NeuroImage 44 (2009) 62–70

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Technical Note

False discovery rate revisited: FDR and topological inference using Gaussian random fields

Justin R. Chumbley *, Karl J. Friston

Wellcome Trust Centre for Neuroimaging, University College London, UK

ARTICLE INFO

Article history:
Received 22 December 2007
Revised 9 May 2008
Accepted 12 May 2008
Available online 23 May 2008

ABSTRACT

In this note, we revisit earlier work on false discovery rate (FDR) and evaluate it in relation to topological inference in statistical parametric mapping. We note that controlling the false discovery rate of voxels is not equivalent to controlling the false discovery rate of activations. This is a problem that is unique to inference on images, in which the underlying signal is continuous (*i.e.*, signal which does not have a compact support). In brief, inference based on conventional voxel-wise FDR procedures is not appropriate for inferences on the topological features of a statistical parametric map (SPM), such as peaks or regions of activation. We describe the nature of the problem, illustrate it with some examples and suggest a simple solution based on controlling the false discovery rate of connected excursion sets within an SPM, characterised by their volume.

Introduction

In this note, we consider the detection of distributed signals in image data, using statistical parametric mapping (SPM). The notion of a distributed signal is critical, in that it forces us to consider signal (as well as noise) as spatially continuous without compact support. Examples of distributed signals include induced responses in EEG studies that are distributed over frequency and peristimulus time or hemodynamic responses in fMRI that are mediated by molecules that diffuse rapidly over space. When signal is distributed, one might intuitively *define* the signal at a spatial location (e.g., voxel) as the value of the signal process at that location. However, *voxel-wise* approach this leads to several problems. *A priori,* all points have signal (see Table 1), so it is illogical to examine a null hypothesis of no signal. This compels us to define treatment effect or activation as signal above some *ad hoc* threshold. Second, the multiple-comparison problem becomes severe (with thousands of voxels contributing to family-wise error). These considerations lead to the notion of an activation that is defined in terms of the signal's topological features (*e.g.*, maxima, spatial extent etc). This converts a continuous signal into a discrete set of features, whose statistics can be examined in the usual way. The notion of a topological response finesses the interpretation of inference and allows for rigorous control of a smaller multiple comparison problem (Friston et al., 1991; Worsley et al., 1992). Under the topological perspective, a response or activation is an attribute of the signal profile over voxels; it is therefore a category error[1] to call a voxel activated. For example, we refer to a peak in a SPM as "an activation" not a collection of activations at voxels subtending the peak. The implications of this category error can be quite profound, because it permits images to be treated collections of discrete voxels or statistical tests that do not consider the continuity constraints under which the data were generated. In this work, we look at FDR procedures from the topological perspective. In particular, we show that, in the context of smooth distributed signals, conventional FDR procedures do not control the FDR of either voxels or topological features. The purpose of this note is to promote discussion of current public-software implementations of voxel-wise FDR and their usefulness.

False discovery rate procedures were introduced to neuroimaging by Genovese et al. (2002). Since their introduction, they have enjoyed considerable use. Controlling false discovery rate (FDR) provides a more sensitive analysis than the conventional control of family-wise error. This is particularly important for neuroimaging, which faces a severe and rather complicated multiple comparison problem. However, the problem faced by conventional FDR procedures is that they regard SPMs as a collection of discrete tests. This is in contrast to random field theory approaches, which consider an SPM to be a lattice approximation to an underlying continuous process. This distinction is not trivial. Inference, using random field theory, is about topological features of the SPM, such as the number of maxima or regions, their spatial extent or their peak height. On the other hand, inference using false discovery rate treats each voxel as a separate feature. This can lead to the following problem, which is best illustrated with an example:

- Imagine that we declare a hundred voxels significant using an FDR criterion. 95 of these voxels constitute a single region that is truly

* Corresponding author. Wellcome Trust Centre for Neuroimaging, 12 Queen Square, London WC1N 3BG, UK.
E-mail address: j.chumbley@fil.ion.ucl.ac.uk (J.R. Chumbley).

[1] Assigning an attribute to something that cannot possess that attribute.

1053-8119/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.05.021

> active. The remaining five voxels are false discoveries and are dispersed randomly over the search space. In this example, the false discovery rate of voxels conforms to its expectation of 5%. However, the false discovery rate in terms of regional activations is over 80%. This is because we have discovered six activations but only one is a true activation. This is a contrived example but illustrates nicely the problem we want to address.

In brief, when we make an inference using SPM it is about a topological feature *e.g.* inflection points, or clusters above a threshold. It is not about each voxel in that cluster (or more formally the excursion set). This is why one only reports the cluster, usually in terms of its maximum value and location. Conventional family-wise procedures (*e.g.*, Bonferroni correction) cannot support this sort of inference because they have no notion of topology. In other words, the fact that two voxels are part of the same cluster is incidental to both inference and the way the results are reported. This limits the usefulness of procedures like the Bonferroni correction and FDR in imaging and was the motivation for random field theory approaches to topological inference based on differential topology (Friston et al., 1991; Worsley et al., 1992). In short, the Bonferroni correction controls the false positive rate of voxels, whereas SPM controls the false positive rate of features. Conventional FDR procedures control the false discovery rate of voxels, whereas they should be controlling the false discovery rate of features.

This problem with FDR is articulated nicely by Heller et al. (2006); “Recognizing that the fundamental units of interest are the spatially contiguous clusters of voxels that are activated together, we set out to approximate these cluster units from the data by a clustering algorithm especially tailored for fMRI data” (see also Pacifico et al., 2004; Benjamini and Heller, 2007). We pursue the same theme but using a simple approach and standard results from random field theory.

This paper comprises two sections. The first presents the theoretical background to conventional inference in neuroimaging, false discovery rate and a quantitative illustration of the problem introduced above. We then consider alternative formulations of FDR based on the topology of excursion sets. The second section provides some worked examples and evaluates the procedures using simulated and real data.

