $\min\{1, a\}$, where

$$a = \frac{p(\boldsymbol{u}_i|\tilde{\boldsymbol{Y}}_i, \mu, \sigma^2)N(\boldsymbol{u}_i^t; 0, \nu_i^2 \boldsymbol{I}_3)}{p(\boldsymbol{u}_i^t|\tilde{\boldsymbol{Y}}_i, \mu, \sigma^2)N(\boldsymbol{u}_i; 0, \nu_i^2 \boldsymbol{I}_3)} = \frac{N(\tilde{\boldsymbol{Y}}_i(\boldsymbol{v} + \boldsymbol{u}_i); \mu, \sigma^2)}{N(\tilde{\boldsymbol{Y}}_i(\boldsymbol{v} + \boldsymbol{u}_i^t); \mu, \sigma^2)}.$$

If the proposal is not accepted, then the current value is retained: $\boldsymbol{u}_i^{t+1} = \boldsymbol{u}_i^t$. In the special case of no spatial uncertainty ($\nu_i \equiv 0$), $\boldsymbol{u}_i$ is frozen to 0, $a \equiv 1$ and this step may be dropped out from the sampling scheme.

When observations are exact ($\nu_i \equiv 0, s_i^2 \equiv 0$), the posterior distribution of the population mean μ in a given voxel is tractable, and given by

$$\mu \overset{D}{=} \frac{n\overline{Y}}{n + \lambda} + \sqrt{\frac{S^2 + \lambda\overline{Y}^2 + 2\frac{\beta}{n}}{n + 2\alpha}}\, T,$$

where $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$, $S^2 = n^{-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$, and T is a Student variate with $n + 2\alpha$ degrees of freedom.

References

Aerts, M., Claeskens, G. and Hart, J. D. (2004). Bayesian-motivated tests of function fit and their asymptotic frequentist properties. *Ann. Statist.* **32**, 2580-2615.

Andrade, A., Kherif, F., Mangin, J.-F., Worsley, K., Paradis, A.-L., Simon, O., Dehaene, S. and Poline, J.-B. (2001). Detection of fMRI activation using cortical surface mapping. *Hum. Brain Mapp.* **12**, 79-93.

Ashburner, J. and Friston, K. (1999). Nonlinear spatial normalization using basis functions. *Hum. Brain Mapp.* **7**, 254-66.

Beckmann, C., Jenkinson, M. and Smith, S. (2003). General multi-level linear modelling for group analysis in fMRI. *Neuroimage* **20**, 1052-1063.

Bernardo, J. and Smith, A. (2000). Bayesian Theory. John Wiley & Son Ltd.

Brett, M., Johnsrude, I. and Owen, A. (2002). The problem of functional localization in the human brain. *Nature Reviews Neuroscience* **3**, 243-249.

Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G. and Ashburner, J. (2002). Classical and bayesian inference in neuroimaging: Theory. *Neuroimage* **16**, 465-483.

Friston, K. J. (1997). *Human Brain Function* (Chapter 2, pages 25-42). Academic Press.

Good, I. J. (1992). The Bayes/non-Bayes compromise: a brief review. *J. Amer. Statist. Assoc.* **87**, 597-606.

Good, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, 3rd Edition. Springer.

Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Le Goualher., G., Collins, D. L., Evans, A., Malandain, G., Ayache, N., Christensen, G. E. and Johnson, H. J. (2003). Retrospective evaluation of intersubject brain registration. *IEEE Trans. Med. Imag.* **22**, 1120-1130.

Jeffreys, H. (1961). *The Theory of Probability.* Oxford University Press.

Kim, H. Y., Giacomantone, J. and Cho, Z. H. (2005). Robust anisotropic diffusion to produce enhanced statistical parametric map from noisy fMRI. *Computer Vision and Image Understanding* **99**, 435-452.

Kim, S., Smyth, P. and Stern, H. (2006). A nonparametric Bayesian approach to detecting spatial activation patterns in fMRI data. In: *Proc. 9th MICCAI.* LNCS 4190, 217-224. Springer Verlag, Copenhagen.

Mériaux, S., Roche, A., Dehaene-Lambertz, G., Thirion, B. and Poline, J.-B. (2006). Combined permutation test and mixed-effect model for group average analysis in fMRI. *Hum. Brain Mapp.* **27**, 402-410.

Natarajan, R. and Kass, R. E. (2000). Reference bayesian methods for generalized linear mixed models. *J. Amer. Statist. Assoc.* **95**, 227-237.

Nichols, T. and Holmes, A. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum. Brain Mapp.* **15**, 1-25.

Roche, A., Mériaux, S., Keller, M. and Thirion, B. (2007). Mixed-effects statistics for group analysis in fMRI: A nonparametric maximum likelihood approach. *Neuroimage* **38**, 501-510.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996). BUGS 0.5: Bayesian Inference Using Gibbs Sampling - Manual. Tech. rep., MRC Biostatistics Unit, Cambridge.

Talairach, J. and Tournoux, P. (1988). *Co–Planar Stereotaxic Atlas of the Human Brain. 3-Dimensional Proportional System : An Approach to Cerebral Imaging.* Thieme Medical Publishers, Inc., Georg Thieme Verlag, Stuttgart, New York.

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S. and Poline, J.-B. (2007a). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *Neuroimage* **35**, 105-120.

Thirion, B., Tucholka, A., Keller, M., Pinel, P., Roche, A., Mangin, J.-F. and Poline, J.-B. (2007b). High level group analysis of FMRI data based on Dirichlet process mixture models. In: *IPMI*. Vol. 4584 of LNCS, 482-494. Springer Verlag.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *Ann. Statist.* **22**, 1701-1728.

Woolrich, M., Behrens, T., Beckmann, C., Jenkinson, M. and Smith, S. (2004). Multi-level linear modelling for fMRI group analysis using Bayesian inference. *Neuroimage* **21**, 1732-1747.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F. and Evans, A. (2002). A general statistical analysis for fMRI data. *Neuroimage* **15**, 1-15.

Xu, L., Johnson, T. and Nichols, T. (2007). Bayesian spatial modeling of fMRI data: A multiple-subject analysis. Tech. rep., The University of Michigan Department of Biostatistics.

CEA, Neurospin, Gif-sur-Yvette, France.

E-mail: merlin.keller@cea.fr

INRIA, Saclay, France

E-mail: alexis.roche@cea.fr

CEA, Neurospin, Gif-sur-Yvette, France.

E-mail: alan.tucholka@cea.fr

CEA, Neurospin, Gif-sur-Yvette, France.

E-mail: bertrand.thirion@cea.fr

(Received April 2007; accepted June 2008)