Table 1
Why signal is a smooth analytic function of its support

Physical reasons (point-spread functions)
• In fMRI, signal is acquired in *k*-space or Fourier space; this means signal and noise are measured in terms of Fourier coefficients and, after projection onto anatomical space, have continuous support.
• Similarly, for PET and other modalities that reconstruct images using filtered back-projection from scintillation counts.
• Space-time EEG and MEG data are continuous functions of space and time, if they are formed from interpolation using continuous basis functions over space. In the time domain, signal is continuous in the sense that electromagnetic activity is never zero.
Data-feature and processing reasons
• All time-frequency data from EEG and MEG are smooth because they are formed from (windowed) Fourier transforms of the times-series, from sensor or source space (*e.g.*, Kilner et al., 2005
• Smoothing of fMRI constant images for between-subject (second-level) analyses with Gaussian filters render signal continuous. This is required to account for inter-subject variability in functional anatomy
• Smoothing of grey-matter segments required by voxel-based morphometry makes data a continuous measure of grey matter probability density (*e.g.*, Ashburner and Friston, 2000)
• Source reconstructed data from EEG and MEG uses spatial smoothness priors to provide a unique solution to the ill-posed inverse problem (*e.g.*, Baillet and Garnero, 1997; Mattout et al., 2006).
Biological reasons
• Population neuronal activity propagates through intrinsic and lateral connections and is modelled in neural-field models as a function of space and time using wave equations (*e.g.*, Nunez, 1974; Jirsa and Haken, 1997; Breakspear et al., 2006)
• Hemodynamic signals are initiated by rapidly diffusing signals, which have infinite support in finite time (e.g. nitric oxide; Friston, 1995)
• Neuronal activity is inherently distributed by extrinsic connections in the brain.

Theory

This section examines a commonplace procedure: voxel-wise FDR on smooth data. Our purpose is to show that the implicit assumptions about signal and noise may be untenable. Regarding the former, *voxel-wise* FDR on data with continuous signal is strictly illogical[2]; *i.e.*, if the signal does not have compact support, the FDR must be zero. We nevertheless use a tolerant measure of FDR and show that voxel-wise thresholds fail to control the FDR of voxels or regions. Second, these procedures assume that correlations among the noise are independent of their spatial relation, which is not true. These observations mean that voxel-wise FDR cannot support or reject the claim that “that some parts of the brain are activated under treatment”. With a view to a remedy, we note that FDR control using random field theory relaxes the assumption that signal has compact support. Inference is conditioned on the null everywhere; if there is no signal anywhere before smoothing, there is no signal anywhere after. It also relaxes the assumption that noise correlations are independent of location. We therefore exploit random field theory to implement reasonable FDR control on topological features of interest.

Topological inference

Conventionally, neuroimaging uses classical inference to protect against false positives in the context of multiple dependent comparisons (see Nichols and Hayasaka (2003) for a review of common approaches). This usually uses some form of statistical parametric mapping, which entails the adjustment of *p*-values using random field theory. This spatially adjustment plays the same role as a Bonferroni correction for discrete data and controls family wise error (FWE); *i.e.*, the rate of making one or more false positive declarations over the search volume. Critically, random field theory regards the data as realizations of a continuous process in spatial or other dimensions. This is in sharp contrast to procedures like the Bonferroni correction, which consider images to be collections of discrete voxels, with no continuity properties. The random field theory adjustment is more appropriate for continuous data, whose continuity properties place special constraints on spatial dependencies. For smooth data, these constraints are harnessed by random field theory to provide a much more accurate and sensitive adjustment than the equivalent Bonferroni correction.

P-value adjustments for continuous data were introduced using the theory of level-crossings in stochastic processes to control the false positive rate of statistical maxima or peaks (Friston et al., 1991). The distributional approximations were easy to formulate because a peak is defined by a negative second derivative and a zero-crossing of the first derivative (*i.e.* a peak is flat at its top). Controlling the number of peaks, or regional activations, by thresholding is not the same as controlling the false positive rate of voxels. This is because each region can contain many voxels. Put simply, the expected number of supra-threshold voxels is the expected number of regions (*i.e.*, peaks) times the expected number of voxels per region (Friston et al., 1994). Clearly, this means the expected number of voxels is greater than the expected number of peaks. Another way to look at this is in terms of topological inference; one is not making an inference about a voxel, but a topological feature of a process or function of position in the image. These features are generally attributed to a region or cluster (more formally, a connected excursion set above threshold; Worsley et al., 1992). At high thresholds, the number of clusters is equal to the

[2] FDR is defined as FDR = $E(V/R)$, the expected fraction of all positive decisions R, that are false positive decisions V. To be applicable, this definition requires that a subset of the image is signal-free. In images with continuous signal (e.g. after smoothing), all voxels have signal and consequently there are no false positives; FDR (and FWE) must be zero.

number of maxima. The key thing here is that we are not making inferences about voxels but about features of an underlying topology.[3] This means controlling the false positive rate of voxels is irrelevant. The critical thing to control is the false positive rate of the features we are making inferences about.

Random field theory

Shortly after the introduction of statistical parametric mapping, Worsley et al. (1992) showed, in a seminal paper, that random field theory (RFT) could be used to provide adjusted p-values based on the expected Euler characteristic. The Euler characteristic is, effectively, the number of clusters minus the number of holes. At high thresholds, this is roughly equal to the number of maxima or regions. In this context, the distributional approximations for maxima (Friston et al., 1991) and those using the Euler characteristic converge (Worsley et al., 1992). The advantage of the latter was that they could be generalized to any number of dimensions. Over the next few years, this approach to topological inference was extended and refined to provide distributional approximations for the maxima and spatial extent of activated regions, for a whole series of statistical maps, in an arbitrary number of dimensions. An annotated bibliography can be found at http://www.fil.ion.ucl.ac.uk/spm/bibliography.html. (see Worsley et al., 2004).

In the present context, it is important to reiterate that inference under random field theory pertains to clusters or maxima. This is because the FWE of clusters (*i.e.* the Euler characteristic) is controlled, not the FWE of voxels (which is expected Euler characteristic times the expected number of voxels per cluster). These clusters are reported in terms of their maxima (in tables) or constituent voxels (in a graphical SPM). However, one is not saying each voxel is significant (if one were interested in a pre-specified voxel there would be no multiple comparison problem). One is simply saying the probability of getting a cluster with the observed attributes (peak height or volume) by chance is sufficiently small to warrant reporting.

In summary, RFT represents the statistical behaviour of SPMs by modelling noise as Gaussian random fields. At no point do we *require* an assumption about the alternative hypothesis; the underlying signal process. For example, we can simply ask "is there a surprising number of suprathreshold voxels in this cluster, assuming there is no signal anywhere in the image?" Nevertheless, a raft of physical and biological considerations suggests that signal is indeed continuous over space (see Table 1). This means we can make the following interpretation; any topological data-feature, deemed improbable under the null, provides evidence for the existence of such a feature in the signal. Happily, we can then call on (discrete) multiple-test formalisms from mainstream statistics to control the FWE or FDR over the *discrete* set of topological features.

False discovery rate procedures

FDR was introduced by Genovese et al. (2002) as an alternative family-wise error procedure. FDR procedures are more sensitive because they do not control the false positive rate but the false discovery rate. The false discovery rate is simply the proportion of tests declared significant that have been falsely declared significant. These procedures were developed fairly recently for families of discrete tests (Benjamini and Hochberg, 1995). Their principal use is as a screening procedure that helps identify candidates for further analysis. For example, in drug screening, it is important not to miss potentially interesting drugs that can be evaluated further. Although FDR is extremely elegant and simple, we want to focus on a fundamental shortcoming, in the context of continuous imaging signals. Like the Bonferroni correction, FDR procedures were designed for families of discrete tests. They do not cover data with continuous signal or discrete data, which have been smoothed *post hoc*. Because they cannot represent topological features, they cannot furnish inferences about regional effects like random field theory. This means that although the expected number of falsely discovered voxels can be controlled, there is no way of controlling the expected number of falsely discovered regions. This can be demonstrated quite easily:

A toy simulation

To demonstrate, quantitatively, the difference between controlling the false discovery rate of voxels and clusters, we performed a set of two-dimensional simulations. We simulated eight images, each with $V=128\times128$ voxels, whose values were sampled from the normal distribution. A true signal of 24×24 voxel and height 0.75 was placed in the centre of this two-dimensional image (see rationale below). The resulting innovations were smoothed with a Gaussian kernel (full width half maximum of six voxels). Note that, as with real data, signal is therefore propagated from its (24×24 voxel) source to all areas of the image (with attenuation according to the distance from signal source and the form of the convolution kernel). This is formally equivalent to using unsmoothed data with a continuous underlying signal. A one-sample t-test was applied to each voxel to form a simple SPM with seven degrees of freedom. The uncorrected p-values, associated with the ensuing t-values were ordered and a threshold controlling FDR was computed. This procedure is described in Genovese et al (2002) and entails finding the largest p-value that is below a line of slope q as shown in Fig. 1a. q is the upper bound on the FDR. We used $q=0.05$, which means we aim to control the expected false discovery rate at 5%.

To avoid a nonsensical implementation of FDR, in the case where signal is continuous, we have two options: Define a heuristic threshold indicator that labels "weak" signal as "no signal", or define true discoveries as voxels that belong to some compact signal before smoothing. Both are *ad hoc* but to explicate FDR we chose the latter: Once the t-fields had been thresholded, true-positive discoveries were defined as voxels, which lay within the original $576=24\times24$ signal domain (recall that due to smoothing, this is necessarily an underestimate[4]).

Fig. 1b displays one realization of t-values over the search space. Beneath this, Fig. 1c depicts the same data as a contour map, where filled white areas indicate discovered voxels. In this example, there are 4 clusters, comprising 1017 voxels. The largest region encompasses the true signal (before smoothing). This cluster contained 931 voxels. For this example, the false discovery rate, in terms of voxels, was 40%. In terms of regions, there are 4 clusters but only one is a true discovery. Therefore the false discovery rate in terms of clusters is 75%.

We repeated this procedure 500 times for twenty different spatial widths of the signal. Estimated false discovery rates for voxels and clusters are shown in Fig. 2. The key thing to observe is that as the width of the activation increases, relative to the smoothness of the noise, the false discovery rate for clusters increases markedly from around 46% to 89%. In other words, the majority of *regions* declared significant under this criterion would be false. Furthermore, the false discovery rate of *voxels* is substantially bigger than expected; rising from about 20% to 50% as the signal volume gets smaller relative to smoothness (or relative smoothness increases). This is much bigger than the expected value of 5%; so what has gone wrong with FDR?

The reason, as alluded to above, is that our definition of "signal" ignores any signal propagated elsewhere in the image by the convolution or smoothing. This unsatisfactory state of affairs arises because there is no principled way to assign voxels as being truly

[3] The measures of signal we use ('spatial extent', 'number of activated regions') are not unique characterisations and depend on an arbitrary cluster-forming threshold. This is an inherent aspect of these measures.

[4] In smoothed images or images with continuous signal, all voxels have signal and consequently there are no false positives; FDR (and FWE) must be zero.

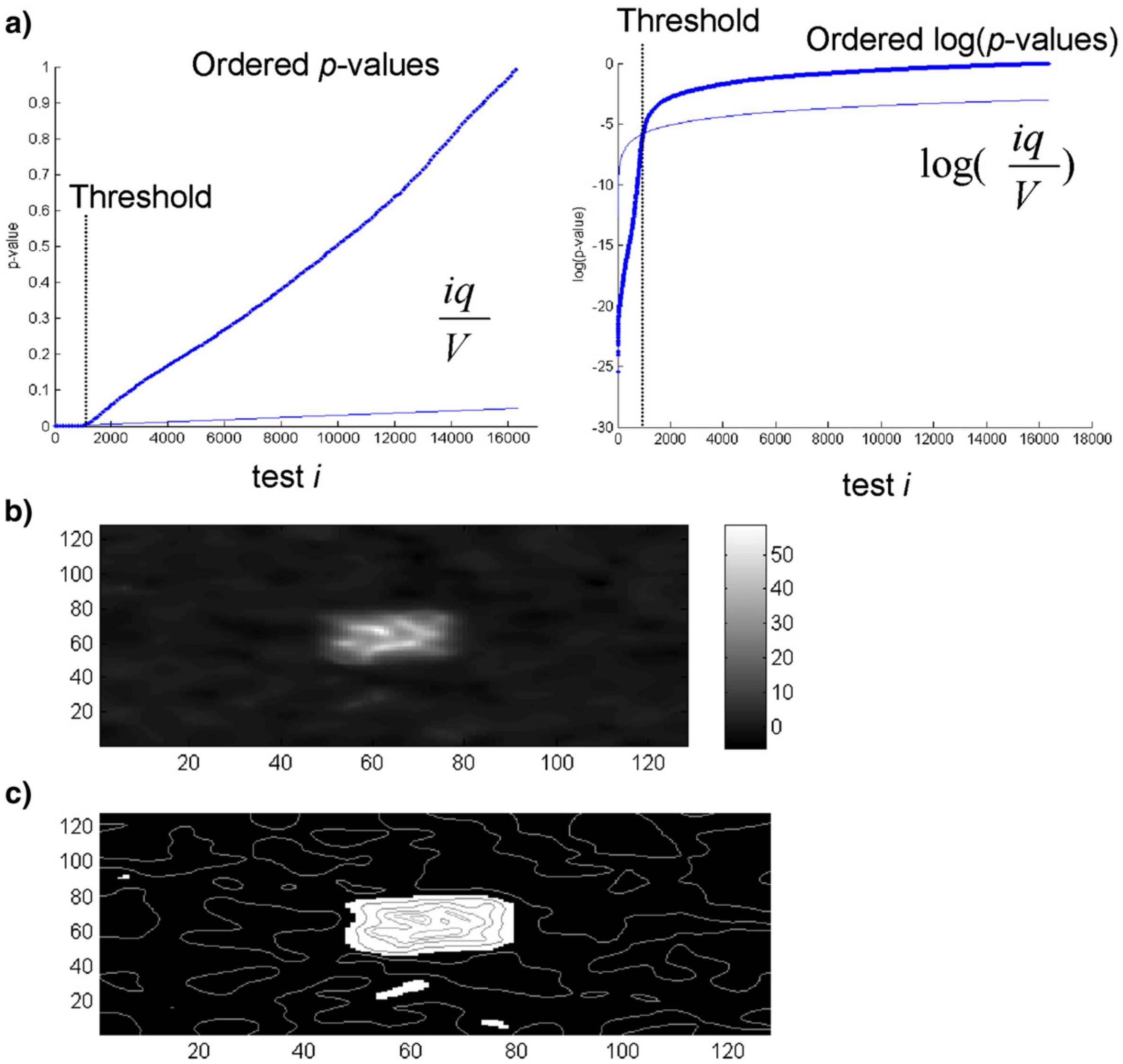

Fig. 1. Panel a displays the ordered uncorrected p-values of every voxel in the 128 × 128 t-image shown in panel b: with (right) and without (left) log scaling. The point of intersection between this empirical curve and the function iq/V constitutes the (adaptive) FDR threshold. Supra-threshold voxels are displayed in panel c.

active or not in the context of continuous processes (e.g. these data after smoothing). This means that unless signal has compact support (*i.e.*, a compact subset of the volume contains signal), the FDR of voxels is ill-defined. In this case, the notion of an "activated voxel" has no meaning (*i.e.*, it represents a category error).

Volume of activation vs. activated volume

This is a fundamental problem for any inference device that borrows from conventional statistics. When the signal or response to treatment is continuous, it is a category error to assign the attribute "activated" to a voxel because a voxel encodes one point in space. In this context, FDR on voxels is not defined. In the example above, convolution renders every voxel part of an activation, in the sense signal is never exactly zero anywhere. As mentioned above, FDR on voxels treats each voxel as categorically active or not (and does not consider things like smoothness). The topological perspective does not have to contend with this category error because the spatial extents of clusters or the number of maxima are well-defined (to within a defining threshold or level).

It is important to realise that all the key papers on FDR (*e.g.*, Genovese et al., 2002; Pacifico, et al., 2004; Heller et al., 2006; Benjamini and Heller, 2007) only examine signals with compact support. At no point in the devolvement of FDR procedures for neuroimaging has anyone considered the case of smooth analytic signals and the failure of FDR under these conditions. For example, Pacifico et al. (2004) develop false discovery control for random fields, by formulating effects in terms of (Lebesgue) measures or volumes; this "extends false discovery rates to random fields, for which there are uncountably many hypothesis tests". These procedures control the FDR of the volume of search space or the clusters comprising the excursion set. FDR on volume corresponds to a FDR of voxels and, as such, is prey to the same category error introduced above. Indeed, the examples used to illustrate the approach use geometric shapes as "signal" so that each voxel can be labelled as signal or not signal. These are reminiscent of the models used in "image-restoration" (Rosenfeld and Kak, 1976) and machine-vision problems (Davies, 2005); however they are not appropriate models of distributed signals (*e.g.*, neuronal activations). The cluster-based FDR procedure described in Pacifico et al. (2004) defines a cluster as true if the proportion of its volume that is truly activated exceeds some threshold τ. This is an *ad hoc* definition and again rests on the assumption that some measure (*i.e.*, volume) can be labelled as activated. Although the excursion set of an activation can have volume, volume itself cannot have the attribute

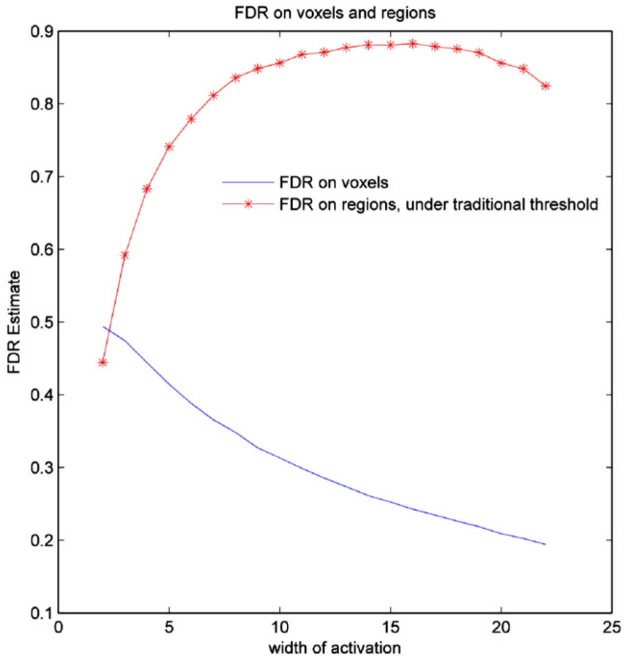

Fig. 2. Depicts the estimated FDR of voxels (solid curve) versus clusters (studded curve) as attained by the conventional voxel-wise FDR procedure. These graphs were obtained by averaging over repeated noisy, realisations, as described in the text.

activated; in the sense that a mountain can cover a large area but a location in that area cannot be a mountain. The result is a procedure that controls "the proportion of clusters in which false rejections occur". The problem is that a true cluster is defined in terms of the proportion of its volume that is activated but a volume cannot be activated. In the next section, we take a much simpler and tenable approach to the controlling the FDR of clusters, using the volume of activation.

Activations as functions

It could be argued that activation in brain images can be defined on voxels because the data are just pictures of objects with compact support (*e.g.*, ocular dominance columns in visual cortex). In fact, the signals in PET and fMRI data do not have compact support but are continuous and analytic (differentiable) functions of location. This is because the images are formed by back-projection and inverse Fourier transform of the sampled data respectively (*i.e.*, the signal and noise are mixtures of basis functions with infinite support; Twieg, 1983). Even in the absence of these mathematical considerations, hemodynamic signals are mediated by rapidly diffusing substances that preclude a compact support (Friston, 1995). Indeed, most adaptive smoothing algorithms that try to reserve functional topology, use diffusion-based kernels (*e.g.*, Huang and Cheng, 2005; Harrison et al., 2007).

In other applications of topological inference, the distributed nature of the signal is more self-evident. For example, in the detection of induced responses using time-frequency decompositions of electromagnetic data (Kilner et al., 2005), responses are defined explicitly in topological terms (*e.g.*, peaks of gamma band activity late in peristimulus time; Kaiser et al., 2002). In short, inference in neuroimaging calls on distributed models of signal that do not have compact support. This means a particular voxel or volume element cannot be labelled as signal or not signal. In this context, the FDR of voxels or volumes has no meaning and inference on discrete topological features is required. If signal can be defined on voxels (*e.g.*, in image-processing and computer vision problems), then one can use conventional statistics and SPM is unnecessary.

Summary

In summary, although false discovery rate procedures represent a good idea, when each test or voxel has meaning in its own right, their unqualified use in detecting regional activations may not be appropriate. This is because a regional activation can be a topological feature of a continuous or extended process that is not covered by current FDR procedures. In the context of large activations, the majority of regional effects reported using current FDR procedures could easily be false discoveries. It may be important to appreciate this when using the FDR in neuroimaging when signals do not have compact support. We next consider adaptations of the FDR procedures to accommodate topological inference.

Controlling topological false discovery rate

FDR based on spatial volume

One simple solution to the problem highlighted above is to control the false discovery rate of topological features as opposed to voxels. In other words, apply the FDR procedure to the null distribution of features such as cluster-volume or peak height. In this work, we focus on FDR for cluster-volume. As stated in Genovese et al. (2002) any valid statistical test with a known null distribution can be subject to FDR control. In this section, we use this procedure with simulated and real data to compare and contrast it with conventional voxel-based FDR.

The null distribution of the number of voxels in each cluster (conditional on that cluster existing) above some *ad hoc* threshold is already known for Gaussian processes (Friston et al., 1994). In fact, this p-value is provided routinely in the SPM tables, under uncorrected p-values based on spatial-extent (the product of this p-value and the Euler characteristic is the corrected p-value at the cluster level). These uncorrected p-values enable FDR control on the proportion of false positive clusters that exceed size s, by thresholding the p-values as described in Genovese et al. (2002). In this context, the topology is characterized with a list of clusters (as opposed to a list of tests or voxels). FDR, in this instance, can be regarded as a step-wise procedure, where a Bonferroni correction is applied to the p-value of the largest cluster. Note that a Bonferroni correction to all the uncorrected p-values based on cluster-volume is a "cheap and cheerful" way of controlling FWE within random field theory. The step-wise component converts it into a FDR control procedure. For detailed discussion of which stepwise procedures control the false discovery rate in multiple hypotheses testing, see Sarkar (2002). These conventional procedures are now viable because the SPM has been reduced to a list of cluster-volumes that has no inherent topology.

As noted by one of our reviewers; in the multiple testing framework used to develop FDR on voxels, the number of tests (*i.e.*, voxels) is fixed; and in a continuous setting uncountably many (Pacifico et al., 2004). However, for FDR on clusters, the number of tests is a finite random variable, even for a fixed threshold. If we condition on the number of tests (clusters) observed, we can argue that the conditional FDR given the observed number of clusters is less than 5% and hence, after averaging over all possible numbers of clusters, the FDR is also less than 5%. This rests on the assumption that the volume of a cluster is independent of the number of clusters in any search volume (i.e., ignoring boundary effects; Friston et al., 1994).

Cluster-wise FDR has been addressed previously in the context of fMRI (Heller et al., 2006; Benjamini and Heller, 2007). For example, Benjamini and Heller (2007) control the FDR on contiguous clusters (as defined by independent data) before proceeding to voxel-wise FDR. As in Pacifico et al. (2004), we use random field theory as a model of the data but in a much more straightforward way. To motivate these approaches, we present numerical results that contrast the voxel-wise and cluster-wise FDR approaches and show how the former fails in relation to the latter.

Quantitative evaluations on simulated data

We repeated the simulations of the previous section using both voxel and cluster-wise FDR. The upper-left panel of Fig. 3a shows discovered *voxels*, as attained by the FDR procedure on voxels. As emphasized above, this results in excessive false regional discoveries (here 8 discovered regions, as opposed to just one underlying true region). The lower panels of Fig. 3a depict the same data. Here however, FDR thresholding has been performed at the cluster level. Discovered *clusters* are shown in lower-left frame of Fig. 3a. These were obtained by performing FDR thresholding on the *p*-values associated with the

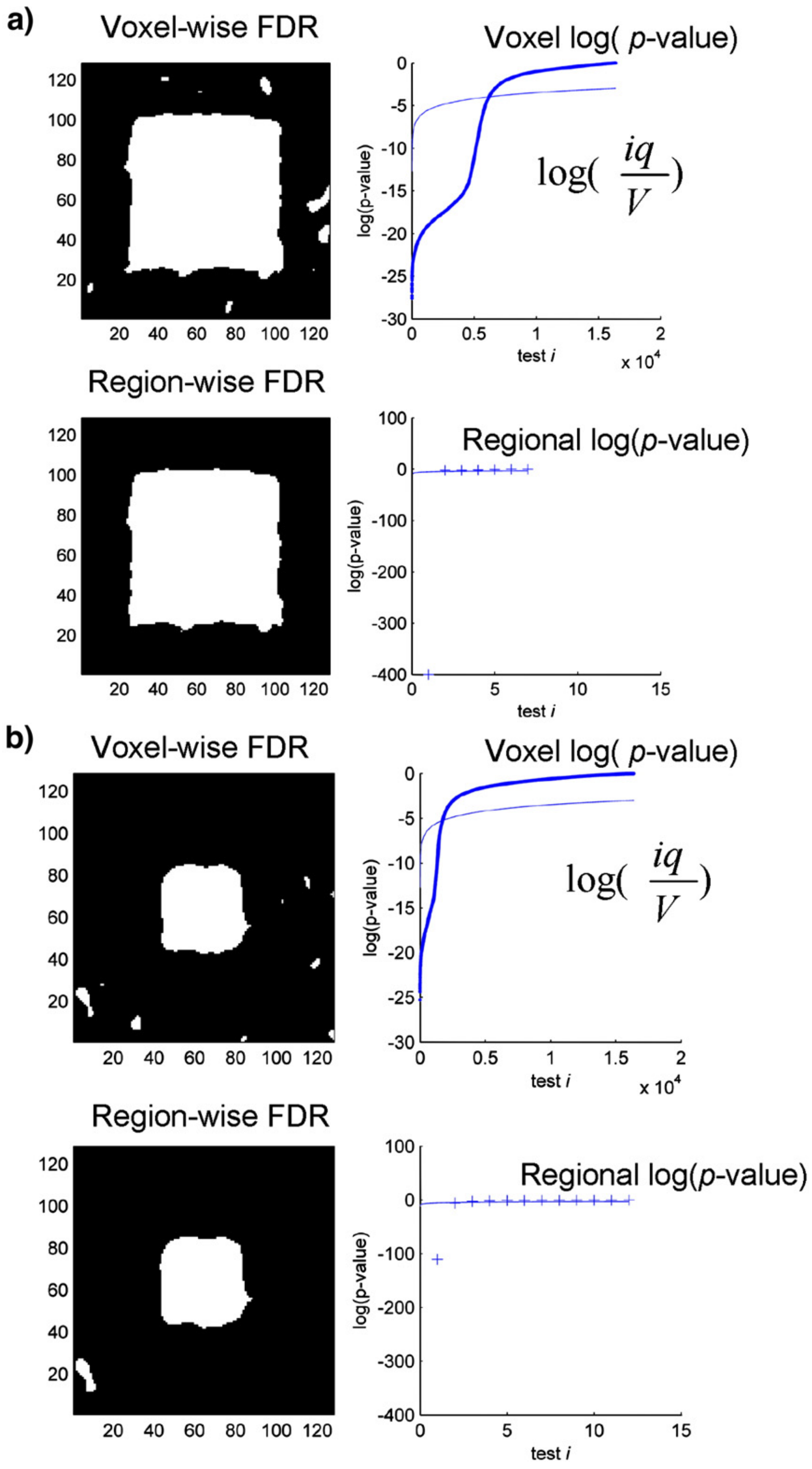

Fig. 3. The left panels compare the results obtained from the voxel-wise FDR thresholding (upper) versus spatial-extent FDR thresholding (lower). The log (p-values) underlying these thresholds are displayed to their right. Panel a differs from panel b in having a true underlying signal with more restricted support. The FDR threshold obtains at the intersection of the ordered log(p-values) and the thin line corresponding to logic (iq/V).

volume of each region (lower-right panel); in particular, the volume of each connected subset of voxels surpassing an (arbitrary) height threshold of three. This spatial extent was assigned a *p*-value indicating how improbable it was, assuming that no signal is present (null images are modelled as Gaussian random fields). Approximate functional forms used to compute these *p*-values will be found in Friston et al. (1994) and Worsley (2003). Note that this procedure results in notably fewer falsely discovered regions than under the conventional FDR procedure over voxels. Fig. 3b uses the same format as Fig. 3a but presents results using a smaller spatial support for the pre-convolved signal. Results (not reported) were qualitatively similar for all other signal configurations.

We repeated the above procedures 500 times for twenty different sizes of the pre-convolved signal. Fig. 4 contrasts the estimated expectation of FDR for clusters under the conventional voxel-wise approach (shown in Fig. 2) and the cluster-wise approach based on random field theory. It is evident that while the former approach leads to very large false discovery rates, the latter is much more robust and remains close to the true upper bound of $q=0.05$ over all signal configurations.

Illustrative application to real data

Finally, we applied the procedures outlined above to real data. This data set comprises whole brain BOLD/EPI images acquired on a 2 T Siemens MAGNETOM Vision system. Each acquisition consisted of 64 contiguous slices (64×64×64, 3×3×3 mm voxels). 96 images were acquired (TR=7 s) from a single subject, in blocks of six, giving sixteen 42 s blocks. Successive blocks alternated between rest and auditory stimulation, starting with rest. Auditory stimulation comprised bi-syllabic words, presented binaurally at a rate of 60 per minute. We discarded the first twelve scans to eschew T1 saturation effects. These images are stored in Analyse format and are available from the SPM site http://www.fil.ion.ucl.ac.uk/spm/data/.

Having realigned and smoothed these data with a Gaussian kernel of 6 mm FWHM, we constructed a single *t*-image for the contrast: active vs. rest. Fig. 5 shows an axial slice, taken from this *t*-image. For the purposes of comparison, we performed both voxel-wise and cluster-wise FDR thresholding on the same axial slice of the *t*-image. Fig. 6 uses the same format as Fig. 3. The upper left panel depicts voxels that have been classified as positive ("discovered") according to voxel-wise FDR (upper right panel). This panel is to be contrasted with the lower left panel which reports clusters discovered under the cluster-wise FDR procedure.

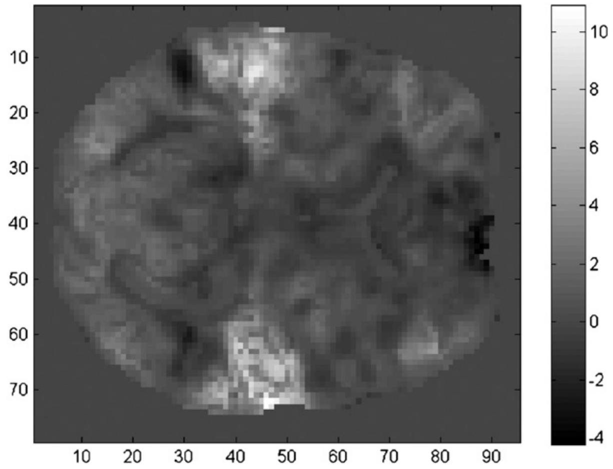

Fig. 5. An axial slice of the whole brain *t*-map from an auditory stimulation study (see main text).

Supra-threshold activations are shown in black, sub-threshold voxels in grey (white areas denote voxels with no brain tissue). Note that numerous small activations present in the voxel-wise FDR have been excluded in the cluster-wise FDR. These include (right) fronto-lateral and medial posterior activations as well as small satellites of the two principally bilateral activations. In sum, voxel-wise FDR discovers eleven regions while cluster-wise FDR discovers six. The five regions in dispute are all between one and three voxels in extent.

To calculate *p*-values for spatial volume from analytic approximations, one requires the smoothness of the data (*i.e.*, the equivalent

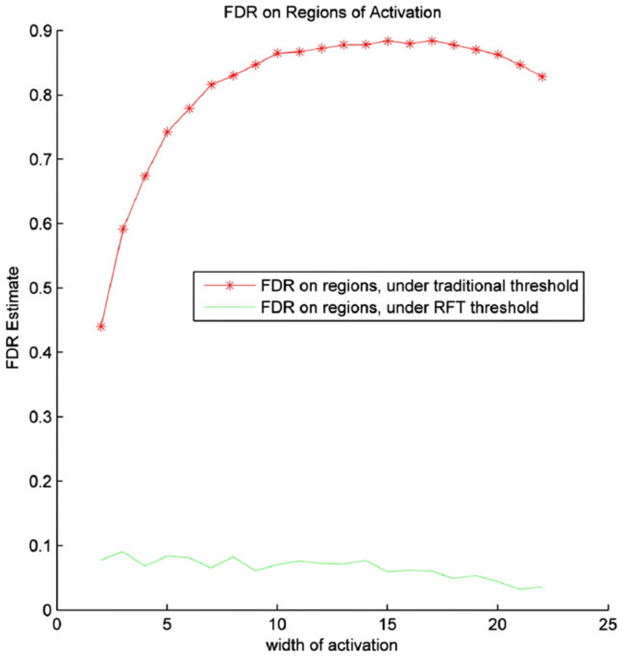

Fig. 4. Depicts the estimated regional FDR as obtained using conventional voxel-wise FDR procedure (studded curve) versus that under spatial-extent FDR (dashed curve). These graphs were obtained by averaging over repeated noisy realisations.

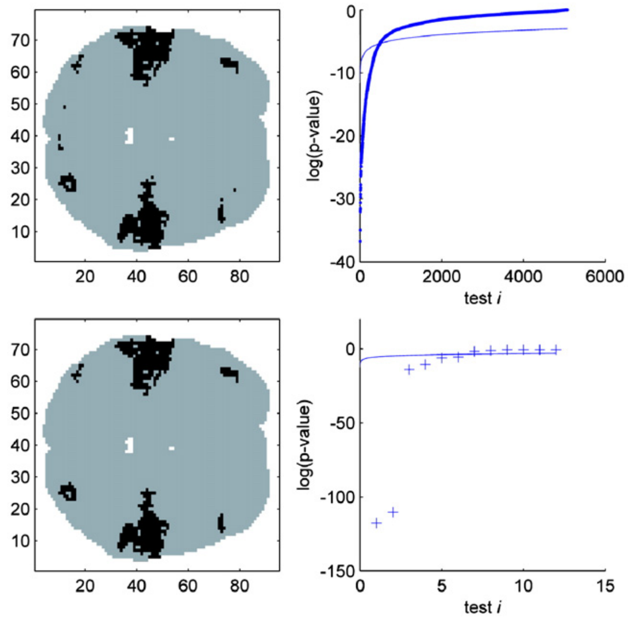

Fig. 6. The leftmost column contrast regions discovered by the conventional voxel-wise. FDR procedure (upper left) versus the spatial-extent FDR procedure (lower left). Discovered voxels are in black, sub threshold brain tissue is in grey and non-brain tissue is in white. The corresponding ordered log (*p*-values) required for these respective FDR procedures are displayed alongside each of the images.

FWHM). We estimated this smoothness via the algorithm presented in Kiebel et al. (1999) that is implemented as a standard part of SPM. It is important to note that the validity of these *p*-values relies on two things. The height threshold chosen should not be data-dependent; *i.e.*, the threshold should be determined prior to looking at the data. Second, we emphasize that the validity of all fMRI inference procedures based on RFT relies on an assumption: that the true smoothness is close to its (unbiased) point estimate. This assumption is tenable because we used an estimator pooled across the entire image (see Kiebel et al., 1999), under the assumption that the spatial correlations are roughly stationary.

Discussion

In this work we have revisited the use of FDR for topological inference on neuroimages. We have shown that there are fundamental problems with the interpretation of voxel-wise FDR. Firstly, because it contains no inherent representation of the spatial structure of signal, voxel-wise FDR cannot control the false discovery rate of regional effects. One consequence is that the *regional* FDR arising from a voxel-wise FDR analysis may be intolerably large (see Fig. 2). Second, in practice, the use of voxel-wise FDR relies on a definition of activation that precludes distributed responses; in other words, signal exists in some discrete, proper subset of voxels, each of which can be activated or not. However, under this assumption, because images are smooth (generally analytic) functions of space and have continuous signal, the existence of activation in *one* voxel can induce activation in *every* voxel. Historically, people have adopted a pragmatic simplification in which continuous data like images are regarded as a “bag” of voxels, each of which can be switched on or off. This heuristic permits the use of voxel-wise FDR. However, it fails completely in the context of image smoothness (see Table 1) that is enforced by many applications (*e.g.*, analysis of PET images, voxel-based morphometry, time-frequency analysis of EEG and second-level or between-subject fMRI analyses). In this regard, it is notable that, in the presence of smoothing, the voxel-wise FDR violates the putative upper bound (q=0.05) by a very large margin (Fig. 2).

We have outlined an approach that combines the False Discovery Rate (FDR) procedure with Random Field Theory (RFT). Following Heller et al. (2006), we have identified the elements of an FDR analysis not with voxels but with some topological property of the underlying signal. We have focused on the volume of topological excursion sets. Furthermore, our results (Fig. 4) indicate that, unlike voxel-wise FDR, estimates of the cluster-wise FDR are consistently close to their theoretical upper bound (q=0.05).

To estimate the rate of true and false discoveries in our simulations, we humoured the assumption of compact signal but, in line with standard practice, smoothed the data (signal and noise). We defined region-wise FDR as the proportion of “discovered contiguous regions” that contained “true signal” (the compact support of the signal before smoothing). Similarly, we defined voxel-wise FDR as the proportion of “discovered voxels” that contained “true signal” (this overlooks the fact that smoothing strictly invalidates FDR). Our simulations show that conventional FDR thresholding procedures do not control either of these FDR estimates. Furthermore, the FDR for a nominal voxel-based threshold changes with signal width (Fig. 2). Because signal width is unknown and non-stationary the ensuing FDR is unknown and uncontrolled. The only procedure that performed reasonably was a cluster-based FDR threshold, which controlled the false discovery rate of clusters, over all signal widths (Fig. 4).

We have emphasised that extant FDR procedures are predicated on a model of signal that is an attribute of every point in the image (*i.e.*, every location can be labelled as either “activated” or not). This is fine for images of discrete objects with compact spatial support because it allows one to infer that an object exists at a particular location. However, this is not an appropriate model for distributed signals like neuronal activity that are present everywhere in the search space. This is not a semantic nicety; it can lead to seriously misleading inferences about regional effects in neuroimaging, as we have demonstrated. The distributed nature of neuronal signals is not an approximation; it is a fundamental property of the brain, certainly at the level of population dynamics that are modelled with neural field equations (which are explicit wave-functions of space and time; *e.g.*, Nunez, 1974; Jirsa and Haken, 1997; Breakspear et al 2006) or are inferred using distributed source reconstruction techniques from electromagnetic data (*e.g.,* Baillet and Garnero, 1997; Mattout et al., 2006).

If signal is smooth and does not have bounded support, signal exists everywhere in an image. This means that all voxels contain signal and it is unhelpful to think of a voxel as activated or not. In this context, activations are an attribute of the profile over voxels and we require FDR control on clusters or maxima. However, there are situations (*e.g.*, in image restoration), where signal has bounded support (*e.g.*, an object is ether present at a particular location in a photograph or it is not). One perspective, on the distinction between signal as a continuous function of position and signal confined to a proper subset of locations, is provided by the smoothness of the signal, in relation to noise (Keith Worsley; personal communication).

When the signal's smoothness reduces to zero, the image restoration model becomes plausible; in the sense that signal falls off very quickly and is effectively bound to each voxel. In this context, voxel-based FDR thresholds are valid and can be more sensitive than cluster-based thresholds: simulations using a point signal, whose width is much less than the point spread-function inducing noise correlations, show that voxel-based FDR thresholds are more sensitive and specific than equivalent cluster-based thresholds, provided one assesses the FDR of clusters (Keith Worsley; personal communication). This recapitulates the well-known behaviour of FWE procedures, where the sensitivity of cluster-based inference is surpassed by peak-based tests when the smoothness of the signal falls below the smoothness of noise. Furthermore, it concurs with the results in Figs. 2 (and 4), which show the FDR of clusters increases with signal width, under a voxel-based threshold (note these results consider only signals that have the same, or greater, width as noise).

These observations suggest that the FDR of both voxels and clusters, defined by conventional voxel-based thresholds, depends on signal width. This is interesting but it presents a problem, because we do not know the width of signal and, even if we did, it would not be stationary, even in an image restoration setting. Furthermore, in an imaging context, the signal is always more dispersed than noise, because the effective point-spread function of the imaging device is applied to both signal and noise. Under this lower bound on signal dispersion (*i.e.*, when signal and noise have the same width), both voxel and cluster FDR, using conventional voxel-based FDR thresholds, are unacceptably high. Taken together, these considerations suggest that voxel-based FDR thresholding may have a limited role in imaging.

It should be noted that the utility of FDR procedures, in general, relies on there being a large number of tests, where most of them are null. Moving from the voxel-level to the cluster-level effectively reduces the number of tests considered, especially for smooth data. One might anticipate that cluster-level FDR thresholds may reverse the trend to capricious reporting associated with conventional FDR control.

We emphasize that FDR based on cluster-volume is a specific example of a potentially broad class of FDR procedures that perform inference on discrete topological features. Most obviously, inference might be performed on the maxima of an image. We note that one cannot turn to routine random field theory to do this because the Euler characteristic is a [scalar] topological measure of the entire excursion set (*i.e.* all clusters). However, it is possible to return to the original formulation presented in Friston et al. (1991; see also Worsley, 2005) and compute the null distribution of maxima. The order statistics for the *p*-values can then be based on the null distribution of cluster-size or maxima-values.

Acknowledgments

This work was funded by CoMPLEX, UCL and the Wellcome trust. We thank Emily for her help in preparing this manuscript and our reviewers for their help and guidance is addressing these issues.

References

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. NeuroImage 11 (6 Pt 1), 805–821 . Jun.

Baillet, S., Garnero, L., 1997. A Bayesian approach to introducing anatomo-functional priors in the EEG/MEG inverse problem. IEEE Trans. Biomed. Eng. 44 (5), 374–385.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. J. R. Stat. Soc., B. Methodol. 57 (1), 289–300.

Benjamini, Y., Heller, R., 2007. False discovery rates for spatial signals. J. Am. Stat. Assoc. 102 (480), 1272–1281 (10).

Breakspear, M., Roberts, J.A., Terry, J.R., Rodrigues, S., Mahant, N., Robinson, P.A., 2006. A unifying explanation of primary generalized seizures through nonlinear brain modeling and bifurcation analysis. Cereb. Cortex 16 (9), 1296–1313. Sep.

Davies, E.R., 2005. Machine Vision: Theory, Algorithms, Practicalities. Morgan Kaufmann.

Friston, K.J., 1995. Regulation of rCBF by diffusible signals: an analysis of constrains on diffusion and elimination. Hum. Brain Mapp. 3, 56–65.

Friston, K.J., Frith, C.D., Liddle, P.F., Frackowiak, R.S., 1991. Comparing functional, (PET) images: the assessment of significant change. J. Cereb. Blood Flow Metab. 11 (4), 690–699 Jul, 1991.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., Evans, A.C., 1994. Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. 1, 214–220.

Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15, 870–878.

Harrison, L.M., Penny, W., Ashburner, J., Trujillo-Barreto, N., Friston, K.J., 2007. Diffusion-based spatial priors for imaging. NeuroImage 38 (4), 677–695. Dec.

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., Benjamini, Y., 2006. Cluster-based analysis of FMRI data. NeuroImage 1;33 (2), 599–608. Nov.

Huang, F., Cheng, H., 2005. Vijayakumar S.Gradient weighted smoothing for MRI intensity correction. Conf. Proc. IEEE Eng. Med. Boil. Soc. 3, 3016–3019.

Jirsa, V.K., Haken, H., 1997. A derivation of a macroscopic field theory of the brain from the quasi-microscopic neural dynamics. Physica, D 99, 503–526.

Kaiser, J., Lutzenberger, W., Ackermann, H., Birbaumer, N., 2002. Dynamics of gamma-band activity induced by auditory pattern changes in humans. Cereb. Cortex 12 (2), 212–221. Feb.

Kiebel, S.J., Poline, J.B., Friston, K.J., Holmes, A.D., Worsley, K.J., 1999. Robust smoothness estimation in statistical parametric maps using standardized residuals from the General Linear Model. NeuroImage 10, 756–766.

Kilner, J.M., Kiebel, S.J., Friston, K.J., 2005. Applications of random field theory to electrophysiology. Neurosci. Lett. 21;374 (3), 174–178. Feb.

Nichols, T., Hayasaka, S., 2003. Controlling the family-wise error rate in functional neuroimaging: a comparative review. Stat. Methods Med. Res. 12 (5), 419–446. Oct.

Nunez, P.L., 1974. The brain wave equation: a model for the EEG. Math. Biosci. 21, 279–297.

Mattout, J., Phillips, C., Penny, W.D., Rugg, M.D., Friston, K.J., 2006. MEG source localization under multiple constraints: an extended Bayesian framework. NeuroImage 30, 753–767.

Pacifico, M.P., Genovese, C., Verdinelli, I., Wasserman, L., 2004. False discovery control for random fields. J. Am. Stat. Assoc. 99 (468), 1002–1014.

Rosenfeld, A., Kak, A.C., 1976. Digital Picture Processing, Computer Science and Applied Mathematics. Academic Press, New York.

Sarkar, S.K., 2002. Some results on false discovery rate in stepwise multiple comparison procedures. The Annals of Statistics, vol. 30, pp. 239–257. No. 1.

Twieg, D., 1983. The *k*-trajectory formulation of the NMR imaging process with applications in analysis and synthesis of imaging methods. Med. Phys. 10 (5), 610–621.

Worsley, K.J., 2003. Developments in random field theory, In: Frackowiak, R.S.J., Friston, K.J., Frith, C., Dolan, R., Friston, K.J., Price, C.J., Zeki, S., Ashburner, J., Penny, W.D. (Eds.), Human Brain Function, 2nd edition. Academic Press. 2003.

Worsley, K.J., 2005. An improved theoretical *P* value for SPMs based on discrete local maxima. NeuroImage 28 (4), 1056–1062 Dec.

Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12, 900–918.

Worsley, K.J., Taylor, J.E., Tomaiuolo, F., Lerch, J., 2004. Unified univariate and multivariate random field theory. NeuroImage 23 (Suppl 1), S189–S195.